Analyzing and Modeling Mass Casualty Events in Hospitals – An Operational View via Fluid Models

Noa Zychlinski

Analyzing and Modeling Mass Casualty Events in Hospitals – An Operational View via Fluid Models

Research Thesis

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Industrial Engineering

Noa Zychlinski

Submitted to the Senate of the
Technion - Israel Institute of Technology

Tamuz, 5772 Haifa July 2012

The Research Thesis Was Done Under
the Supervision of
Dr. Izhak Cohen and Prof. Avishai Mandelbaum
in the Faculty of Industrial Engineering and
Management

The Generous Financial Help of the Technion is
Gratefully Acknowledged

I would like to express my deep appreciation and gratitude to my supervisors, Dr. Izhak Cohen and Prof. Avishai Mandelbaum, for their endless support, advice and guidance through our research.
Great thanks to Dr. Moshe Michelson and Dr. Shlomi Israelit from Rambam hospital in Haifa, Prof. Haya Kaspi from the Technion and Roni Leker from Mathworks.

Finally, I would like to thank my family, especially Moshe, my husband, the one and only.

CONTENTS

Abstract: **1**

1 Introduction **3**

1.1 Research Objectives and structure of the thesis **6**

2 Background and Literature Review **7**

2.1 Mass Casualty Events **7**

2.2 Fluid Models **9**

3 Problem Definition **11**

4 Fluid Model - Possible Approaches **13**

4.1 First Fluid Model **13**

4.1.1 Single station analysis **13**

4.1.2 A network analysis **15**

4.2 Second Fluid Model **17**

4.2.1 Single Station Analysis **17**

4.2.2 A Network Analysis **19**

4.3 Fluid Models' Comparison and Simulation results **20**

5 Minimizing Mortality **28**

5.1 Optimal solution **31**

5.2 Greedy Problem **35**

5.3 Comparison between Optimal and greedy solutions **40**

5.3.1 Sensitivity Analysis **42**

5.4 Problem Analysis **44**

5.5 Minimal time window for resource allocation **45**

6 Summary and Conclusions **49**

7 Few Direction for Future Research **49**

8 Rreferences **50**

Appendix I: **54**

Appendix II: **54**

Appendix III: **60**

LIST OF FIGURES

1 Activity Chart of ER in conventional MCE……..……....5
2 A Black Box Model……...…..9
3 The network flow for immediate casualties……...….11
4 Cumulative arrivals and Departures from Queue and from Treatment……...…….18
5 A quadratic arrival rate function…….…….....20
6 Total number of Casualties in each station and Queue Simulation vs. First Fluid model – First Scenario…….……..21
7 Total number of Casualties in each station and Queue Simulation vs. Second Fluid model- First Scenario…….……..22
8 A two surges arrival rate……...…….….23
9 Total number of Casualties in each station and Queue Simulation vs. First Fluid model-Second Scenario……..24
10 Total number of Casualties in each station and Queue Simulation vs. Second Fluid model – Second Scenario……...……...25
11 Cumulative Arrivals and Departures - Second Fluid Model……..……....26
12 Cumulative Arrivals and Departures – Station 1 – Second Fluid Model…….....….26
13 Optimal Surgeons allocation – First Scenario……...……...……..32
14 Optimal Surgeons allocation – Second Scenario……...…….....……..33
15 Optimal Surgeons allocation – Third Scenario……..…….……34
16 Optimal Surgeons allocation – Greedy solution - First Scenario……..…….38
17 Optimal Surgeons allocation – Greedy solution - Second Scenario……..…….39
18 Optimal Surgeons allocation – Greedy solution - Third Scenario……...……...40
19 Total Number of Casualties – Optimal vs. Greedy Solutions…….…..……...41
20 Mortality Rate – Optimal vs. Greedy Solutions…….…….……41
21 Cumulative Mortality – Optimal vs. Greedy Solutions…….…….……42
22 Optimal Surgeons allocation – minimal time window – First Scenario…….47
23 Optimal Surgeons allocation – minimal time window – Second Scenario…….....….48
24 Optimal Surgeons allocation – minimal time window – Third Scenario……...….48

LIST OF TABLES

1 MSE results for the two Fluid Models……...…….……23
2 Sensitivity Analysis of Greedy vs. optimal solution…….……...……43-44
3 Optimal surgeons' allocation policies…….……44

ABSTRACT:

In a Mass-Casualty Event (MCE) the work brought by casualties exceeds the capacity for taking care of them. Such events, unfortunately, happen all the time. They may have either a world-wide effect (e.g., the 2004 Indian Ocean tsunami that killed over 200,000 people) or a local one (e.g., a terror attack or a train accident that sends tens of casualties to a hospital) – in both cases there is a continuous imbalance between the workload and the available resources. Therefore it is very important to prepare for such events.

We concentrate on the operational aspects of MCE in hospitals. A central problem in establishing a hospital's emergency plan is the inability to forecast the performances of healthcare services. When an MCE occurs, suddenly and all at once, the demand for healthcare staff and facilities increases, and an emergency plan must be quickly implemented. Such a plan has clinical and operational components. The focus of this research is on the latter.

The mainstream approach for modeling MCEs is through simulation. We, on the other hand, develop a mathematical model (a fluid model) that captures the operational performance of a hospital during and after an MCE. The results from the fluid model are in agreement with a simulation model that was developed for validation purposes.

We formulate optimization problems with the objective of minimizing the mortality of casualties. We then solve the problems by combining theory with numerical analysis. Our research enhances the understanding of the operational effects of MCEs. It provides managerial insights that support dynamic resource allocation throughout the MCE. We then capture these insights in terms of managerial guidelines.

NOMENCLATURE

$\lambda(t)$ – Arrival rate of immediate casualties to the ED at time t (number of casualties per minute).

μ_i – Individual treatment (service) rate in Station i (number of casualties per minute).

θ_i – Individual mortality rate in Station i.

p_{ij} – transition probability from Station i to Station j.

R_i – The required number of surgeons per patient at Station i.

$Q_i(t)$ – Number of casualties in Station i (queue + service) at time t.

$Lq_i(t)$ – Queue Length in Station i at time t.

$N_i(t)$ – Number of resources in Station i at time t.

$A_i(t)$ – Cumulative arrivals to Station i at time t.

$Ds_i(t)$ – Cumulative departures from Station i at time t.

$Dq_i(t)$ – Cumulative departures from Station i's queue at time t.

1 INTRODUCTION

A mass-casualty event (MCE) is defined as an unusual, unpredictable situation in which, at a certain moment, there are more casualties than the system is able to manage. During MCEs, the hospital emergency services must treat a large number of patients that suddenly arrive [1]. When a disaster occurs, the number of patients who require treatment in Emergency Departments could easily overwhelm hospitals' resources [2].

The major challenges that hospitals face in an MCE include surge capacity issues, the fact that they are already at or near capacity for emergency and trauma services, a lack of on-call specialists and nurses, the need to coordinate between competing health care systems and incompatibilities in communications systems [3].

MCEs can be classified into two categories: (1) those that result in an immediate or sudden impact on healthcare services and (2) those that result in a developing or sustained impact [3]. From a Service Engineering point of view, this classification is based on the arrival rate of casualties to receive medical service.

The first category of MCEs includes events such as the detonation of a conventional bomb, NBC (nuclear, biologic, chemical) attack, airplane or train crashes and natural disasters such as earthquakes or tsunamis. This immediate impact category is characterized by a large numbers of casualties at the outset of the event. In some cases there may be a second wave of casualties due to secondary exposure.

The second MCE category features events such as a massive exposure to anthrax or smallpox, or a potential case of influenza pandemic, in which there would be a gradual increase in the number of people affected, possibly rising to a catastrophic number of patients. In this type of MCE, the number of cases may decline due to treatment and prophylactic efforts. This type of MCE would necessitate a more sustained response, as the impact would be felt over a much longer period than the immediate-impact MCE. In our research, we focus on the first category of MCEs with sudden immediate impact.

A different classification of MCEs is to conventional MCEs vs. non-conventional. The hospitals disaster plan in a non-conventional MCE (an MCE that includes toxic exposure - chemical, biological, radiological, nuclear or contamination threat) differs from a conventional disaster plan. A non-conventional MCE requires a decontamination procedure and different medical treatments ([4], [5], [6], [7]). Our research focuses on conventional MCEs.

According to Hirsberg et al. [8] and Aylwin et al. [9], Mass Casualty Event differ from Multiple Casualty Incidents (MCI) in that the former overwhelms the emergency systems, hospitals and community infrastructure, and exceeds the capability of available resources to provide optimum

trauma care. For example, human caused MCEs such as the New York World Trade Center attack in 2001 [10], the Madrid commuter train bombings in 2004 [11] and the London bombing in 2005 [9] or in nature disasters such as the Haiti earthquake in 2010 [12,13], the Turkish earthquake in 2011 [14] and the Japanese tsunami in 2011 [15].

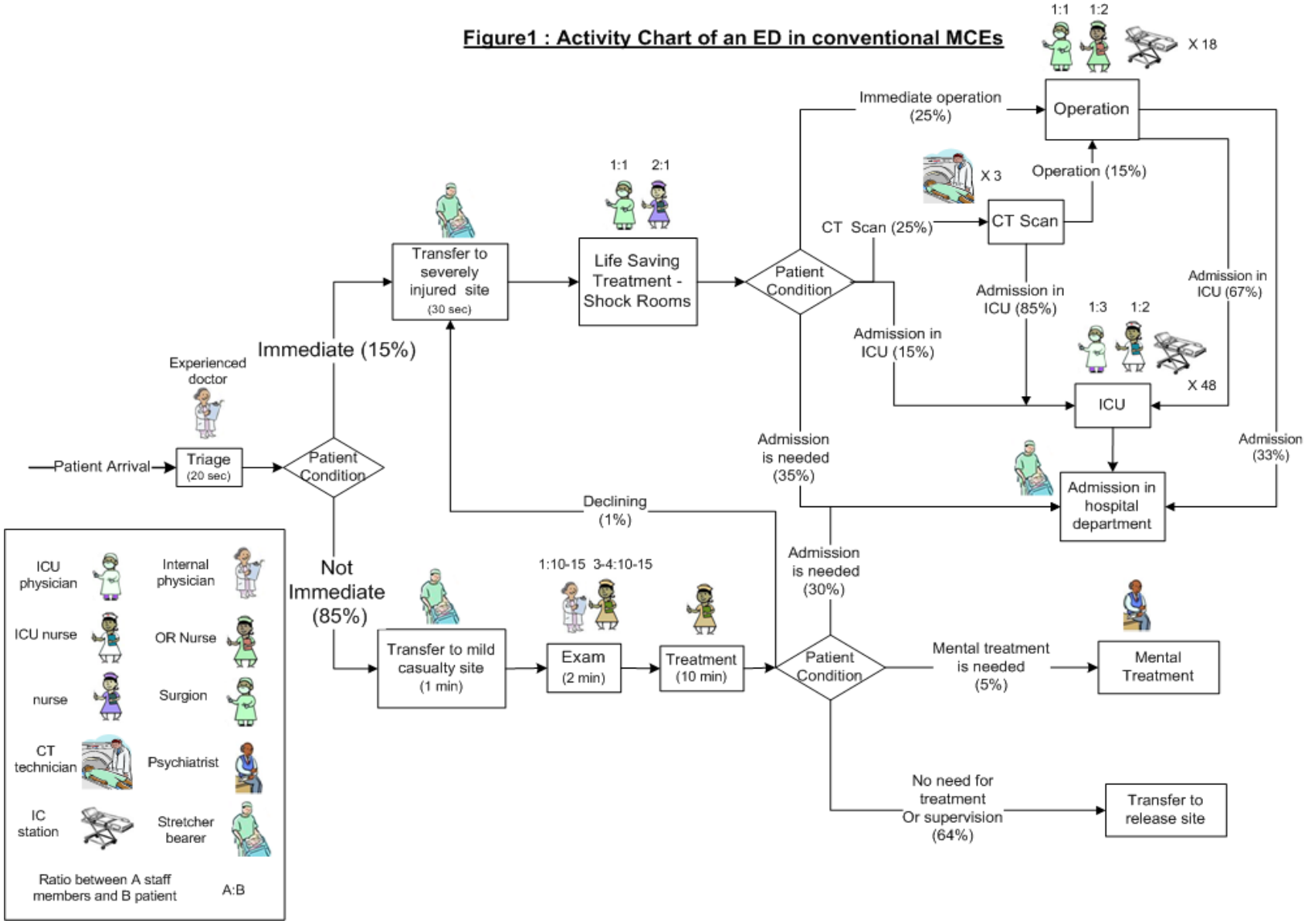
An MCE affects nearby hospitals, which have been working in steady state while treating a reasonable amount of patient, but now must start treating a large number of patients. The lack of adequate staff and equipment may cause injured patients to not receive the same level of medical treatment that would have been provided had they been treated as an individual rather than as one of multiple casualties arriving simultaneously. In order to overcome the temporary lack of qualified staff and resources, the hospital prepares itself to work under a triage strategy. Upon arrival, patients are triaged by the severity of their injuries. Severely injured must be treated immediately, since any delay can endanger their lives. Ideally, hospital allocates separate locations for each group so as to ensure that the majority of resources are allocated to the most severe injuries.

When an MCE occurs, the hospital usually receives an advanced notification at which time it immediately activates its emergency plan. Figure 1 shows the activity flow which casualties go through when entering the ED, following the hospital emergency plan. The resources include healthcare staff and equipment. Healthcare staff includes (a) physicians: internal, surgical and intensive care, (b) nurses: ED, Operation Room (OR) and Intensive Care Unit (ICU) (c) stretcher bearers. The equipment category covers ORs, Intensive Care positions and stretchers.

The main goal of a hospital's emergency response in MCEs is "to provide severely injured patients with a level of care that approximates the care given to similar patients under normal conditions" [8, 16]. In particular, the main goal of the hospital is to reduce the mortality of critically injured patients. That is the reason we chose to focus on immediate casualties and on the upper branch in Figure 1.

Previous research, which has been based either on an ED simulation during Mass Casualty Incident [8, 17] or on the analysis of the surgical response during real MCEs [9,10], found that the hospital surge capacity depends primarily on the number of available surgeons. Therefore, we chose to focus on the bottleneck resource which is surgeons. Surgeons treat patients in two stations: Shock Rooms and Operation Rooms. We seek to find a dynamic policy for allocating the surgeons between the two stations during the event.

Figure1 : Activity Chart of an ED in conventional MCEs



Modeling healthcare systems or emergency departments in general and in MCEs in particular is difficult due to the complexity, size and dynamic of the systems involved.

According to Paul et al. [18], most MCE research has focused on modeling and simulating emergency department in their steady state, while the hospital model in disaster management differs from the model of normal operations in the characteristics of patient arrivals, which changes the system nature to a transient one.

The most widespread method in operation research of MCEs is simulation [8, 16,17, 19]. We, on the other hand, offer a mathematical, dynamic model that reflects the number of casualties in each station at any time. All research that we are familiar with and used simulation model, assumed a constant arrival rate through the event. Data collected during Mass Casualties Incidents in Israel [20] and after the London bombing in 2005 [9] show a **time-varying** surge pattern in casualty arrivals. While the use of averages does not reach the actual maximum surge rate, the fluid model we suggest and the simulation model we use take into account the time-varying arrivals. According to Aylwin et al. [9], a disaster plan that is based on estimates of average surge will fail in an MCE with large number of critical casualties.

1.1 Research Objectives and Structure of the Thesis

The objectives of our research are to develop a mathematical (fluid) model for a hospital's Emergency Department (ED) during an MCE. The model will allow predicting operational performances and determine the optimal dynamic policy for resource allocations.

The thesis is organized as follows: In Section 2, we review the related literature on MCEs and fluid models. In Section 3 we define our problem and assumptions. In Section 4 we describe and compare two mathematical (fluid) models for our problem. In Sections 5 we formally define and solve the optimization problem for reducing mortality in two network stations and then describe its solution. We also illustrate the solution by a few examples which give insight into the structure of the optimal policy. In addition, we propose a heuristic greedy algorithm for the problem and compare it with the optimal solution. In Section 6 we summarize our conclusions and in Section 7 we suggest several directions for future research.

2 Background and Literature Review

2.1 Mass Casualty Events

In the literature, various aspects of MCEs are analyzed: clinical aspects [8, 9, 16, 20], social sciences aspects [13, 21, 22] and operational aspects. The latter, according to a literature survey conducted by Altay and Green [22], is yet limited, despite the fact that analyzing MCEs requires dynamic, real-time, effective and cost efficient solutions, which are most suitable for Operations Research. In addition, according to [23], the main challenges of MCEs are organizational and logistic problems, rather than trauma care problems.

When an MCE occurs in a place with poor or no medical infrastructure, the local population and government are helpless and need assistance from other countries. An example of such an event is the earthquake that struck Haiti on January 2010 [12, 13]. The number of deaths that were caused by the earthquake is estimated to be 230,000, plus approximately 250,000 injured people. Israel Defense Forces Medical Corps sent a delegation to Haiti, consisting of 121 medical personnel who set up a field Hospital. During the 10 days that the hospital was operational, its staff treated 1,111 patients, hospitalized 737 patients, and performed 244 operations on 203 patients [12].

The Israeli field hospital in Haiti managed to treat 100 patients a day despite the fact that its bed capacity was only 60 (later expanded to 72) [13]. The reason for this was efficiency and flexibility in resource allocation and staffing. For example, the distribution of injuries had changed during the time and the hospital had constantly re-balanced its resources accordingly.

Einav et al. [20] analyzed data collected at level 1 Trauma centers in Israel, during 32 MCIs caused by suicide bombings, it concluded that high staffing demand for Emergency Department (ED), Operation Rooms (OR) and Intensive Care Units (ICU) overlap, hence, surgeons are needed immediately. Therefore, the hospital emergency plan should include simultaneous appropriate resource allocation.

The hospital's surge capacity for multiple casualties in Israel, as was defined by the Emergency and Disaster Medicine Division in the Israeli Ministry of Health, is 20% of each hospital's bed capacity. Recent research disagrees with this definition and offers alternative definitions, either by a fixed number of casualties [24], or by the rate of casualty arrivals [16].

Operations Research regarding MCEs focuses on four operational stages: mitigation [25,26], preparedness [27, 28], response [29, 30] and recovery [31]. Most of these works, according to Altay and Green [22], focus on preparedness and response.

Various problems and methodologies are described in the literature regarding hospital operations during MCEs.

Simulation is widely accepted as an effective method for assisting management in evaluating different operational alternatives [32]. Sinreich and Marmor developed a general flexible simulation tool for Emergency Departments, which provides estimates regarding the current operational state and enables short term operational planning. Hirshberg et al. [17] used a discrete-event computer model of an emergency room during an urban terrorism bombing, based on accumulated data from 12 urban terrorist bombing incidents in Israel, while assuming a constant average arrival rate. They concluded that the admitting capacity of the hospital depends primarily on the number of available surgeons and defined an optimal staff profile for surgeons, residents, and trauma nurses. The researchers concluded that the major bottlenecks in the flow of critical casualties are the shock rooms and the computed tomography scanner but not the operating rooms.

Simulation models were also used to evaluate the realistic hospital capacity, considering the actual quality of care provided to severe casualties [8] and to define a quantitative relation between an increasing casualty load the level of trauma care [16] and the Time To Saturation (TTS) of the trauma teams [19], which is the time interval between the beginning of the simulation until all trauma teams reach their capacity. The TTS was used by the researches as a convenient substitute for the processing capacity of the ED trauma system. A definition of the surge capacity of hospitals as a rate of casualty arrival was also determined by a simulation model [16].

A transient model of hospital operation for a disaster response was developed by Paul et al. [18] by using a generic simulation model and Meta-model for predicting patients' waiting times and estimate hospitals' capacities for all the hospitals in the disaster region.

Mathematic Models and dynamic optimization for minimizing the fatalities in MCEs were used for setting priority assignment and scheduling casualties base on their lifetime (their tolerance to wait) and their service time in a clearing system with multiple classes of impatient jobs [33] and for medical or logistic resource allocation [34]. A dynamic optimization model was used by Fiedrich et al. [35] in order to best assign available resources to operational areas after earthquake disasters.

Planning the transportation of vital first-aid commodities to disaster-affected areas during emergency response is done using stochastic programming [36]. Logistic planning of supplies was done by solving a dynamic problem which combines the multi-commodity network flow problem and the vehicle routing problem [37]. Sherali et al. [38] prescribed an evacuation plan under hurricane or flood conditions, which minimizes the total congestion-related evacuation time by a nonlinear mixed-integer programming model and heuristic algorithm and exact implicit enumeration algorithm.

2.2 FLUID MODELS

Fluid models provide useful approximations that support performance analysis and control of large systems (high arrival rate and large number of servers) that vary in time. For example, fluid models have been used to design and analyze Markovian service network, in which large demand is being served by a large number of servers. The approximation becomes more accurate as the system grows [39]. The basic fluid model (Figure 2) refers to the system as a black box having an arrival rate function $\lambda(t)$ and a departure rate function $\delta(t)$, $t \geq 0$.

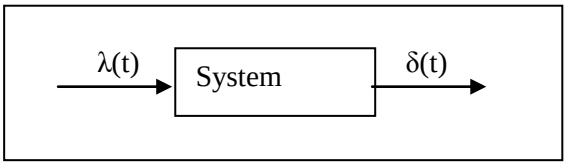


Figure 2 – A Black Box Model

Let Q(t) represent the total amount of "fluid" in the system, it can be calculated by solving the equation:

$$\dot{Q}(t) = \lambda(t) - \delta(t), \quad t \geq 0, \quad Q(0) = Q^0$$

If the system consists of N(t) servers at time t, each one with service rate μ, then we get the following differential equation:

$$\dot{Q}(t) = \lambda(t) - \mu \cdot \min(Q(t), N(t)), \quad t \geq 0, \quad Q(0) = Q^0$$

Mandelbaum et al. [40,41] used fluid approximation for modeling a multi-server queue in a single service system station with abandonment and retrials. The model was proven accurate both in steady state and in transient state, the latter caused by a sudden peak in the arrival rate. Solving the model's equations yields an estimation of the total number of people in the system at any time, the total number in queue and in service. The model allows predicting the time until the system returns to steady state. In [41] Mandelbaum et al. develop fluid and diffusion approximations for the waiting time in the system. Both approximations are "asymptotically exact as the size of the system grows large".

Oliver and Samuel [42] present an application of the fluid approach for analyzing the mail sorting procedure, modeled as a flow network. The network consists of serial and parallel processing stations

in post offices. The minimal delay for two sequential activities is found by the researchers, by equaling the processing rates of the two stations.

Analyzing time-varying queueing networks is presented by Vandergraft [43], who models the flow in a network by a set of ordinary differential equations, one for every station, and solves them by numerical methods. Vandergraft [43] characterizes five operational measurements (productivity, queue length, utilization of resources, waiting time and sojourn time) and expresses the staffing level for each station in terms of the number of customers in each station.

Whitt [44] analyzes a multi server queue with abandonment. In his research, Whitt develops a many server heavy–traffic limits in the Efficiency Driven (ED) regime for the M(n)/M(n)/s/r + M (n) model, on which arrival rates, service rates and abandonment rates are state-dependent, there are s servers and r extra waiting spaces. The development includes fluid limit, diffusion limit and limits for the steady-state distribution. In [45] Whitt develops a deterministic fluid approximation for the *G/GI/s + GI* queueing model with large *s* (queueing model for a general multi-server queue with customer abandonment) while focusing on steady state behavior.

Based on the fluid approximation developed by Whitt [44], Green et al. [46] analyze time varying service systems while focusing on methods for setting staffing requirements. They show that analyzing systems that are overloaded for long periods by using deterministic fluid models is useful and implement fluid approximation to an overloaded financial service call center. The approximation allows estimating the waiting time and queue length (which vary in time) and the total time until the system recovers from congestion that occurs during rush hours.

Yom-Tov [47] expands the framework of Mandelbaum et al. [39] on time-varying queues and develops fluid and diffusion limits for the Erlang-R model. Erlang-R captures the behavior of Re-Entrant customers, who cycle between **need** for service and being **content**: such reentrant customers are prevalent in healthcare. Yom-Tov [47] shows that fluid approximations are not only useful in analyzing time-varying systems, but also help understand the transient behavior of systems in steady-state. More specifically, the Erlang-R model is a stochastic queueing process which consists of 2-node state-dependent queues: $Q(t) = (Q_1(t),Q_2(t))$. $Q_1(t)$ represents the number of Needy patients in the system (i.e., those either waiting for service or being served), and $Q_2(t)$ the number of Content patients in the system, at time t. The fluid model yields an approximation for Q(t), $t \geq 0$.

The developed model is then used to analyze mass-casualty events in which the arrival rate changes rapidly during a short time. A numerical example, in which arrival rate is multiplied fivefold over two hours, was simulated and compared to the fluid and diffusion approximation. The comparison showed high accuracy, under the assumption that the time in critically-loaded state is negligible.

Liu and Whitt [48,49] analyze a deterministic fluid model for systems which alternate between overloaded and under-loaded intervals, having time-varying arrival rate and staffing, exponential or non-exponential service and a non-exponential abandonment time. When the system is under-loaded, the total system content is less than its service capacity, hence, there is no waiting and external input flows directly into service at time-varying rate $\lambda(t)$, $t \geq 0$. When the system is overloaded, there is no spare service capacity, so that the input is buffered in a queue, where abandonment occurs. Liu and Whitt [48,49] develop algorithms to describe time-dependent performance. They determine the time-varying potential waiting time, i.e., the virtual waiting time of an arrival at a specified time, assuming that it will not abandon. Simulations of queueing systems confirmed that the algorithm and the approximation were effective even when the number of servers was as low as 20. They also show that non-exponential service distribution played an important role in the fluid dynamics.

3 PROBLEM DEFINITION

We focus on immediate casualties, who go through the upper branch of Figure 1. The analyzed sub-network of three stations is presented in Figure 3.

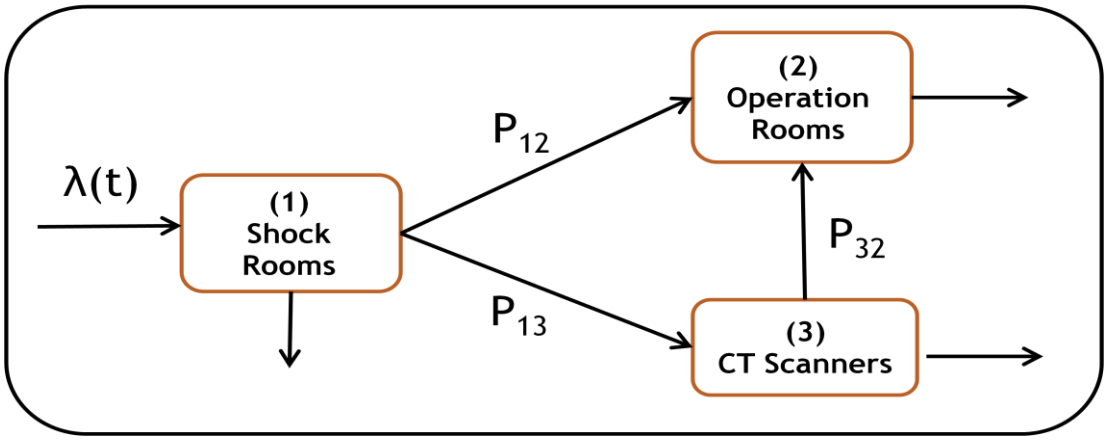


Figure 3 – The network flow for immediate casualties

Model Assumptions:

1. Immediate patients arrive to Station 1 after triage according to a general arrival rate. We assume that all immediate patients are identical in their clinical situation thus their flow throughout the network is according to First Come First Served (FCFS) priority.
2. An immediate patient that enters Station 1 at time t receives a life-saving treatment if at least one of the $N_1(t)$ surgeons is available, or otherwise queues for treatment.
3. With probability p_{12} the patient is routed into an operation room, where $N_2(t)$ surgeons are allocated.
4. With probability p_{13} the patient is routed to CT scan, where $N_3(t)$ scanners are allocated. Treatment starts immediately by an available scanner or the patient waits until one completes a previously started treatment.
5. Treatment rates are μ_1, μ_2 and μ_3 for Stations 1,2 and 3 , respectively.
6. The "effective" treatment time for a patient treated in a station includes its treatment duration and any delay time caused by unavailable surgeons.

A treatment may take place at a station only if the necessary resources (e.g., surgeons, operation room, CT scanners and medical equipment) are available. We assume that the only constraining resources are the N surgeons that are available at the hospital so, at any time, $N_1(t) + N_2(t) \leq N$ must hold.

The value of the parameters used in our analysis is based on previous research, both from data collected in trauma centers in Israel during MCIs [17,20] and on data from international MCEs [9,10,11]. According to Kosashvily et al. [24], and Einav et al. [20], recent terror assaults involve high rates of blast, penetrating injuries and unpredictable trajectories, which make the CT scans necessary. Hirshberg et al. [17], based on a simulation model, suggests that the bottlenecks in the flow of critical casualties are the shock rooms and the CT scans, but not the operating rooms.

One surgeon is required for each casualty in the shock room and for performing an operation. A CT scanner and a technician are required for performing a CT scan. The durations of service in each station are taken from previous research [17,16,19].

Our analysis includes two parts. In the first part we examine and compare two fluid models for describing the dynamics of the system. In the second part, we optimize system performance based on the fluid model we chose, by determining the optimal dynamic surgeons' allocation in Stations 1 and Station 2.

4 FLUID MODEL - POSSIBLE APPROACHES

4.1 FIRST FLUID MODEL

The first fluid model we present is based on characterizing the number of casualties in each station by a differential equation. The solution of these differential equations set is the number of casualties in each station at any time.

4.1.1 SINGLE STATION ANALYSIS

The number of casualties in Station 1 (in queue and in treatment) is expressed by the following differential equation:

$$(1)\qquad \begin{aligned} &\dot{Q}_1(t) = \lambda(t) - \mu_1(t)\cdot(Q_1(t) \wedge N_1(t)) \\ &Q_1(0) = Q_1^0 \end{aligned}$$

$$(A \wedge B) \stackrel{d}{=} \min(A, B)$$

The rate change in the number of casualties in the station at time t is given by the difference between the number of casualties that entered the station at time t and the number of casualties that departed the station at time t.

The number of departures is determined by the treatment rate, $\mu_1(t)$, multiplied by the number of casualties in treatment. When there is no queue, all casualties in the station are in treatment and, therefore, the number of casualties in treatment is $Q_1(t)$. When there is a queue, all surgeons work at full capacity and the number of casualties in treatment equals the number of surgeons, $N_1(t)$. Therefore, the number of casualties in treatment is the minimum between the total number of casualties and the total number of available surgeons in the station.

There are two options for practically solving equation (1). The first option is achieved by dividing the time interval to incremental time intervals dt, which yields the following difference equation:

$$Q_1(t+dt) = Q_1(t) + dt\cdot[\lambda(t) - \mu_1(t)\cdot(Q_1(t) \wedge N_1(t))]$$

The approximation becomes more accurate as the size of the incremental time intervals diminishes.

The second option for solving equation (1) uses the assumption that first there is no queue and from a certain time point, queue starts to form. The solution is done by dividing the total time interval into two intervals: before a queue starts to build up and after. We define the time at which a queue starts to form as the **critical time point** of the station and denote it as s_1. During the time interval until s_1 there is no queue and therefore there are no more casualties than surgeons. During the time interval after s_1

a queue starts to form and therefore there are no more surgeons than casualties. The equation can be written for the two intervals by the following non-homogeneous linear differential equations:

$$(Q_1(t) \wedge N_1(t)) = \begin{cases} Q_1(t), & 0 < t \le s_1 \\ N_1(t), & s_1 < t \le T \end{cases}$$

$$\dot{Q}_1(t) = \begin{cases} \lambda(t) - \mu_1(t) \cdot Q_1(t), & 0 < t \le s_1 \\ \lambda(t) - \mu_1(t) \cdot N_1(t), & s_1 < t < T \end{cases}$$

The solution for a non-homogeneous linear differential equation of the form:

$$\dot{f}(t) = \alpha(t) - \beta(t) \cdot f(t), \quad t \ge 0,$$

is

$$f(t) = \frac{\int_0^t \alpha(u) \cdot e^{\int_0^u \beta(w)dw} du}{e^{\int_0^t \beta(w)dw}} + \frac{C}{e^{\int_0^t \beta(w)dw}} = \int_0^t \alpha(u) \cdot e^{-\int_u^t \beta(w)dw} du + C \cdot e^{-\int_0^t \beta(w)dw}, \quad t \ge 0.$$

This yields the following solution for (1):

$$Q_1(t) = \begin{cases} Q_1(0) \cdot e^{-\int_0^t (\mu_1(w))dw} + \int_0^t (\lambda(u) \cdot e^{-\int_u^t (\mu_1(w))dw})du, & 0 < t \le s_1, \\ Q_1(s_1) + \int_{s_1}^t ((\lambda(u) - \mu_1(u) \cdot N_1(u)))du, & s_1 < t \le T. \end{cases}$$

The queue length at any time t can be expressed by the following:

$$Lq_1(t) = (Q_1(t) - N_1(t))^+, \quad t \ge 0$$

where $(x)^+ = \max(x, 0)$

If the total number of casualties in the station exceeds the total number of surgeons, the difference between the two is the number of casualties in queue. If the number of casualties is less than the number of surgeons, then this difference expresses the number of idle surgeons and, in this case, the queue length equals to zero.

4.1.2 A NETWORK ANALYSIS

The extension of the model from one station to three stations is carried out by defining a set of three differential equations, one for each station. Each equation represents the rate of change in the number of casualties at time t.

The first equation for Station 1 (Shock Rooms) remains the same. The second and third equations represent the change in the number of casualties at time t by adding the number of casualties that entered the station at time t and subtracting the number of casualties that departed the station at that moment. Casualties entering Station 3 (CT scans) come solely from Station 1, since the use of CT scans in MCEs is limited to very specific indications of severe casualties [17]. Therefore, the number of casualties that enter Station 3 (CT) at time t equals the number of casualties that departed from Station 1 (shock rooms) at time t multiplied by the transition probability p_{13} from Station 1 to Station 3. (about 25% based on previous research [9,17,20]).

Casualties entering Station 2 (Operation Rooms) come either directly from Station 1, with probability p_{12}, or from Station 3, with probability p_{32}. (Based on previous research p_{12} = 25% and p_{32} = 15% [9,17,20]).

We formalize the above as follows for t≥0:

$$\dot{Q}_1(t) = \lambda_1(t) - \mu_1(t)\cdot(Q_1(t) \wedge N_1(t))$$
$$\dot{Q}_2(t) = \lambda_2(t) - \mu_2(t)\cdot(Q_2(t) \wedge N_2(t))$$
$$\dot{Q}_3(t) = \lambda_3(t) - \mu_3(t)\cdot(Q_3(t) \wedge N_3(t))$$

$$\lambda_2(t) = p_{12}\mu_1(t)\cdot(Q_1(t) \wedge N_1(t)) + p_{32}\mu_3(t)\cdot(Q_3(t) \wedge N_3(t))$$
$$\lambda_3(t) = p_{13}\mu_1(t)\cdot(Q_1(t) \wedge N_1(t))$$
$$Q_1(0) = Q_1^0, \quad Q_2(0) = Q_2^0, \quad Q_3(0) = Q_3^0$$

One can summarize the flow equations more compactly, for t≥0:

$$(2)\quad \dot{Q}_1(t) = \lambda_1(t) - \mu_1(t)\cdot(Q_1(t) \wedge N_1(t))$$
$$\dot{Q}_2(t) = p_{12}\mu_1(t)\cdot(Q_1(t) \wedge N_1(t)) + p_{32}\cdot\mu_3(t)\cdot(Q_3(t) \wedge N_3(t)) - \mu_2(t)\cdot(Q_2(t) \wedge N_2(t))$$
$$\dot{Q}_3(t) = p_{13}\mu_1(t)\cdot(Q_1(t) \wedge N_1(t)) - \mu_3(t)\cdot(Q_3(t) \wedge N_3(t))$$
$$Q_1(0) = Q_1^0, \quad Q_2(0) = Q_2^0, \quad Q_3(0) = Q_3^0$$

Solving (2) can be achieved, as in the case of a single station, via two approaches. It is important to notice that the solution of the first equation remains the same as for the one station model and is independent of the other two equations. The analytical solution is more difficult, since the equations

are dependent. The first approach is carried out by dividing the time interval into small intervals of length dt. This yields the following:

$$Q_1(t+dt) = Q_1(t) + dt \cdot [\lambda_1(t) - \mu_1(t) \cdot (Q_1(t) \wedge N_1(t))]$$
$$Q_2(t+dt) = Q_2(t) + dt \cdot [p_{12}\mu_1(t) \cdot (Q_1(t) \wedge N_1(t)) + p_{32}\mu_3(t) \cdot (Q_3(t) \wedge N_3(t)) - \mu_2(t) \cdot (Q_2(t) \wedge N_2(t))]$$
$$Q_3(t+dt) = Q_3(t) + dt \cdot [p_{13}\mu_1(t) \cdot (Q_1(t) \wedge N_1(t)) - \mu_3(t) \cdot (Q_3(t) \wedge N_3(t))]$$

In the second approach we define the time at which a queue begins to form before station i as the **critical time point** of that station and denote it as s_i.

The third equation does not depend on the second one but not vice versa. We begin therefore, with the third equation:

$$(Q_3(t) \wedge N_3(t)) = \begin{cases} Q_3(t), & 0 < t \le s_3 \\ N_3(t), & s_3 < t \le T, \end{cases}$$

$$\dot{Q}_3(t) = p_{13}\mu_1(t) \cdot (Q_1(t) \wedge N_1(t)) - \mu_3(t) \cdot (Q_3(t) \wedge N_3(t)).$$

The solution for the third equation depends on the relation between s_1 and s_3, e.g. which station will first reach saturation and will start to form a queue. The first option is that the surgeons in the Station 1 (shock rooms) work in full capacity before the CT scanners and technicians in Station 3 do: $s_1 < s_3$. Then one obtains:

$$\dot{Q}_3(t) = \begin{cases} p_{13}\mu_1(t) \cdot Q_1(t) - \mu_3(t) \cdot Q_2(t), & 0 < t \le s_1 \\ p_{13}\mu_1(t) \cdot N_1(t) - \mu_3(t) \cdot Q_3(t), & s_1 < t < s_3 \\ p_{13}\mu_1(t) \cdot N_1(t) - \mu_3(t) \cdot N_3(t), & s_3 < t < T, \end{cases}$$

$$Q_3(t) = \begin{cases} Q_3(0) \cdot e^{-\int_0^t (\mu_3(t))dw} + \int_0^t (p_{13}\mu_1(u) \cdot Q_1(u) \cdot e^{-\int_u^t \mu_3(w))dw})du, & 0 < t \le s_1 \\ Q_3(s_1) \cdot e^{-\int_{s_1}^t (\mu_3(t))dw} + \int_{s_1}^t (p_{13}\mu_1(u) \cdot N_1(u))du, & s_1 < t \le s_3 \\ Q_3(s_3) + \int_{s_2}^t [(p_{13}\mu_1(u) \cdot N_1(u) - \mu_3(u) \cdot N_3(u))]du, & s_3 < t \le T. \end{cases}$$

The second option is that a queue begins to form in Station 3 before building up in Station 1: $s_1 > s_3$:

$$\dot{Q}_3(t) = \begin{cases} p_{13}\mu_1(t)\cdot Q_1(t) - \mu_3(t)\cdot Q_3(t), & 0 < t \le s_3 \\ p_{13}\mu_1(t)\cdot Q_1(t) - \mu_3(t)\cdot N_3(t), & s_3 < t < s_1 \\ p_{13}\mu_1(t)\cdot N_1(t) - \mu_3(t)\cdot N_3(t), & s_1 < t < T, \end{cases}$$

$$Q_3(t) = \begin{cases} Q_3(0)\cdot e^{-\int_0^t(\mu_3(t))dw} + \int_0^t (p_{13}\mu_1(u)\cdot Q_1(u)\cdot e^{-\int_u^t \mu_3(w))dw})du\ , & 0 < t \le s_3 \\ Q_3(s_3) + + \int_{s_2}^t [(p_{13}\mu_1(u)\cdot Q_1(u) - \mu_3(u)\cdot N_3(u))]du, & s_3 < t \le s_1 \\ Q_3(s_1) + \int_{s_1}^t [(p_{13}\mu_1(u)\cdot N_1(u) - \mu_3(u)\cdot N_3(u))]du, & s_1 < t \le T. \end{cases}$$

The solution for the second equation is performed in a similar way, except that there are now more options regarding the order of s_1, s_2 and s_3.

4.2 Second Fluid Model

Following Hall's model [50] for systems with long service times (relative to the waiting times), we distinguish between the departures of casualties from the queue into treatment and the departures from treatment. Since the service time in the stations is assumed long, the delay between the time a casualty leaves the queue and the time leaving the station cannot be ignored.

4.2.1 Single Station Analysis

The two departure curves, one from the queue and the other from treatment, must, according to Hall [50] p.192, satisfy the following two conditions:

$$(3) \quad Ds(t + \frac{1}{\mu}) = Dq(t)$$

$$Dq(t) = \min(A(t),\ Ds(t) + N(t))$$

The vertical difference between the two curves is given by the second condition. When there is no queue, the total number of casualties that departed from the queue equals the total number that arrived. If a queue exists, the total number that departed from the queue equals the total number that already departed the station plus the total number that are being treated at that moment. Since a queue exists, the surgeons work at full capacity and, thus, the number of casualties in treatment equals the number of surgeons.

If we assume that Q(0)=0, then until one service time ($t = \frac{1}{\mu}$), no casualties leave the station:

$$Ds(t) = 0, \qquad 0 \le t \le \frac{1}{\mu}$$

$$Ds(t) = Dq(t - \frac{1}{\mu}) = \min(A(t - \frac{1}{\mu}),\ Ds(t - \frac{1}{\mu}) + N(t - \frac{1}{\mu})), \qquad t \ge \frac{1}{\mu}$$

An illustration of the three cumulative curves is presented in Figure 4:

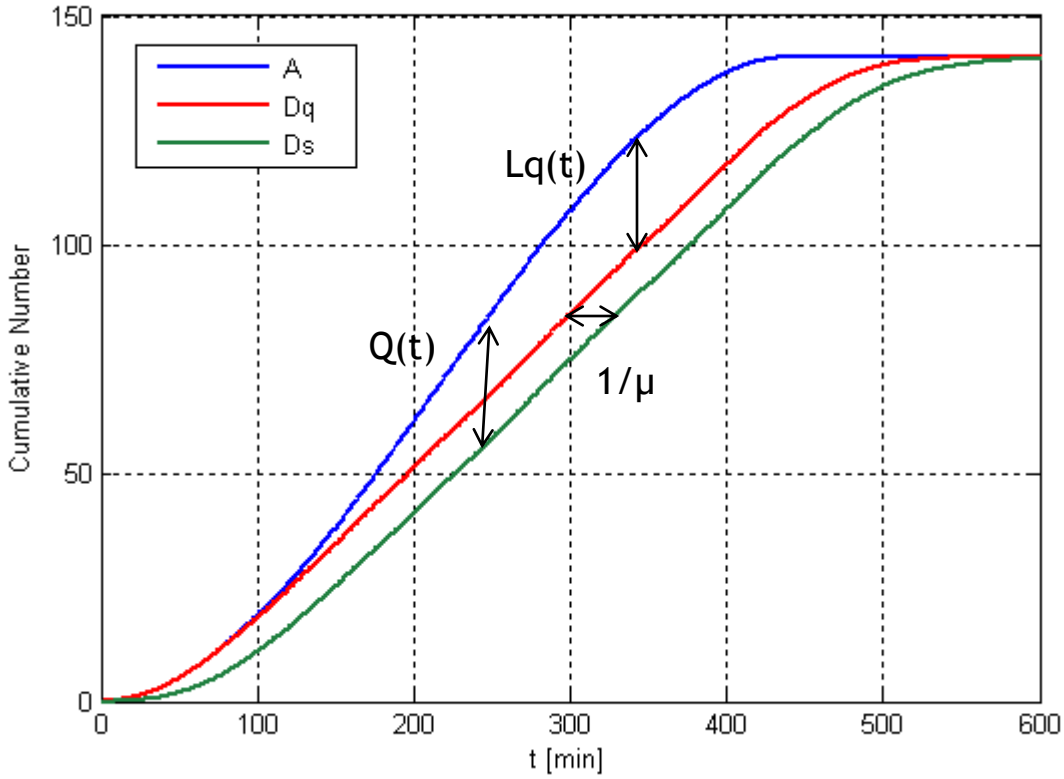


Figure 4 – Cumulative Arrivals and Departures from Queue and from Treatment

The horizontal distance between the two cumulative departure curves equals one service time and is given by the first condition. At any time t, the number of casualties that departed from the queue equals the number of casualties that will depart the station at time t plus one service time (t+1/μ).

The vertical distance between the cumulative arrivals and the cumulative departures from queue is the queue length and the vertical distance between the cumulative arrivals and the cumulative departures from the station is the number of casualties in station (in queue and in treatment).

4.2.2 A NETWORK ANALYSIS

The expansion of the second fluid model for a three stations network is based on the construction of the two departure curves (from queue and from treatment) for each station. Ds_i denotes the cumulative departures from station i and Dq_i denotes the cumulative departures from station i's queue. The two conditions for the first station remain the same as in the single station model:

$$Ds_1(t + \frac{1}{\mu}) = Dq_1(t)$$
$$Dq_1(t) = \min(A_1(t),\ Ds_1(t) + N_1(t))$$

The cumulative number of arrivals for Station 3 is the cumulative number of departures from Station 1multiplied by p_{13}, the transition probability from the Station 1 to the Station 3. In other words, $A_3(t) = p_{13}Ds_1(t)$.

Therefore, the two conditions that must be satisfied for Stations 3 are:

$$Ds_3(t + \frac{1}{\mu}) = Dq_3(t)$$
$$Dq_3(t) = \min(p_{13}Ds_1(t),\ \ Ds_3(t) + N_3(t))$$

Following the same principles, the cumulative arrivals to the second station is constructed by the cumulative departures from Station 1 and Station 3 multiplied by the corresponding transition probabilities:

$A_2(t) = p_{12}Ds_1(t) + p_{32}Ds_3(t)$

The two conditions for Station 2 are:

$$Ds_2(t + \frac{1}{\mu}) = Dq_2(t)$$
$$Dq_2(t) = \min((p_{12}Ds_1(t) + p_{32}Ds_3(t)),\ \ Ds_2(t) + N_2)$$

4.3 FLUID MODELS' COMPARISON AND SIMULATION RESULTS

In order to validate the fluid models against a discrete simulation and compare the two models, a network queueing simulation model of Figure 3 was created. The simulation was developed in SimEvent - Matlab discrete simulation tool, with 200 replications for each simulation run. Several MCE scenarios were used for the comparison. Each scenario was represented by an arrival rate function $\lambda(\cdot)$ which indicates the number of immediate casualties arriving per minute. Arrivals were sampled from a non-homogeneous Poisson process with an intensity function $\lambda(t)$, as described in Appendix I. The service times in each station were sampled from exponential distributions with rates μ_i, i=1,2,3. Transitions from stations in the network were determined according to transition probabilities. We use minutes as our time units.

We demonstrate the comparison between the models and the simulation results in two MCE scenarios distinguished by their arrival rate. The first is a quadratic arrival rate and the second is an arrival rate with two surges.

Scenario 1 - Quadratic Arrival Rate

The quadratic arrival rate, which represents the number of immediate casualties that arrive to the ED per minute, is presented in Figure 5.

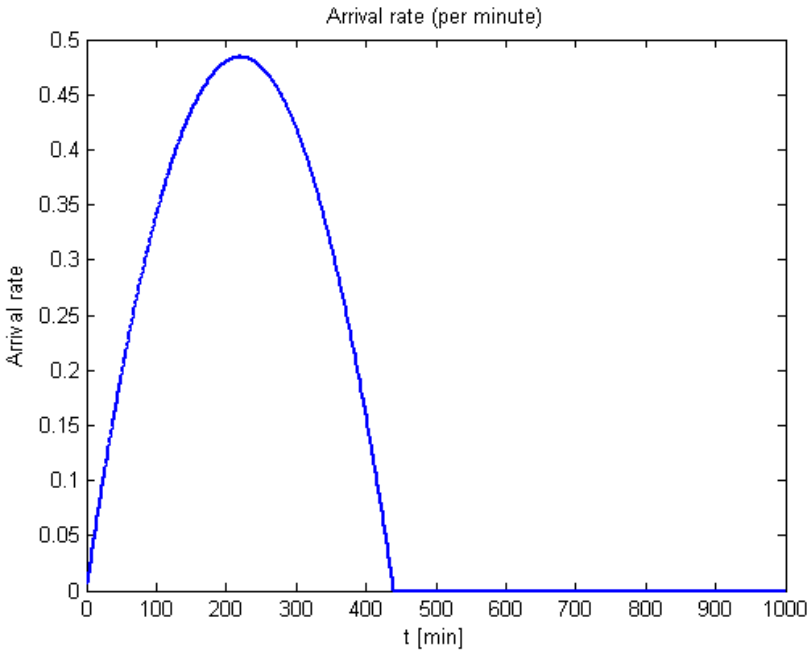


Figure 5 – A quadratic arrival rate function

The parameters used in the First Scenario are as follows:

$\lambda(t) = -1 \cdot 10^{-5} t^2 + 0.0044t$, $t \in [0,440]$,

$\mu_1 = 1/30$, $\mu_2 = 1/30$, $\mu_3 = 1/20$, $p_{12} = 0.25$, $p_{13} = 0.25$, $p_{23} = 0.15$, $N_1 = 10$, $N_2 = 5$, $N_3 = 3$

The arrival rate (per minute) is $\lambda(t)$, average treatment time is 30 minutes in the Station 1, 100 minutes in the Station 2 and 20 minutes in Station 3. 25% of the casualties in Station 1 are transferred to Station 2, 25% of the casualties in Station 1 are transferred to Station 3 and 15% of the casualties in Station 3 are transferred to Station 2. The number of available surgeons is 10 in Station 1 and 5 in Station 2. There are 3 available CT scanners in Station 3.

Figures 6 and 7 show the time-varying total number of casualties in each station and in queue for each model. The continuous line represents the number of casualties given by the model and the dashed lines represent the simulation results: average over the 200 replications and 95% confidence interval.

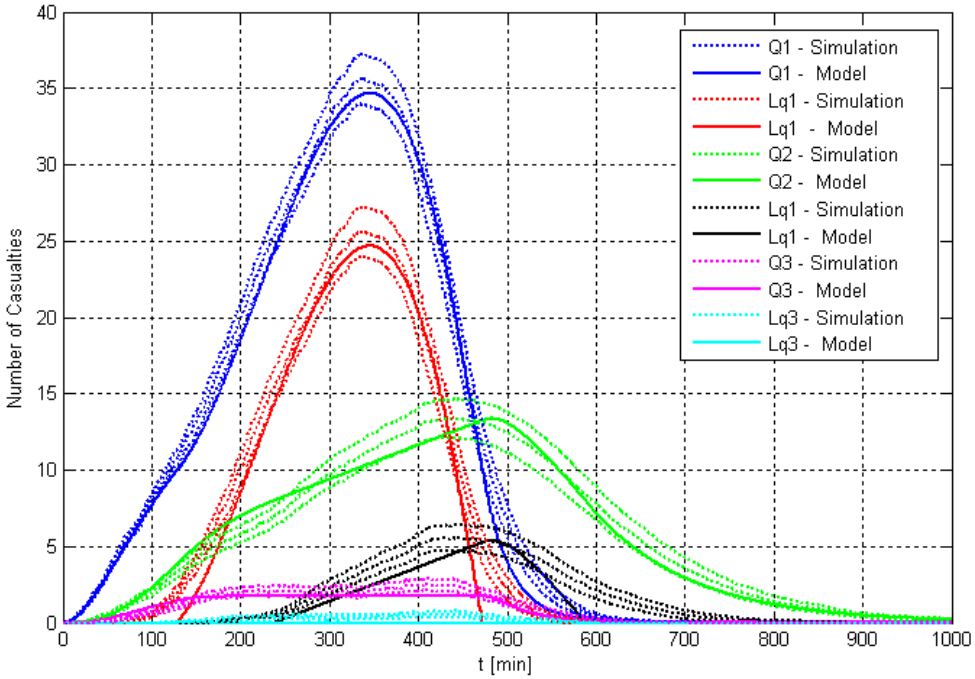


Figure 6 – Total number of Casualties in each station and Queue - Simulation vs. First Fluid model – First Scenario

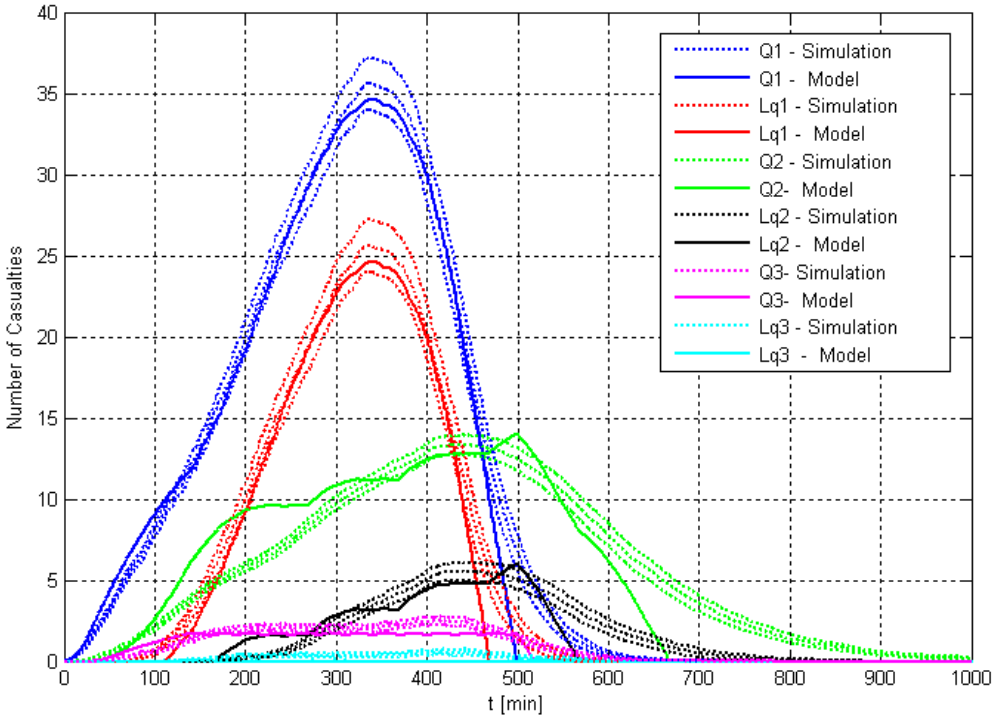


Figure 7 – Total number of Casualties in each station and Queue - Simulation vs. Second Fluid model – First Scenario

Mean Square Error (MSE) for cumulative arrivals, cumulative departures, total number in each station and queue length in each station were calculated for each model according to the following formula:

$$MSE = \sqrt{\frac{\sum_{t=1}^{T}(m_t - s_t)^2}{T}}$$

Where T denotes the total duration of the MCE scenario in minutes, m_t denotes the value calculated by the model at time t and s_t denotes the average value over the 200 replications measured by the simulation at time t.

The MSE results are summarized in the following table:

Station	Model	Cumulative Arrivals	Cumulative Departures	Total number	Queue Length
(1) Shock Room	First Fluid Model (4.1)	0.167	1.084	1.003	1.273
	Second Fluid Model (4.2)	0.167	1.699	1.62	1.243
(2) OR	First Fluid Model (4.1)	0.757	0.43	0.634	0.828
	Second Fluid Model (4.2)	0.732	1.32	1.9	0.888
(3) CT	First Fluid Model (4.1)	0.966	1.074	0.326	0.271
	Second Fluid Model (4.2)	0.961	1.153	0.427	0.271

Table 1 – MSE results for the two Fluid Models

Although the two models both predict the time of the peak and the maximal number of casualties and queue length, according to the graphs and MSE results it is apparent that the first fluid model is more accurate than the second model.

Scenario 2 – Two-surge (peaks) Arrival Rate

The two-surge arrival rate is presented in Figure 8.

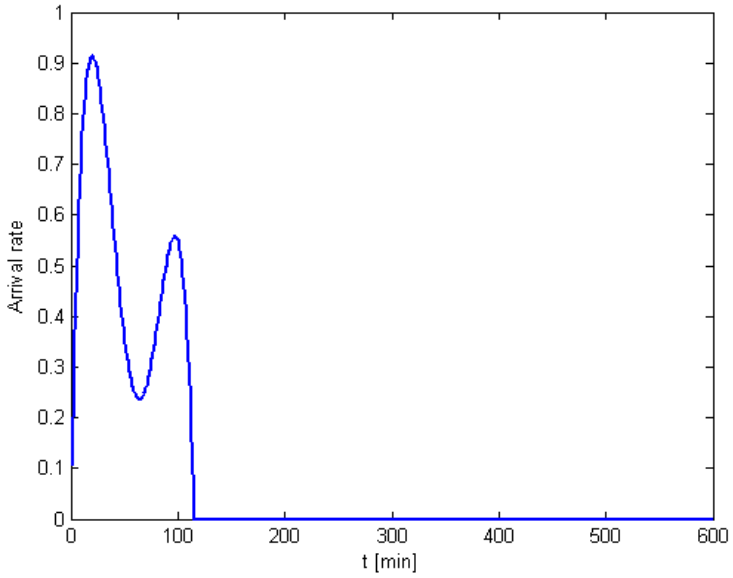


Figure 8 – A two-surge arrival rate

All parameters used in the Second Scenario, besides the arrival rate, are the same as were used in the First Scenario:

$\lambda(t) = -2.16\cdot10^{-7}t^4+5.23\cdot10^{-5}t^3-0.00410\cdot t^2+0.1085t,\ t\in[0,115]$,

$\mu_1=1/30,\ \mu_2=1/100,\ \mu_3=1/20,\ p_{12}=0.25,\ p_{13}=0.25,\ p_{23}=0.15,\ N_1=10,\ N_2=5,\ N_3=3$

Figures 9 and 10 show the total number of casualties in each station and in queue for each model. The continuous line represents the number of casualties given by the model and the dashed lines represent the simulation results: average over the 200 replication and 95% confidence interval.

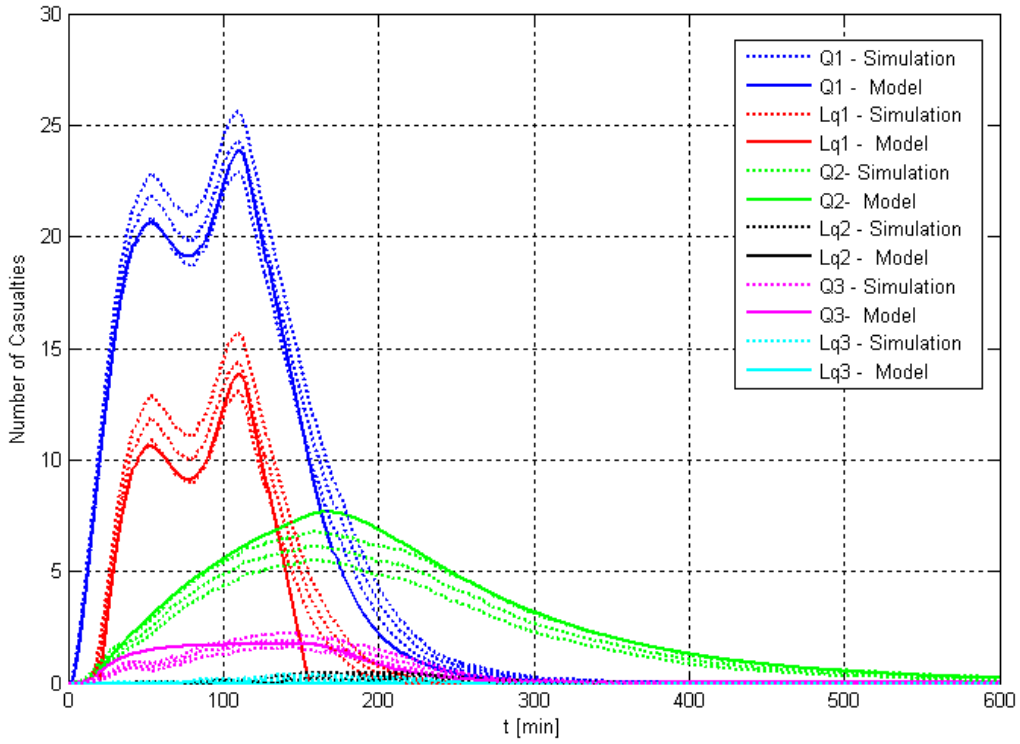


Figure 9 – Total number of Casualties in each station and Queue - Simulation vs. First Fluid model – Second Scenario

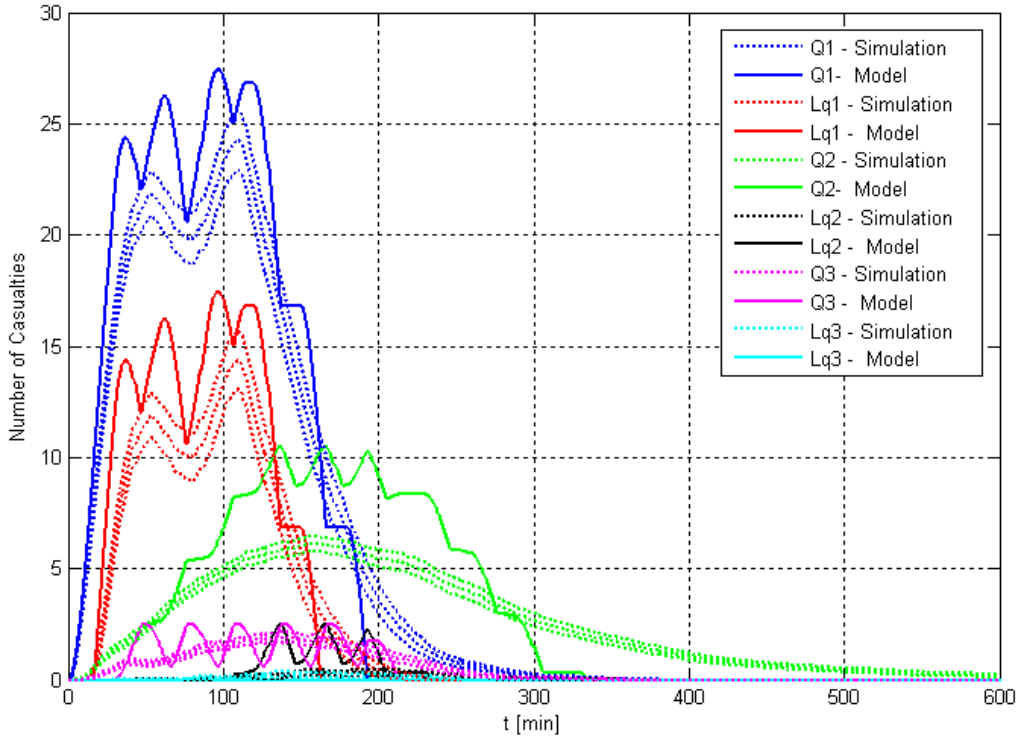


Figure 10 – Total number of Casualties in each station and Queue - Simulation vs. Second Fluid model – Second Scenario

While the first fluid model is quite accurate in predicting the two peaks in the number of casualties, the second fluid model fails to do so. In order to understand the reason for this, we go back to the cumulative arrivals and departures curves constructed by the second fluid model in Figure 11. The departure curves in the figure have a piecewise-constant shape. In trying to analyze this shape, we focus on the first station, as presented in Figure 12. In Phase 1 in Figure 12, there is no queue and therefore, the total number that departed the queue equals the total number that arrived. In addition, no casualty has yet departed the station:

$$Dq(t) = A(t),$$
$$Ds(t) = 0.$$

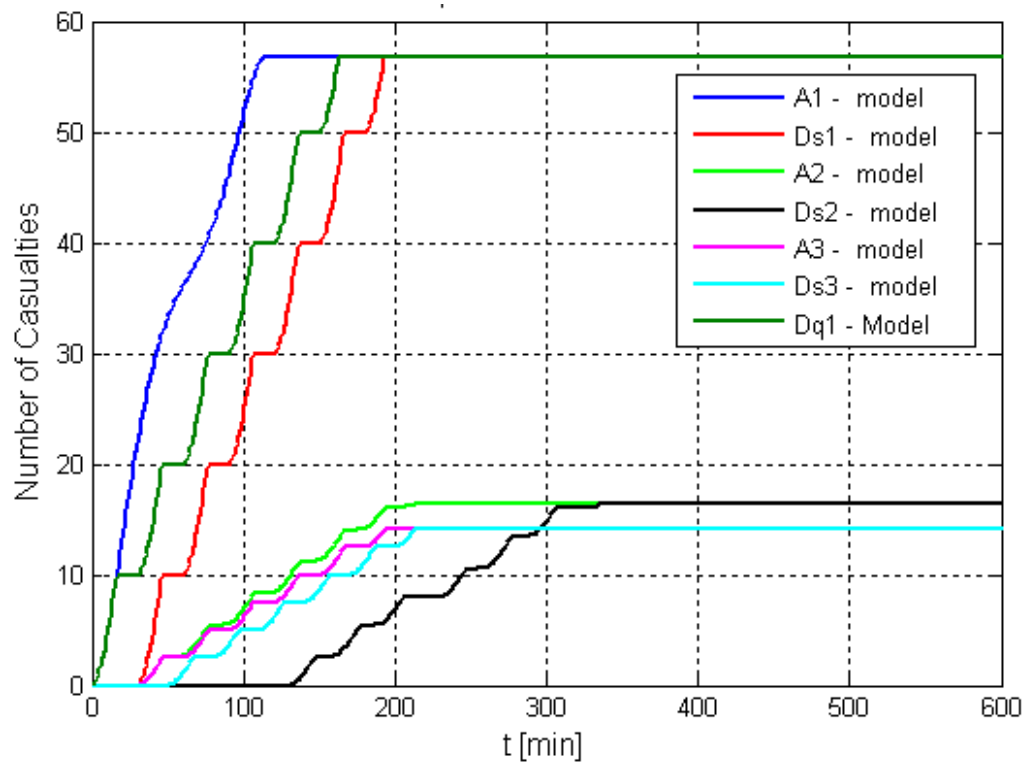


Figure 11 – Cumulative Arrivals and Departures - Second Fluid Model

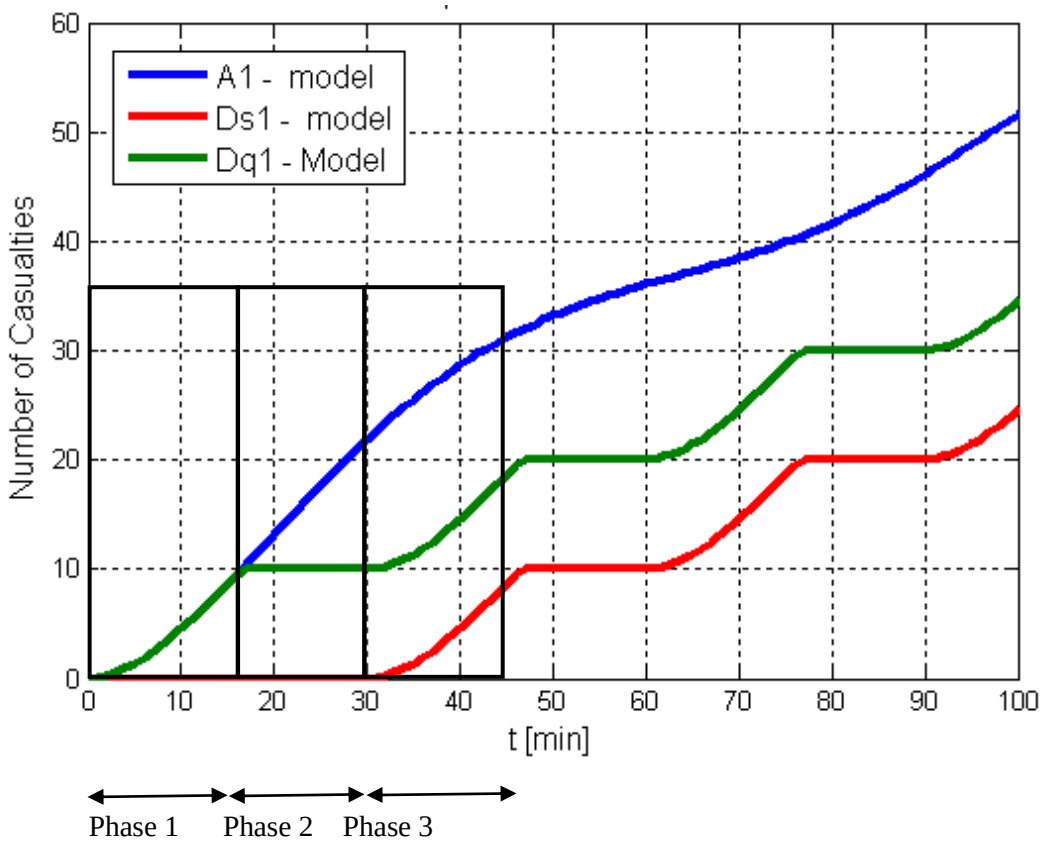


Figure 12 – Cumulative Arrivals and Departures – Station 1 – Second Fluid Model

In the start of Phase 2 in Figure 12, a queue begins to form but no casualty has yet departed the station, therefore

$$Dq(t) = N_1(t),$$
$$Ds(t) = 0.$$

At Phase 3 in Figure 12, one period of service time has ended ($t > 1/\mu$) and the total number of departures from the station equals the total departures from the queue one service time earlier, hence:

$$Dq(t) = Ds(t) + N_1(t),$$
$$Ds(t) = Dq(t - 1/\mu).$$

From the start of Phase 3 and on, the piecewise-constant pattern repeats itself in a 30 minutes ($1/\mu$) cycle.

The reason why the second fluid model fails in the two-surge scenario, in comparison with the quadratic arrival rate scenario, is that in the latter a queue began to form after one service time, and therefore there were no intervals in which the number of departures remained constant, e.g. Phase 2 in Figure 12 did not appear. The conclusion from this is that the second fluid model is not accurate in cases where the resources reach full capacity and a queue starts to form before one service time ends. When Ds is strictly increasing, one can deduce that the second fluid model will be accurate.

Following the reasons above, we chose to focus on the first fluid model when solving the optimization problem in the next chapter.

Sojourn Time from the models

We denote the sojourn time of a casualty that entered a station at time t by $\tau(t)$. Under the FCFS queue regime assumption, all casualties that arrived until time t, leave Station 1 until time $t+\tau(t)$, therefore $\tau(t)$ can be obtained as a solution to the following equation:

$$D_s(t + \tau(t)) = A(t), \qquad t \geq 0.$$
$$\tau(t) = D_s^{-1}(A(t)) - t, \qquad t \geq 0.$$

The sojourn time and waiting time can be derived by calculating the horizontal distance between the cumulative arrival curve and the cumulative departure curves from the station and from the queue respectively.

5 MINIMIZING MORTALITY

The main goal of hospitals during MCEs is to reduce mortality [5]. Hence, the objective function we chose is minimizing total mortality. Our model treats mortality as abandonments, which can occur both during waiting time and during treatment. The mortality rate differs from one station to another and the overall mortality is a summation of mortality in all stations throughout the MCE. We assume that the mortality rate is θ_1 and θ_2 for Stations 1 and Station 2, respectively. Based on the literature, we also assume that $\theta_1, \theta_2 \ll 1$. For example if θ_i=1/300, then the average time to death in station i is 5 hours (300 minutes). From this point on, we focus on two stations: Station 1 represents the Shock Rooms, Station 2 represents the Operation Rooms and the transition probability between the two is P_{12}.

The optimization problem we define for the two stations is:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \int_0^T [\theta_1 Q_1(t) + \theta_2 Q_2(t)]\, dt$$

S.T.

$$\dot{Q}_1(t) = \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t)) - \theta_1 \cdot Q_1(t), \qquad 0 \le t \le T$$

$$\dot{Q}_2(t) = p_{12} \cdot \mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t)) - \theta_2 \cdot Q_2(t), \qquad 0 \le t \le T$$

$$N_1(t) + N_2(t) \le N, \qquad 0 \le t \le T$$

$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0, \qquad 0 \le t \le T$$

$$Q_1(0) = 0, \quad Q_2(0) = 0.$$

The decision variables are $N_1(t)$ and $N_2(t)$ for all t in [0,T]. This is the number of surgeons that must be allocated to each station in every minute. The first two constraints are the flow constraints, according to which the number of casualties is calculated for every t. The third constraint is a resource constraint, ensuring that the maximal number of surgeons does not exceed N. The two last constraints define the variables to be non-negative and set the initial conditions. For simplicity, we assume that the system starts empty, e.g. there was early enough notice to clear the stations before casualties begin to arrive.

We transform the continuous problem into a discrete one by dividing the time interval into small subintervals, each of length 1 (we think in terms of time unit being 1 minute):

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \sum_{t=0}^{T} [\theta_1 Q_1(t) + \theta_2 Q_2(t)]$$

S.T.

$$Q_1(t+1) = Q_1(t) + \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t)) - \theta_1 \cdot Q_1(t) \qquad t = 0,\ldots T\text{-}1$$

$$Q_2(t+1) = Q_2(t) + p_{12} \cdot \mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t)) - \theta_2 \cdot Q_2(t) \qquad t = 0,\ldots T\text{-}1$$

$$N_1(t) + N_2(t) \le N \qquad t = 0,\ldots T\text{-}1$$

$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0 \qquad t = 0,\ldots T\text{-}1$$

$$Q_1(0) = 0,\ \ Q_2(0) = 0.$$

We thus implicitly assume that the allocation of surgeons can change every minute.

The above problem can also be rewritten as follows:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \sum_{t=0}^{T-1} [\theta_1 Q_1(t+1) + \theta_2 Q_2(t+1)]$$

S.T.

$$Q_1(t+1) = (1-\theta_1) Q_1(t) + \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t)) \qquad t = 0,\ldots T\text{-}1$$

$$Q_2(t+1) = (1-\theta_2) Q_2(t) + p_{12} \cdot \mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t)) \qquad t = 0,\ldots T\text{-}1$$

$$N_1(t) + N_2(t) \le N \qquad t = 0,\ldots T\text{-}1$$

$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0 \qquad t = 0,\ldots T\text{-}1$$

$$Q_1(0) = 0,\ \ Q_2(0) = 0.$$

Substituting the flow constraints into the objective function yields the following:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \sum_{t=0}^{T-1} \{(\theta_1[(1-\theta_1) Q_1(t) + \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t))] + \theta_2[(1-\theta_2) Q_2(t) + p_{12} \cdot \mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t))]\}$$

S.T.

$$Q_1(t+1) = (1-\theta_1) Q_1(t) + \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t)) \qquad t = 0,\ldots T\text{-}1$$

$$Q_2(t+1) = (1-\theta_2) Q_2(t) + p_{12} \cdot \mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t)) \qquad t = 0,\ldots T\text{-}1$$

$$N_1(t) + N_2(t) \le N \qquad t = 0,\ldots T\text{-}1$$

$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0 \qquad t = 0,\ldots T\text{-}1$$

$$Q_1(0) = 0,\ \ Q_2(0) = 0.$$

The following two sets of constraints can be added without changing the optimization problem. They ensure that the number of surgeons in each station will not exceed the number that is needed:

$N_1(t) \le Q_1(t) \qquad t = 0,...,T\text{-}1$
$N_2(t) \le Q_2(t) \qquad t = 0,...,T\text{-}1$

Those constraints reduce the number of possible solutions while the value of the objective function is not affected, since adding more surgeons than needed does not change the departure rate from the station.

Another reason for adding those constraints is to leave the redundancy of surgeons at the hands of the hospital and use only the necessary minimum as indicated by the solution.

Adding the above two sets of constraints will transform the problem into a linear one, by the following substitutions:

$[N_1(t) \wedge Q_1(t)] = N_1(t)$
$[N_2(t) \wedge Q_2(t)] = N_2(t)$

The linear programming problem we now wish to solve is:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \sum_{t=1}^{T-1}\{(\theta_1[(1-\theta_1)Q_1(t) + \lambda(t) - \mu_1 N_1(t)] + \theta_2[(1-\theta_2)Q_2(t) + p_{12}\cdot\mu_1 N_1(t) - \mu_2 N_2(t)]\}$$

S.T.

$Q_1(t+1) = (1-\theta_1)Q_1(t) + \lambda(t) - \mu_1 N_1(t) \qquad t = 0,...\ T\text{-}1$
$Q_2(t+1) = (1-\theta_2)Q_2(t) + p_{12}\cdot\mu_1 N_1(t) - \mu_2 N_2(t) \qquad t = 0,...\ T\text{-}1$
$N_1(t) \le Q_1(t) \qquad t = 0,...\ T\text{-}1$
$N_2(t) \le Q_2(t) \qquad t = 0,...\ T\text{-}1$
$N_1(t) + N_2(t) \le N \qquad t = 0,...\ T\text{-}1$
$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0 \qquad t = 0,...\ T\text{-}1$
$Q_1(0) = 0,\ \ Q_2(0) = 0.$

Few algebraic manipulation steps and omitting the constants yield the following:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \sum_{t=1}^{T-1}\{N_1(t)\mu_1[(1-\theta_1)^{T-t}-1-\ p_{12}[(1-\theta_2)^{T-t}-1]]+\ N_2(t)\mu_2[(1-\theta_2)^{T-t}-1]\}$$

S.T.

$$N_1(1) = 0$$

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

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

$$\vdots$$

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

$$N_2(1) = 0$$

$$\mu_2 N_2(1) - p_{12}\mu_1 N_1(1) + N_2(2) \le 0$$

$$(1-\theta_2)\mu_2 N_2(1) - (1-\theta_2)p_{12}\mu_1 N_1(1) + \mu_2 N_2(2) - p_{12}\mu_1 N_1(2) + N_2(3) \le 0$$

$$\vdots$$

$$(1-\theta_2)^{T-3}\mu_2 N_2(1) + (1-\theta_2)^{T-3}p_{12}\mu_1 N_1(1) + (1-\theta_2)^{T-4}\mu_2 N_2(2) - (1-\theta_2)^{T-4}p_{12}\mu_1 N_1(2) + \cdots$$

$$\cdots + \mu_2 N_2(T-2) - p_{12}\mu_1 N_1(T-2) + N_2(T\text{-}1) \le 0$$

$$N_1(t) + N_2(t) \le N \qquad t = 0,\ldots T\text{-}1$$

$$N_1(t),\ N_2(t) \ge 0 \qquad t = 0,\ldots T\text{-}1.$$

5.1 Optimal solution

We demonstrate the solution for the optimization problem in three scenarios. The solutions were generated by Mosek software installed on Malab (www.mosek.com), which globally solves constraint optimization problems.

First Scenario - Priority is given to Station 1:

The First Scenario is a quadratic arrival rate with the following parameters:

$$\lambda(t) = -1\cdot 10^{-5}t^2 + 0.0044t \qquad t \in [0,440]$$

$$\mu_1 = 1/30, \ \ \mu_2 = 1/100, \ \ \theta_1 = 1/180, \ \ \theta_2 = 1/300, \ \ p_{12} = 0.25, \ \ N = 10$$

The arrival rate $\lambda(t)$ refers to the number of immediate casualties who arrive to the ED per minute. Average treatment time is 30 minutes in Station 1 and 100 minutes in Station 2. Average time to death is 180 minutes in Station 1 and 300 minutes in Station 2. 25% of the casualties in Station 1 are transferred to Station 2 and there are 10 available surgeons.

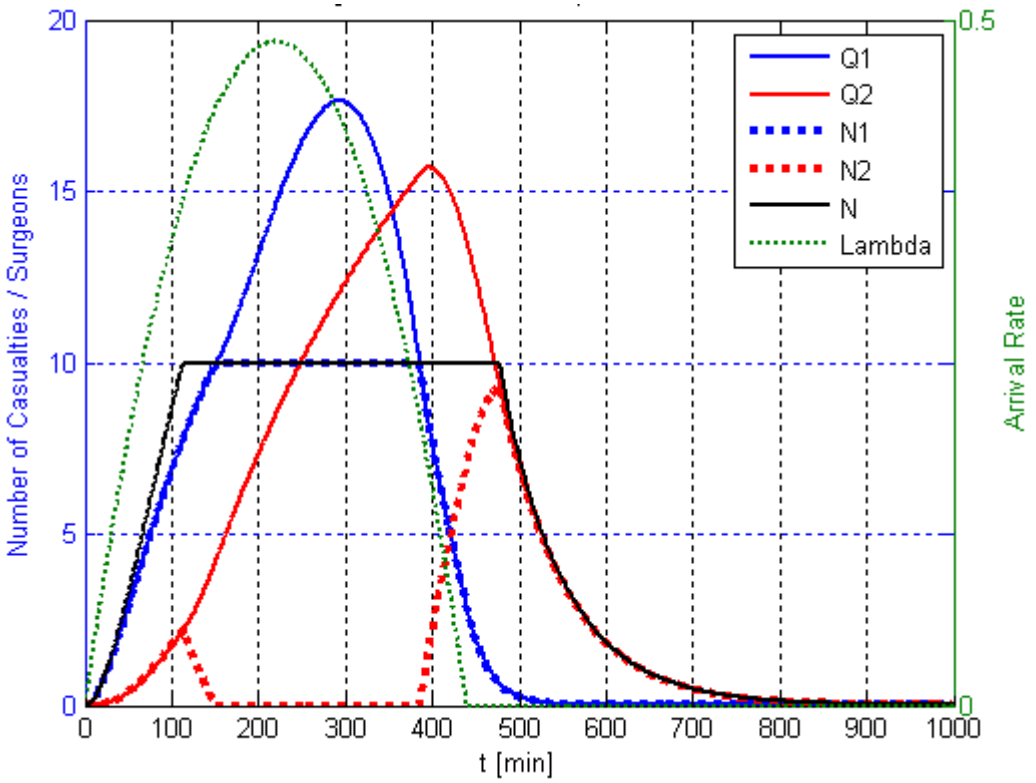


Figure 13 – Optimal Surgeons allocation – First Scenario

At the beginning of the event, until t=113, there is relatively a small number of casualties and, therefore, enough available surgeons for each station. The curves of $Q_1(\cdot)$ and $N_1(\cdot)$ and the curves of $Q_2(\cdot)$ and $N_2(\cdot)$ coincide. At t=113, the surgeons reach full capacity, Station 1 is prioritized before Station 2 and the surgeons are diverted to Station 1, until t=385, when the need for surgeons in Station 2 starts to decrease.

Second Scenario - Priority is given to Station 2:

The Second Scenario is a quadratic arrival rate with the following parameters:

$$\lambda(t) = -1 \cdot 10^{-5} t^2 + 0.0044t \qquad t \in [0,440]$$
$$\mu_1 = 1/30, \quad \mu_2 = 1/100, \quad \theta_1 = 1/180, \quad \theta_2 = 1/180, \quad p_{12} = 0.9, \quad N = 10$$

The arrival rate $\lambda(t)$, treatment rates and number of available surgeons are equal to the First Scenario. Average time to death is 180 minutes both in Station 1 and in Station 2. 90% of the casualties in Station 1 are transferred to Station 2.

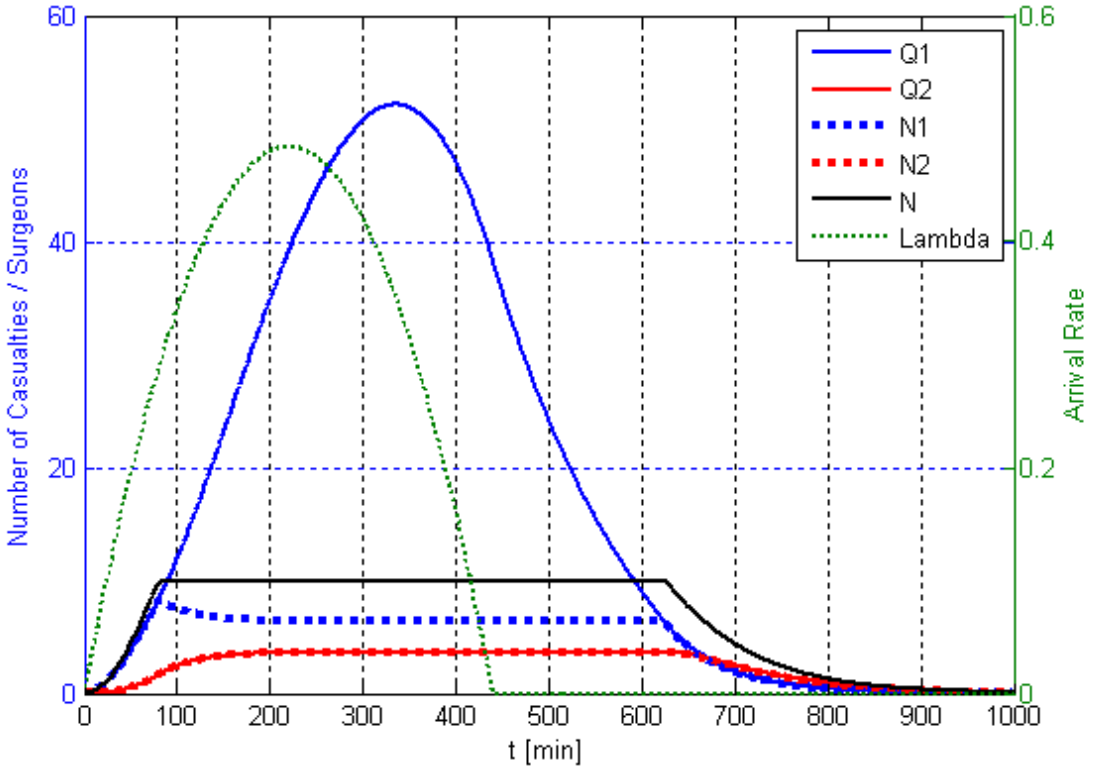


Figure 14 – Optimal Surgeons allocation – Second Scenario

In this scenario, when the surgeons reach full capacity at time t=83, Station 2 is prioritized over Station 1 and receives all the resources it requires. Station 1 receives all the resources that are left. In general, when Station 2 is prioritized, the arrival rate to Station 2 equals the exit rate from it:

$$p_{12} \cdot N_1(t) \cdot \mu_1 = N_2(t) \cdot (\mu_2 + \theta_2).$$

The system during those time interval is overloaded, the surgeons work at full capacity and therefore:

$$N_1(t) + N_2(t) = N \quad \Rightarrow \quad N_1(t) = N - N_2(t)$$

Combining the above equations yields the following constant ratio between the two stations:

$$N_1(t) = \frac{N(\mu_2 + \theta_2)}{\mu_2 + \theta_2 + p_{12} \cdot \mu_1}, \qquad N_2(t) = \frac{N \cdot p_{12} \cdot \mu_1}{\mu_2 + \theta_2 + p_{12} \cdot \mu_1}, \qquad 0 \le t \le T$$

If we generalize the number of surgeons needed in each station to R_1 for Station 1 and R_2 for Station 2, as is common in specific operations or with different resources, we get:

$$R_1 N_1(t) + R_2 N_2(t) = N \quad \Rightarrow \quad N_1(t) = \frac{N - R_2 N_2(t)}{R_1}$$

Combining the above equations yields the following:

$$N_1(t) = \frac{N(\mu_2 + \theta_2)}{R_1(\mu_2 + \theta_2) + R_2 \cdot p_{12} \cdot \mu_1}, \qquad N_2(t) = \frac{N \cdot p_{12} \cdot \mu_1}{R_1(\mu_2 + \theta_2) + R_2 \cdot p_{12} \cdot \mu_1}, \qquad 0 \le t \le T$$

Third Scenario - Priority is Switching:

The Third Scenario is a quadratic arrival rate with the following parameters:

$$\lambda(t) = -1 \cdot 10^{-5} t^2 + 0.0044t \qquad t \in [0,440]$$
$$\mu_1 = 1/30, \quad \mu_2 = 1/100, \quad \theta_1 = 1/180, \quad \theta_2 = 1/300, \quad p_{12} = 0.8, \quad N = 10$$

The arrival rate λ(t), treatment rates and number of available surgeons are equal to the First and Second Scenarios. Average treatment time is 30 minutes in Station 1 and 100 minutes in Station 2. Average time to death is 180 minutes in Station 1 and 300 minutes in Station 2. 80% of the casualties in Station 1 are transferred to Station 2 and there are 10 available surgeons.

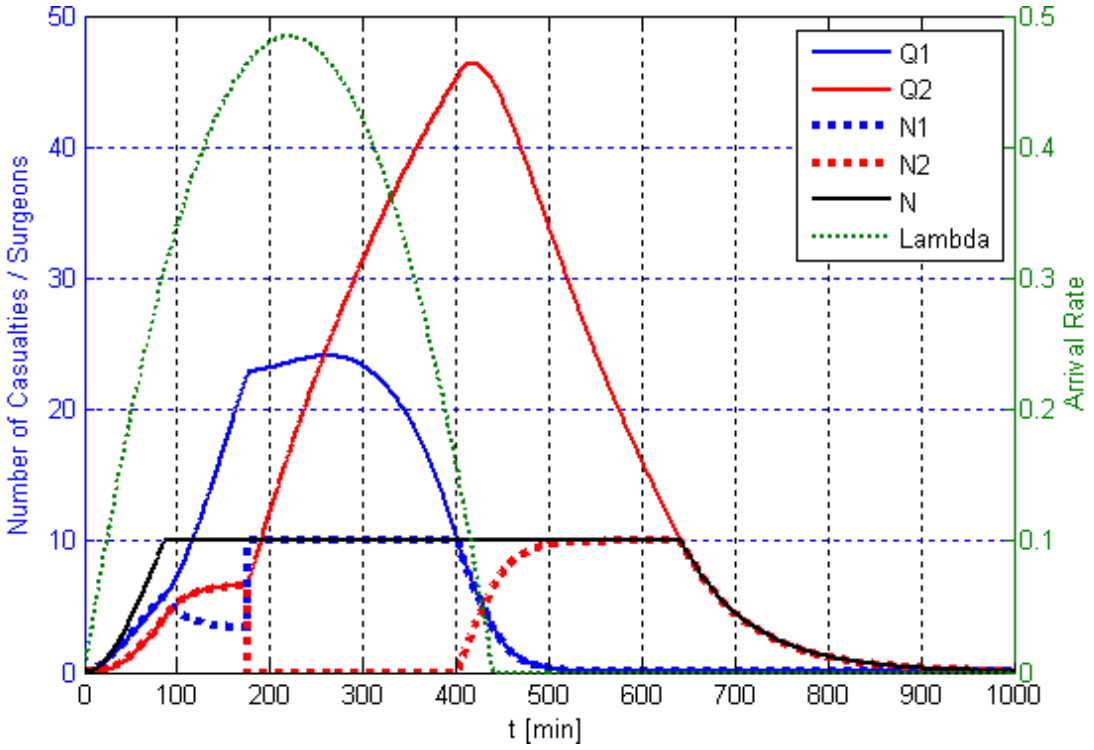


Figure 15 - Optimal Surgeons allocation – Third Scenario

In this example, the priority is changing throughout the event. At the beginning when t<86, there are enough surgeons in order to allocate to each station the number it needs. When 86 < t < 176, Station 2

is prioritized and receives all the surgeons it needs and Station 1 receives the rest. At t=176, and until t=400, Station 1 is prioritized and all surgeons are diverted to it. When t > 400, the total number of casualties is less than the number of available surgeons and, therefore, each station receives the number of surgeons it needs.

In general, the optimal solution gives priority to a station with higher service rate and higher mortality rate. As the transition probability from Station 1 to Station 2 increases, the priority for Station 2 increases since more casualties are being transferred to Station 2. As the required ratio between surgeons and casualties in Station 2 increases, the priority for Station 2 decreases, since it is preferable to allocate surgeons to Station 1 were only one surgeon per casualty is needed.

5.2 GREEDY PROBLEM

The greedy problem would be to solve an optimization problem for every t, namely for every time interval t, determining $N_1(t)$ and $N_2(t)$ in order to minimize the mortality in the next interval (t+1). The optimization problem, in this case would be:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \theta_1 Q_1(t+1) + \theta_2 Q_2(t+1) \qquad \forall t \in [0,\ T\text{-}1]$$

S.T.

$$Q_1(t+1) = Q_1(t) + \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t)) - \theta_1 \cdot Q_1(t)$$
$$Q_2(t+1) = Q_2(t) + p_{12} \cdot \mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t)) - \theta_2 \cdot Q_2(t)$$

$$N_1(t) + N_2(t) \le N$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0$$
$$Q_1(0) = 0, \quad Q_2(0) = 0.$$

Substituting the flow constraints in the objective function yields the following:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \theta_1[(1-\theta_1)Q_1(t) + \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t))] +$$
$$\theta_2[(1-\theta_2)Q_2(t) + p_{12} \cdot \mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t))] \qquad \forall t \in [0,\ T\text{-}1]$$

S.T.

$$N_1(t) + N_2(t) \le N$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0$$
$$Q_1(0) = 0, \quad Q_2(0) = 0.$$

As before, adding the two following constraints will prevent surgeons' redundancy, but will not affect the objective function and will transform the problem into a linear one:

$$N_1(t) \le Q_1(t),$$
$$N_2(t) \le Q_2(t).$$

$$\underset{N_1(\cdot),\, N_2(\cdot)}{\text{Min}} \quad \theta_1[(1-\theta_1)Q_1(t) + \lambda(t) - \mu_1 N_1(t)] + \theta_2[(1-\theta_2)Q_2(t) + p_{12} \cdot \mu_1 N_1(t) - \mu_2 N_2(t)] \qquad \forall t \in [0, T\text{-}1]$$

S.T.

$$N_1(t) \le Q_1(t)$$
$$N_2(t) \le Q_2(t)$$
$$N_1(t) + N_2(t) \le N$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0$$
$$Q_1(0) = 0,\quad Q_2(0) = 0.$$

At any time t, $Q_1(t)$, $Q_2(t)$ and $\lambda(t)$ are already known and do not affect the optimization problem. Therefore, they can be omitted from the objective function.

$$\underset{N_1(\cdot),\, N_2(\cdot)}{\text{Min}} \quad \theta_1[-\mu_1 N_1(t)] + \theta_2[p_{12} \cdot \mu_1 N_1(t) - \mu_2 N_2(t)] \qquad \forall t \in [0, T\text{-}1]$$

S.T.

$$N_1(t) \le Q_1(t)$$
$$N_2(t) \le Q_2(t)$$
$$N_1(t) + N_2(t) \le N$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0$$
$$Q_1(0) = 0,\quad Q_2(0) = 0.$$

The problem can also be written as a maximization problem as follows:

$$\underset{N_1(\cdot),\, N_2(\cdot)}{\text{Max}} \quad N_1(t) \cdot \mu_1[\theta_1 - p_{12}\theta_2] + N_2(t) \cdot \mu_2 \cdot \theta_2 \qquad \forall t \in [0, T\text{-}1]$$

S.T.

$$N_1(t) \le Q_1(t)$$
$$N_2(t) \le Q_2(t)$$
$$N_1(t) + N_2(t) \le N$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0$$
$$Q_1(0) = 0,\quad Q_2(0) = 0.$$

The priority for allocating the resources when the system is overloaded is determined according to the continuous Knapsack Problem. The Knapsack Problem is an optimization problem, where given a set of items, each with a weight and a value, determines the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible [51]. In the continuous Knapsack Problem, the number of items can be a fraction (not

necessary an integer). The solution for the continuous Knapsack problem is given by the following rule:

If $\mu_1(\theta_1 - p_{12}\theta_2) > \mu_2\theta_2$ the priority should be given for Station 1. In other words, allocate all the resources Station 1 needs, and if there are still available resources left, allocate them to Station 2.

$N_1(t) = \min(N, Q_1(t))$, $N_2(t) = N\text{-}N_1(t)$.

Otherwise, the priority should be given to Station 2 and the allocation should be:

$N_2(t) = \min(N, Q_2(t))$, $N_1(t) = N\text{-}N_2(t)$.

If we generalize the constraint $N_1(t) + N_2(t) \le N$ to $R_1N_1(t) + R_2N_2(t) \le N$, that is we allow the ratio between the number of resources and number of casualties to be different then 1:1, we get different rules regarding the priorities:

If $\frac{\mu_1(\theta_1 - p_{12}\theta_2)}{R_1} > \frac{\mu_2\theta_2}{R_2}$, the priority should be given to Station 1. In other words, allocate all the resources Station 1 needs, and if there are still available resources left, allocate them to Station 2.

$N_1(t) = \min(N, Q_1(t))$, $N_2(t) = [N\text{-}R_1N_1(t)]/ R_2$.

Otherwise, priority should be given to Station 2 and the allocation should be:

$N_2(t) = \min(N, Q_2(t))$, $N_1(t) = [N\text{-}R_2N_2(t)]/ R_1$.

We now demonstrate the greedy solution to the three scenarios discussed in the previous chapter.

First Scenario:

$$\lambda(t) = -1\cdot 10^{-5}t^2 + 0.0044t \qquad t \in [0,440]$$
$$\mu_1 = 1/30,\ \ \mu_2 = 1/100,\ \ \theta_1 = 1/180,\ \ \theta_2 = 1/300,\ \ p_{12} = 0.25,\ \ N = 10$$

The arrival rate $\lambda(t)$ refers to the number of immediate casualties who arrive to the ED per minute. Average treatment time is 30 minutes in Station 1 and 100 minutes in Station 2. Average time to death is 180 minutes in Station 1 and 300 minutes in Station 2. 25% of the casualties in Station 1 are transferred to Station 2 and there are 10 available surgeons.

The greedy solution is the optimal one - the priority is given to Station 1 since $\mu_1(\theta_1 - p_{12}\theta_2) > \mu_2\theta_2$ as shown in Figure 16:

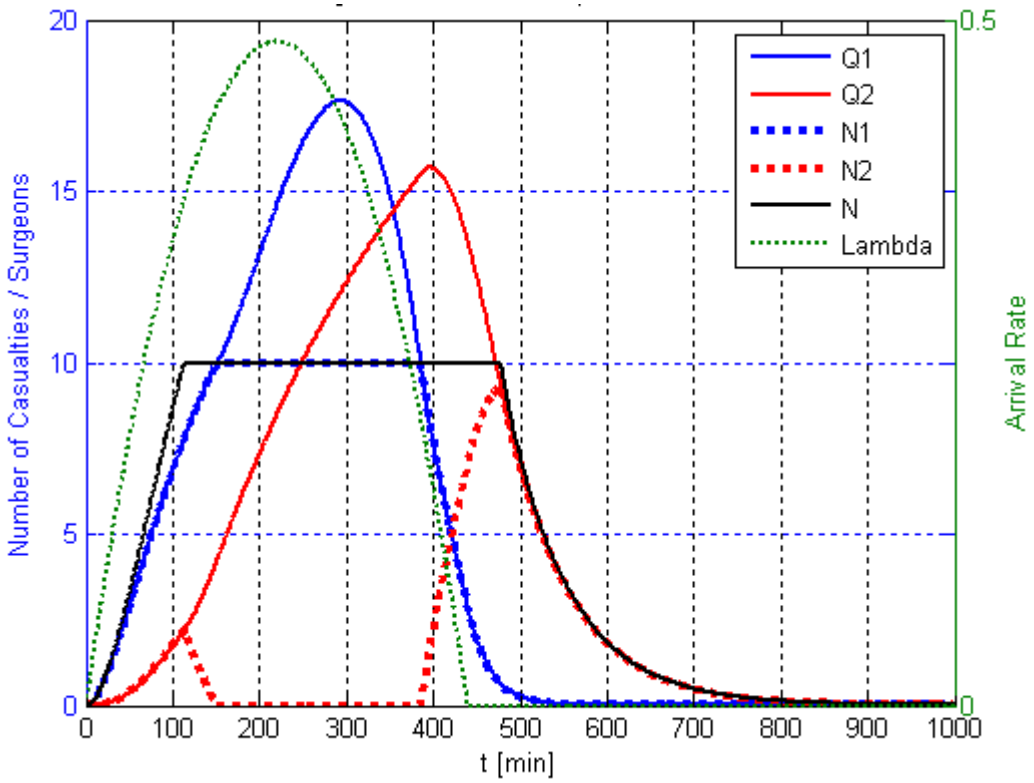


Figure 16 - Optimal Surgeons allocation – Greedy solution - First Scenario

Second Scenario:

$\lambda(t) = -1 \cdot 10^{-5} t^2 + 0.0044t \qquad t \in [0,440]$

$\mu_1 = 1/30, \quad \mu_2 = 1/100, \quad \theta_1 = 1/180, \quad \theta_2 = 1/180, \quad p_{12} = 0.9, \quad N = 10$

Average treatment time is 30 minutes in Station 1 and 100 minutes in the Station 2. Average time to death is 180 minutes both in Station 1 and in Station 2. 90% of the casualties in Station 1 are transferred to Station 2 and there are 10 available surgeons.

The greedy solution is the optimal one - the priority is given to Station 2 since $\mu_1(\theta_1 - p_{12}\theta_2) < \mu_2\theta_2$ as shown in Figure 17:

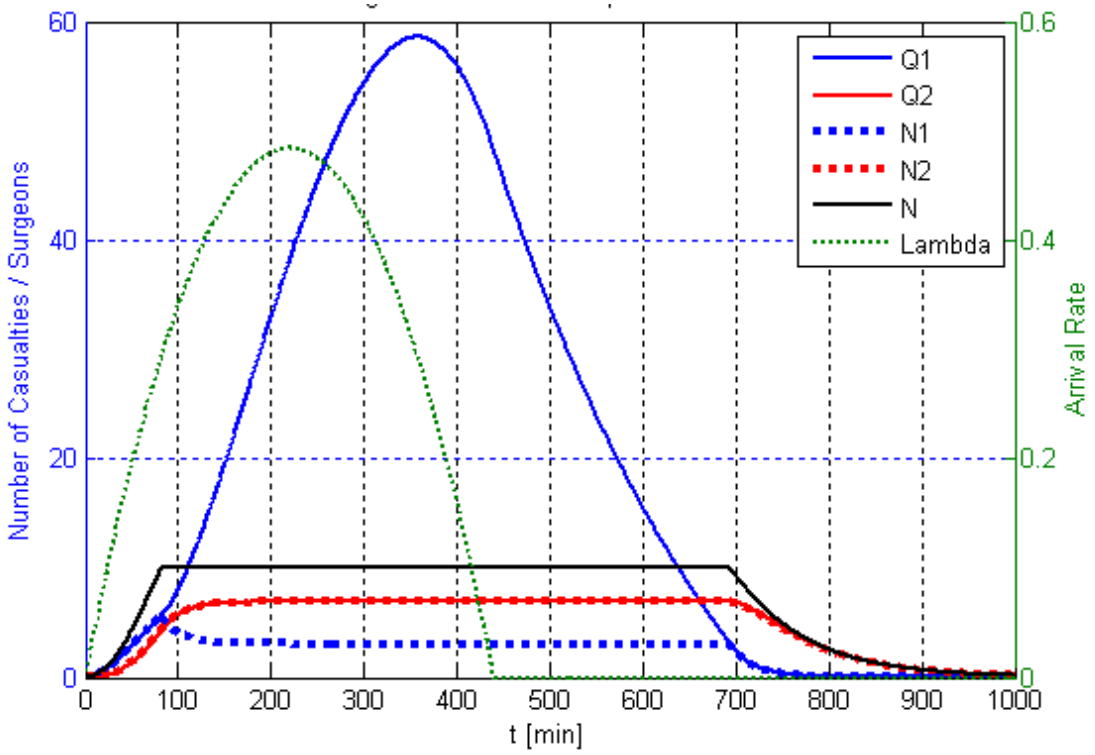


Figure 17 - Optimal Surgeons allocation – Greedy solution - Second Scenario

Third Scenario:

$\lambda(t) = -1 \cdot 10^{-5} t^2 + 0.0044t \qquad t \in [0,440]$

$\mu_1 = 1/30, \quad \mu_2 = 1/100, \quad \theta_1 = 1/180, \quad \theta_2 = 1/300, \quad p_{12} = 0.8, \quad N = 10$

Average treatment time is 30 minutes in Station 1 and 100 minutes in Station 2. Average time to death is 180 minutes in Station 1 and 300 minutes in Station 2. 80% of the casualties in Station 1 are transferred to Station 2 and there are 10 available surgeons.

The greedy solution here is different from the optimal solution. The Priority is given to Station 1 throughout the entire event, since $\mu_1(\theta_1 - p_{12}\theta_2) > \mu_2\theta_2$. The priority does not change as it did in the optimal solution. This is shown in Figure 18:

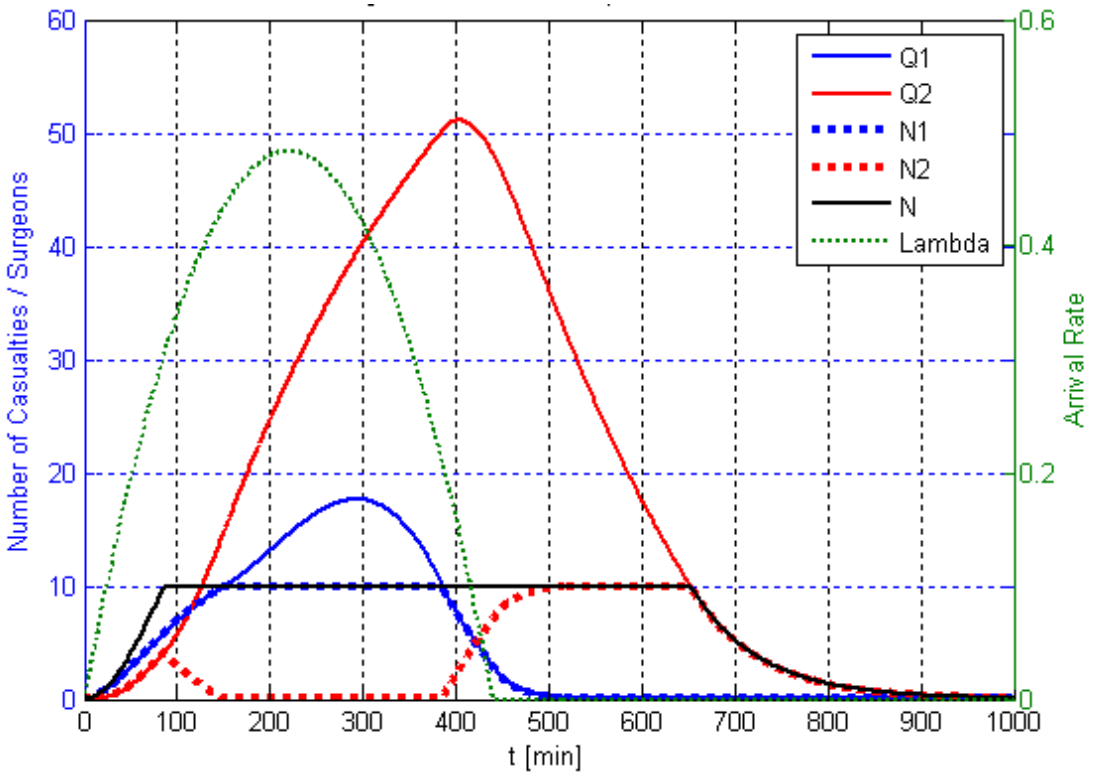


Figure 18 - Optimal Surgeons allocation – Greedy solution - Third Scenario

5.3 Comparison between Optimal and greedy solutions

A comparison between the optimal solution and the greedy solution for the Third Scenario in Chapters 5.1 and 5.2 is presented in Figures 18-20. Figure 18 presents the total number of casualties in each station in each solution. The total number of casualties is similar in both solutions, although in the greedy solution the total number is slightly higher. Despite this, the distribution between the two stations is different: the total number in the first station is higher in the optimal solution than in the greedy solution, but in the second station, the total number is higher in the greedy solution.

Figure 19 presents the mortality rate in each solution. When t <331, the mortality rate is higher in the optimal solution. But when t>331, the mortality rate in the greedy solution is higher.

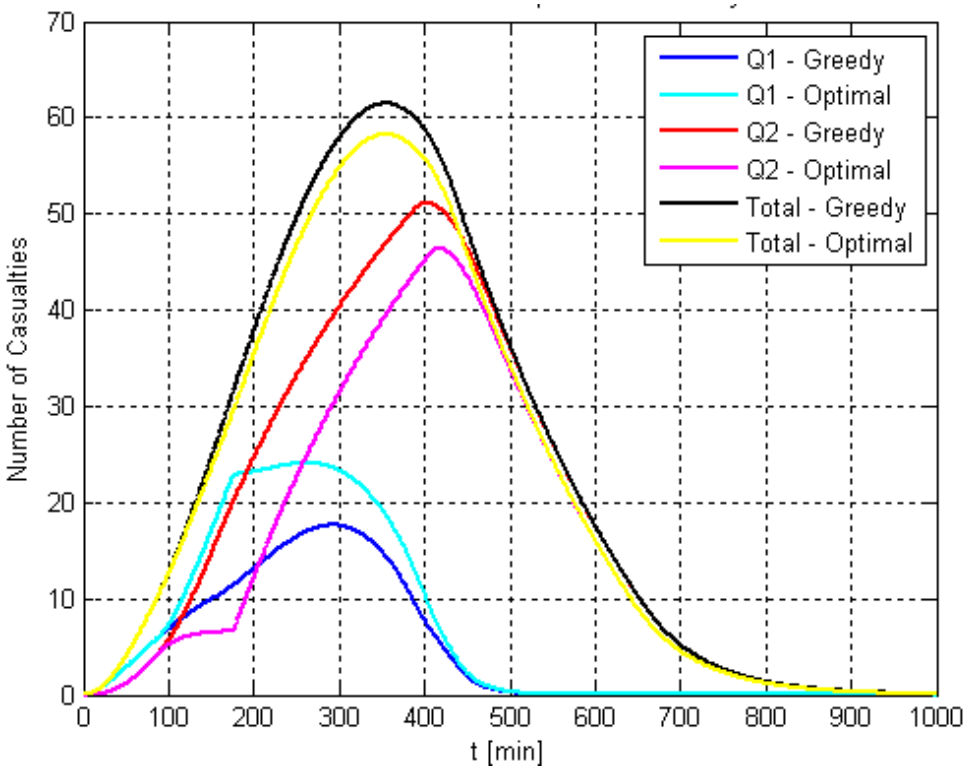


Figure 19 – Total Number of Casualties – Optimal vs. Greedy Solutions

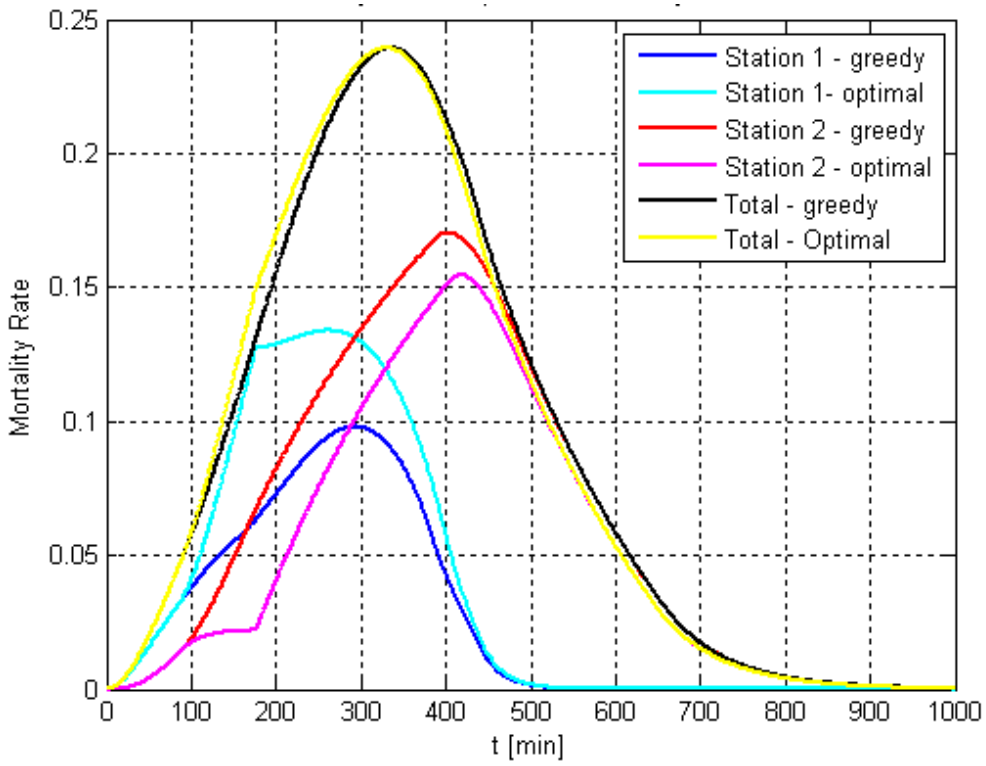


Figure 20 – Mortality Rate – Optimal vs. Greedy Solutions

Figure 20 presents the cumulative mortality in each solution. Until t=729, the total mortality in the optimal solution is higher than in the greedy solution. Only when t>729, the total mortality in the optimal solution improves.

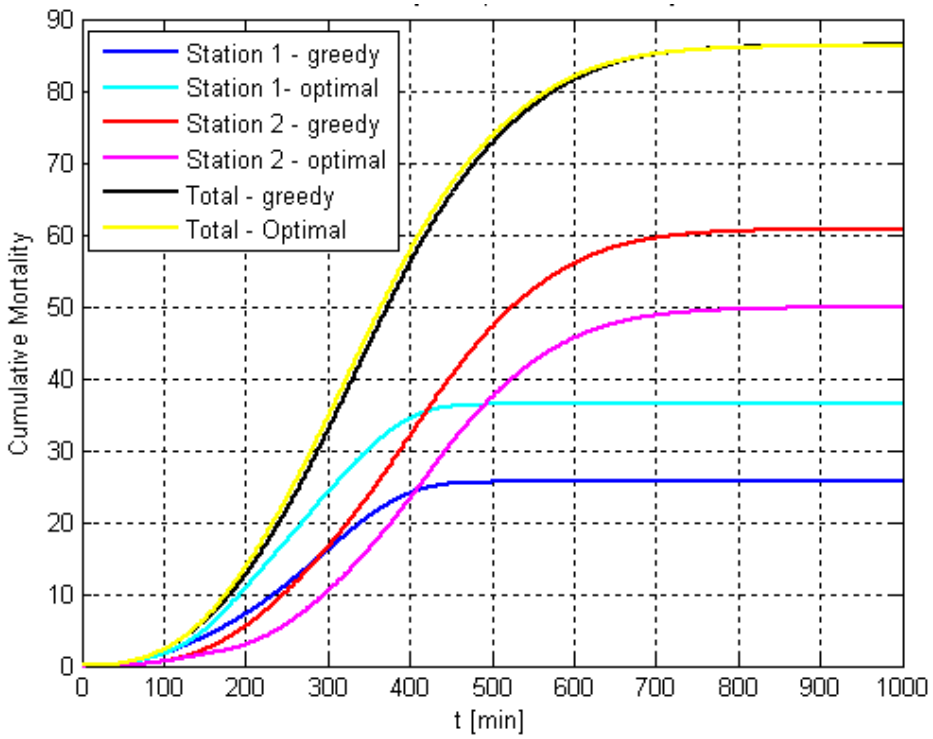


Figure 21 – Cumulative Mortality – Optimal vs. Greedy Solutions

5.3.1 Sensitivity Analysis

A sensitivity analysis of the Greedy solution vs. optimal solution for 24 scenarios is summarized in Table 2. The parameters used in the experiments were chosen after consulting with experts in MCEs and Emergency Medicine and they are adequate for an Emergency Departments during MCE. Station 1 (Shock Rooms) treats casualties at rate of 2 casualties per hour. The treatments rate in Station 2 (Operation Rooms) is lower and is on average 0.6 casualties per hour. The mortality rate in both stations was determined by the average time it takes a casualty until death. In Station 1, an average of 5 hours is taken per casualty. Since the mortality is modeled by an exponential distribution, the probability that a casualty will survive more than 5 hours is ~0.333, and the probability of survival more than 10 hours is ~0.13. In Station 2, averages of 3.33 hours and 1.667 hours are taken. The transition probability from Station 1 to Station 2 is 0.25. We used eight scenarios for the arrival rate (quadratic, linear, constant and two-surges) (Table 2c) and a range of 10-20 available surgeons. The mortality percentages in Tables 2a and 2b are calculated from the total number of casualties that arrive to the ED (out of which 15% are immediate casualties).

The main insight from the results is that the closer the mortality rates of the two stations to each other, the smaller the difference between the optimal and greedy solutions (Table 2a vs. Table 2b). This insight strengthens with the fact that there is no difference between the greedy and optimal solutions when the mortality rates are equal. The maximal difference between the greedy and optimal solution is ~6%. Another insight which can be derived from the results is that the higher the arrival rate and the fewer the number of available surgeons (N), the larger the difference between the two solutions.

No.	Scenario	Mortality - Optimal	Mortality - Greedy	% Diff
1	Arrival Rate1, N=10	39.25 (4.15%)	39.26 (4.15%)	0.01%
2	Arrival Rate1, N=15	27.35 (2.89%)	27.35 (2.89%)	0.01%
3	Arrival Rate2, N=20	45.07 (3.17%)	45.07 (3.17%)	0.01%
4	Arrival Rate3, N=10	27.49 (4.12%)	27.49 (4.12%)	0.02%
5	Arrival Rate4, N=10	62.96 (5.90%)	62.97 (5.90%)	0.01%
6	Arrival Rate4, N=15	46.47 (4.36%)	46.47 (4.36%)	0.01%
7	Arrival Rate5, N=10	23.37 (3.98%)	23.38 (3.98%)	0.02%
8	Arrival Rate5, N=15	17.86 (3.04%)	17.86 (3.04%)	0.03%
9	Arrival Rate6, N=15	57.13 (4.90%)	57.14 (4.90%)	0.01%
10	Arrival Rate7, N=10	13.41 (3.53%)	13.41 (3.53%)	0.04%
11	Arrival Rate7, N=15	10.04 (2.64%)	10.05 (2.64%)	0.06%
12	Arrival Rate8, N=15	26.80 (4.67%)	26.81 (4.68%)	0.01%

Table 2a – Sensitivity Analysis of Greedy vs. Optimal solution

(μ_1=1/30, μ_2=1/100, θ_1=1/300, θ_2=1/200, P_{12}=0.25)

No.	Scenario	Mortality - Optimal	Mortality - Greedy	Diff
1	Arrival Rate1, N=10	43.65 (4.61%)	46.17 (4.88%)	5.77%
2	Arrival Rate1, N=15	31.69 (3.35%)	31.98 (3.38%)	0.92%
3	Arrival Rate2, N=20	51.84 (3.65%)	53.28 (3.75%)	2.78%
4	Arrival Rate3, N=10	30.96 (4.64%)	31.84 (4.78%)	2.86%
5	Arrival Rate4, N=10	67.28 (6.31%)	69.84 (6.55%)	3.81%
6	Arrival Rate4, N=15	51.89 (4.86%)	53.46 (5.01%)	3.03%
7	Arrival Rate5, N=10	26.44 (4.51%)	26.84 (4.58%)	1.52%
8	Arrival Rate5, N=15	21.03 (3.58%)	21.04 (3.59%)	0.08%
9	Arrival Rate6, N=15	62.68 (5.37%)	64.37 (5.52%)	2.69%
10	Arrival Rate7, N=10	15.53 (4.09%)	15.59 (4.10%)	0.37%
11	Arrival Rate7, N=15	12.08 (3.18%)	12.08 (3.18%)	0.01%
12	Arrival Rate8, N=15	29.70 (5.18%)	30.07 (5.24%)	1.26%

Table 2b – Sensitivity Analysis of Greedy vs. Optimal solution

(μ_1=1/30, μ_2=1/100, θ_1=1/300, θ_2=1/100, P_{12}=0.25)

Arrival Rate	$\lambda(t)$
Arrival Rate1	$\lambda(t) = -1\cdot10^{-5}t^2 + 0.0044t, \quad 0 \le t \le 440$
Arrival Rate2	$\lambda(t) = -1.5\cdot10^{-5}t^2 + 0.0066t, \quad 0 \le t \le 440$
Arrival Rate3	$\lambda(t) = 0.5, \quad 0 \le t \le 200$
Arrival Rate4	$\lambda(t) = 0.8, \quad 0 \le t \le 200$
Arrival Rate5	$\lambda(t) = 0.0044t, \quad 0 \le t \le 200$
Arrival Rate6	$\lambda(t) = 0.0088t, \quad 0 \le t \le 200$
Arrival Rate7	$\lambda(t) = -2.16\cdot10^{-7}t^4 + 5.23t^3 - 0.0041t^2 + 0.1085t, \quad 0 \le t \le 115$
Arrival Rate8	$\lambda(t) = -3.24\cdot10^{-7}t^4 + 7.85t^3 - 0.0062t^2 + 0.163t, \quad 0 \le t \le 115$

Table 2c – Arrival Rates (per minute)

5.4 PROBLEM ANALYSIS

When the decision maker encounters an MCE, and the surgeons cannot overcome the load, the options are to prioritize surgeons to Station 1 or to Station 2. Another conflict is if to change priorities (e.g., first prioritize Station 1 and at some time point prioritize Station 2) or keep them constant until the event concludes. In this section, we present Theorem 1 that characterizes optimal priority settings for various scenarios, and we then analyze each scenario.

Theorem 1: The optimal policies are listed in the following Table 3, according to the relationship between $\mu_1(1\text{-}p_{12})$ and μ_2:

Conditions	$\theta_1 = \theta_2$	$\theta_1 > \theta_2$	$\theta_1 < \theta_2$
$\mu_1(1\text{-}p_{12}) = \mu_2$	Station 1 or Station 2 (Case1)	Station 1 (Case 4)	Station 2 (Case 7)
$\mu_1(1\text{-}p_{12}) > \mu_2$	Station 1 (Case 2)	Station 1 (Case 5)	Set priority to Station 1 and switch to Station 2 at *t* (Case 8)
$\mu_1(1\text{-}p_{12}) < \mu_2$	Station 2 (Case 3)	Set priority to Station 2 and switch to Station 1 at *t* (Case 6)	Station 2 (Case 9)

Table 3: Optimal Surgeons allocations

Specifically, the entries in the table indicate which station gets a higher priority: this station gets all the surgeons it needs out of those available, the other station gets the rest of the surgeons, if any.

Proof: The proof involves three steps. In the first step we prove, by induction, that the greedy problem yields optimal results when mortality rates at the two stations are equal and develop analytical conditions for priority policy setting (Cases 1- 3 of Theorem 1). In the second step, we prove by induction that, when the mortality rates are equal and the priority is giving to a specific station, this priority will also be given to it when the mortality rate in that station is higher than in the other (Cases 4,5,7 and 9 of Theorem 1). In the third step we show, empirically, that a switch in priority may occur during the event (Cases 6 and 8 of Theorem 1). As the mathematics involved is rather tedious, we have placed it, when possible, in Appendix II and III while keeping here only the main components.

Step 1: We prove by a Mathematical Induction that the optimal solution of the greedy problem is identical to the optimal solution of the original problem for equal mortality rates $\theta = \theta_1 = \theta_2$. The proof is detailed in Appendix II and it facilitates a static priority rule for optimal surgeons' allocation. The Second Scenario in Chapters 5.1 and 5.2 illustrates Case 3.

Step 2: We expand our proof for equal mortality rates to the cases where the mortality rates are different but still the Greedy solution is the optimal one. Cases 4 and 5 - if no station or Station 1 gets priority when $\theta_1=\theta_2$, then when $\theta_1 > \theta_2$, Station 1 will still get priority, since higher mortality rate in one station strengthens its priority. The same holds for cases 7 and 9 - if no station or Station 2 gets priority when $\theta_1=\theta_2$, then when $\theta_1 < \theta_2$ Station 2 will still get priority. The proof is detailed in Appendix III. The First Scenario in chapters 5.1 and 5.2 illustrates Case 5.

Step 3: We show, empirically, that for Cases 6 and 8 there may be a switch in priority. The priority before the switch is determined according to the priority for the equal mortality rates and the priority after the switch is given to the other station. The Third Scenario in chapters 5.1 and 5.2 illustrates Case 6. The 24 scenarios described in Chapter 5.3.1 are also included under those cases.

5.5 MINIMAL TIME WINDOW FOR RESOURCE ALLOCATION

Changing the allocation of surgeons every minute is not feasible. We assume that the allocation can be changed only every S minutes (usually S=30 or 60 minutes). Then, additional constraints must be added to our original problem. The optimal allocation for each period cannot be found by solving an

optimization problem for every minute, as was done before. The entire time interval must be taken into account while the objective function is minimizing the overall mortality.

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \sum_{t=0}^{T}[\theta_1 Q_1(t) + \theta_2 Q_2(t)]$$

S.T.

$$Q_1(t+1) = (1-\theta_1)Q_1(t) + \lambda_1(t) - \mu_1(Q_1(t) \wedge N_1(t))$$
$$Q_2(t+1) = (1-\theta_2)Q_2(t) + p_{12}\mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t))$$

$$N_1(t) + R_2 N_2(t) \le N \qquad t=1,\ldots T$$
$$N_1(i) = N_1(i+1) = \ldots = N_1(i+S-1) \qquad i = 1, S+1, 2S+1 \ldots \left\lfloor \frac{T}{S} \right\rfloor S+1$$
$$N_2(i) = N_2(i+1) = \ldots = N_2(i+S-1) \qquad i = 1, S+1, 2S+1 \ldots \left\lfloor \frac{T}{S} \right\rfloor S+1$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0.$$

When allocations can be changed every minute, additional constraints preventing surgeons' redundancy were added and transform the problem into a linear one. In present case, such constraints cannot be added since the allocation remains constant for S minutes. The optimal solution can include intervals in which there is redundancy in surgeons in order to prevent shortage in other parts of the interval.

Auxiliary sets of variables $Z_1(t)$ and $Z_2(t)$ for every t are defined by the following:

$$Z_1(t) = Q_1(t) \wedge N_1(t) \qquad \forall t \in [0,T]$$
$$Z_2(t) = Q_2(t) \wedge N_2(t) \qquad \forall t \in [0,T].$$

This can be expressed in terms of the four sets of linear constraints:

$$Z_1(t) \le N_1(t)$$
$$Z_1(t) \le Q_1(t)$$
$$Z_2(t) \le N_2(t)$$
$$Z_2(t) \le Q_2(t).$$

In addition, $Z_1(t)$ and $Z_2(t)$ must appear with a negative sigh in the objective function, in order to assure that the solution sets $Z_1(t)$ and $Z_2(t)$ the maximal value it can. $Z_2(t)$ always appears with a negative sign, but $Z_1(t)$ appears with a negative sign only if the following condition holds:

$$-\theta_1\mu_1 + \theta_2 p_{12}\mu_1 < 0 \quad \Rightarrow \quad \theta_2 p_{12} < \theta_1$$

Substituting the auxiliary variables in the problem yields the following:

$$\underset{N_1(\bullet),\, N_2(\bullet)}{\text{Min}} \quad \sum_{t=0}^{T-1}[\theta_1 Q_1(t+1) + \theta_2 Q_2(t+1)]$$

S.T.

$$Q_1(t+1) = Q_1(t) + \lambda_1(t) - \mu_1 Z_1(t) - \theta_1 \cdot Q_1(t)$$
$$Q_2(t+1) = Q_2(t) + p_{12} \cdot \mu_1 Z_1(t) - \mu_2 \cdot Z_2(t) - \theta_2 \cdot Q_2(t)$$

$$N_1(t) + R_2 N_2(t) \le N$$
$$N_1(i) = N_1(i+1) = \ldots = N_1(i+S-1) \qquad i = 1, S+1, 2S+1 \ldots \left\lfloor \frac{T}{S} \right\rfloor S+1$$
$$N_2(i) = N_2(i+1) = \ldots = N_2(i+S-1) \qquad i = 1, S+1, 2S+1 \ldots \left\lfloor \frac{T}{S} \right\rfloor S+1$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0$$
$$Q_1(0) = Q_2(0) = Q_3(0) = 0.$$

We illustrate the three scenarios described in Chapters 5.1 and 5.2:

First Scenario: Priority is given to Station 1:

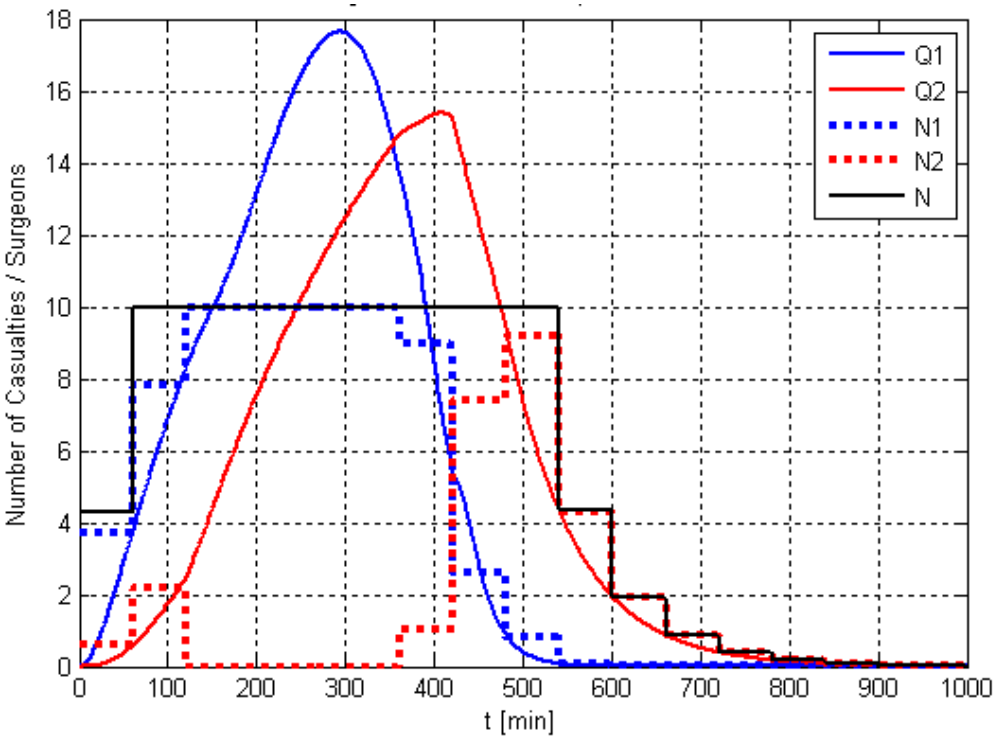


Figure 22 – Optimal Surgeons allocation – Minimal Time Window – First Scenario

Second Scenario: priority is given to Station 2:

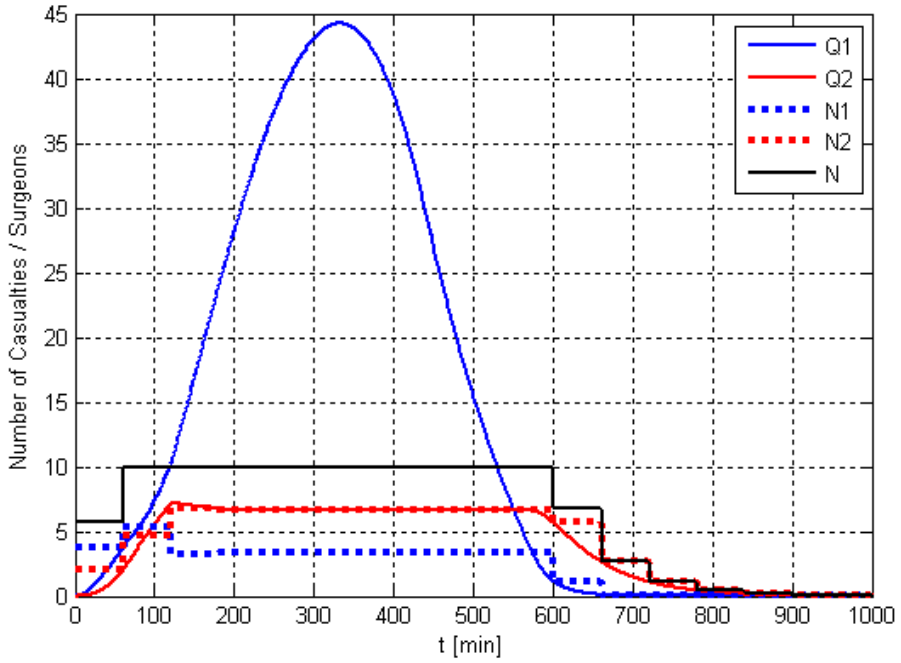


Figure 23 – Optimal Surgeons allocation – Minimal Time Window – Second Scenario

Third Scenario: priority is Switching:

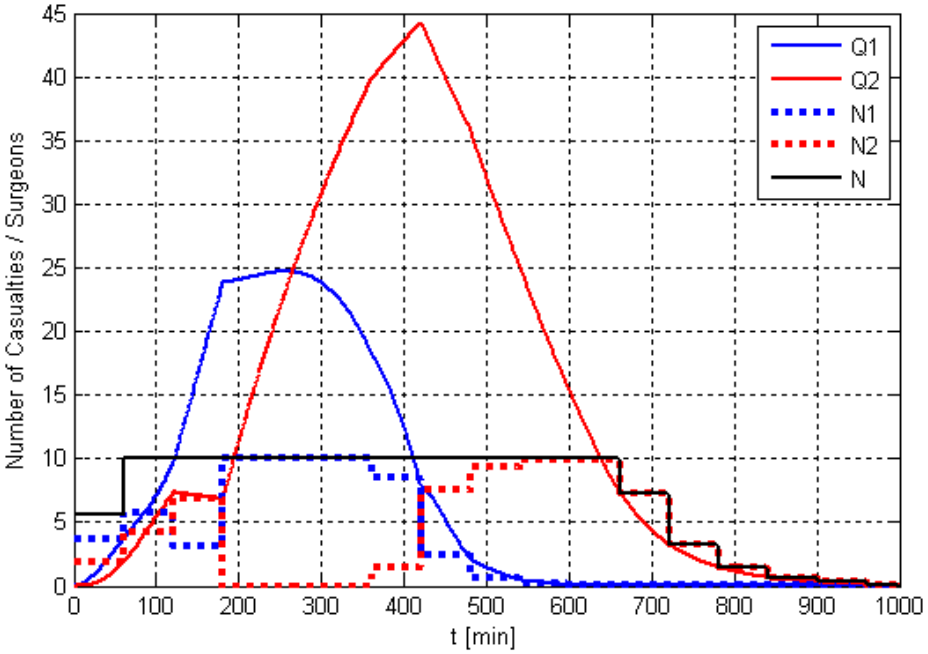


Figure 24 – Optimal Surgeons allocation – Minimal Time Window – Third Scenario

6 SUMMARY AND CONCLUSIONS

The model we propose predicts the number of casualties in a hospital's ED during an MCE. We formulate optimization problems with the objective of minimizing the mortality of casualties and then solve the problems by combining theory with numerical analysis. Our solution approach finds the dynamic allocation of surgeons between the shock rooms and operating rooms during an MCE.

We formulated a greedy counterpart for the original problem and found the conditions under which its solution solves also the original problem. We defined a general approach to predict the structure of the optimal solution of the original problem. Our model is simple enough, yet able to describe a broad range of different MCE scenarios. As such, it can be used to help in preparing for, and managing an MCE.

7 FEW DIRECTION FOR FUTURE RESEARCH

There are several directions for future research.

One type of research can deal with analytical extensions such as finding the optimal t in which resource allocation priority changes between the two stations and bounds to the performance of the greedy algorithm compared to the optimal policy. Our insights can be analyzed also with respect to a change in the structure of parameters, for example if the mortality rates (or alternatively, the survival chances) of casualties increase (decrease) while they are waiting for treatment.

Another possible outlet is to use the developed approach to account for other types of MCEs, such as non-conventional MCEs (biological, chemical, nuclear and radiation), where there are different medical processes and resources.

Extensions to the network and empirical work may be also natural directions for future research. For example, the analyzed two stations network can be expanded to account for the entire ED. Moreover, real data, collected from MCEs can be analyzed and the approach suggested here can be analyzed with this data and extended networks to find managerial insights that will help to prepare and manage the next MCEs.

8 RREFERENCES

[1] Levi, L., Michaelson M., Admi H., Bregman D., Bar-Nahor R. National strategy for mass casualty situations and its effects on the hospital, *Prehospital and Disaster Med*, 17(1): 12-16, 2002.

[2] Henderson AK, Lillibridge SR, Salinas C, Graves RW, Roth PB, Noji EK. Disaster medical assistance terms: Providing health care to a community struck by Hurricane Iniki. *Ann Emergency Med*, 23:726–730, 1994.

[3] Phillips SJ, Knebel A, eds. Providing Mass Medical Care with Scarce Resources: A Community Planning Guide. Prepared by Health Systems Research, Inc. under contract no. 290-04-0010. AHRQ publication no. 07-0001. Rockville, MD: Agency for Health Research and Quality, 2006.

[4] Baker D. Civilian exposure to toxic agents: emergency medical response, *Prehospital and Disaster Med*, 19:174–178, 2004.

[5] Levitin H, Siegelson H, Dickinson S. Decontamination of mass casualties- Re-evaluating existing dogma, *Prehospital and Disaster Med*, 18:200–207, 2003.

[6] Lake W. Chemical Weapons Improved Response Program, Guidelines for Mass Casualty Decontamination During a Terrorist Chemical Agent Incident. Domestic. Preparedness Program, US Soldier Biological and Chemical Command, 2000.

[7] Institute of Medicine. National Research Council. Chemical and Biological Terrorism. Research and Development to Improve Civilian Medical Response.Washington, DC: National Academy Press, 97–109, 1999.

[8] Hirshberg A, Holcomb JB, Mattox KL. Hospital trauma care in multiple-casualty incidents: a critical view. *Ann Emergency Med*, 37:647– 652, 2001.

[9] Aylwin CJ, Konig TC, Brennan NW, et al. Reduction in critical mortality in urban mass casualty incidents: analysis of triage, surge, and resource use after the London bombings on July 7, 2005. *Lancet*, 368:2219 –2225, 2006.

[10] Cushman JG, Pachter HL, Beaton HL. Two New York City hospitals' surgical response to the September 11, 2001, terrorist attack in New York City. J Trauma. terrorist bomb explosions in Madrid, Spain—an analysis of the logistics, *Injuries*, 54:147–154, 2003.

[11] de Ceballos, J P, Turegano-Fuentes F, Perez-Diaz D, et al. Madrid, Spain - an analysis of the logistics, injuries sustained and clinical management of casualties treated at the closest hospital. *Critical Care*, 9: 104–11, 2005.

[12] Kreiss Y., Merin O., Peleg K., Levy G., Vinker S., Sagi R., et al. Early disaster response in haiti: The israeli field hospital experience. *Annals of Internal Medicine,*153(1), 45, 2010.

[13] Merin, O., Ash, N., Levy, G., Schwaber, M. J., & Kreiss, Y. The Israeli field hospital in haiti - ethical dilemmas in early disaster response. *New England Journal of Medicine*, 2010.

[14] McCurry J. Japan: the aftermath. *Lancet*, 377: 1061–62, 2011.

[15] Bar-Dayan A., Beard P., Mankuta, D., Finestone A., Wolf Y., Gruzman C., Levy Y., Benedek P., VanRooyen M., Martonovits G. An earthquake disaster in Turkey: an overview of the experience of the Israeli defence forces field hospital in Adapazari. *Disasters*, 24(3): 262-270, 2000.

[16] Hirshberg A, Scott BG, Granchi T, Wall MJ Jr, Mattox KL, Stein M. How does casualty load affect trauma care in urban bombing incidents? A quantitative analysis. *J Trauma*, 58:686–693, 2005.

[17] Hirshberg A, Stein M, Walden R. Surgical resource utilization in urban terrorist bombing: a computer simulation. *J Trauma*, 47:545–550, 1999.

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

[19] Hirshberg A, Frykberg ER, Mattox KL and Stein M. Triage and Trauma Workload in Mass Casualty: A Computer Model. *Journal of Trauma-Injury Infection & Critical Care*, Volume 69 - Issue 5 - pp 1074-1082, 2010.

[20] Einav SPI, Aharonson-Daniel LPI, Freund HC, Weissmann CC, Peleg KC. In-hospital resource utilization during multiple casualty incidents. *Ann Surg*, 243(4):533-40, 2006.

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

[22] Altay N, Green WG. OR/MS research in disaster operations management. *European J. on Operational Research*, 175:475-493, 2007.

[23] Waeckerle JF. Disaster planning and response. *N Engl J Med*, 324:815-821, 1991.

[24] Kosashvili Y, Aharonson-Daniel L, Peleg K, Horowitz A, Laor D, Blumenfeld A. Israeli hospital preparedness for terrorism-related multiple casualty incidents: can the surge capacity and injury severity distribution be better predicted? *Injury*, 40(7):727-31, 2009.

[25] Atencia I., Moreno P. The discrete-time Geo/Geo/1 queue with negative customers and disasters. *Computersand Operations Research*, 31(9), 1537–1548, 2004.

[26] Dudin A., Semenova O. A stable algorithm for stationary distribution calculation for a BMAP/SM/1 queueing system with Markovian arrival input of disasters. *Journal of Applied Probability*, 41 (2), 547–556, 2004.

[27] Dudin A., Nishimura S. A BMAP/SM/1 queueing system with Markovian arrival input of disasters. *Journal of Applied Probability*, 36(3), 868–881, 1999.

[28] Gregory W.J., Midgley G. Planning for disaster: Developing a multi-agency counseling service. *Journal of the Operational Research Society*, 51 (3), 278–290, 2000.

[29] Barbarosoglu G., Arda Y. A two-stage stochastic programming framework for transportation planning in disaster response. *Journal of the Operational Research Society*, 55 (1), 43–53, 2004.

[30] Belardo S., Karwan K.R., Wallace W.A. Managing the response to disasters using microcomputers. *Interfaces*, 14 (2), 29–39, 1984.

[31] Bryson K.M., Millar H., Joseph A., Mobolurin A. Using formal MS/OR modeling to support disaster recovery planning. *European Journal of Operational Research*, 141 (3), 679–688, 2002.

[32] Sinreich D, Marmor Y. A simple and intuitive simulation tool for analyzing emergency department operations. Proceedings of the 2004 Winter Simulation conference. 2004.

[33] Jacobson E.U., Argon N.T., Ziya S. Priority Assignment in Emergency Response, *Forthcoming Operations Research*, 2011

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

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

[36] Barbarosoglu G., Arda Y. A two-stage stochastic programming framework for transportation planning in disaster response. *Journal of the Operational Research Society* 55 (1), 43–53, 2004.

[37] Oh S.C., Haghani A.Testing and evaluation of a multi-commodity multi-modal network flow model for disaster relief management. *Journal of Advanced Transportation*, 31(3), 249–282, 1997.

[38] Sherali H.D., Carter T.B., Hobeika A.G. A location allocation model and algorithm for evacuation planning under hurricane flood conditions. *Transportation Research* Part B-Methodological, 25 (6), 439–452, 1991.

[39] Mandelbaum A., Massey W. A., Reiman M. I. Strong approximations for Markovian service networks. *Queueing Systems*, 30(1-2), 149-201, 1998.

[40] Mandelbaum, A., Massey W. A., Reiman M. I., Rider R. Time-varying multiserver queues with abandonments and retrials. In Proceedings of the 16th International Teletra±c Congress, P. Key, D. Smith (eds.), 355-364, 1999.

[41] Mandelbaum A., Massey W.A., Reiman M.I., Stolyar A. Waiting Time Asymptotics for Time Varying Multiserver Queues with Abandonment and Retrials. Allerton Conference Proceedings, 1999.

[42] Oliver R.M., Samuel A.H., Reducing letter delays in post offices, *Operations Research*, 10: 839-892, 1962.

[43] Vandergraft J.M. A fluid flow model of networks of queues", *Management Science*, 29, 10, 1198-1208, 1983.

[44] Whitt, W. Two fluid approximations for multi-sever queues with abandonments. *Operation Research Letters*, 33 363–372, 2005.

[45] Whitt W. Fluid models for multiserver queues with abandonments, *Operations Research*, 54 no. 1, 37-54, 2006.

[46] Green L.V., Kolesar P.J., Whitt W. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 2.4, 3, 2005.

[47] Yom-Tov G. Queues in Hospitals: Semi-Open Queueing Networks in the QED Regime, Ph.D. Proposal, IE&M, Technion, 2007.

[48] Liu Y., Whitt W. A fluid model for a large-scale service system experiencing periods of overloading. working paper, Columbia University, New York, NY, 2010. Available at: http://www.columbia.edu/~ww2040/allpapers.html

[49] Liu Y., Whitt W. A Fluid Model for the Many-Server Gt/GI/st + GI Queue. Working paper, Columbia University, New York, NY, 2010. Available at:

http://www.columbia.edu/~ww2040/allpapers.html

[50] Hall R.W. Queueing Methods for Service and Manufacturing, Prentice Hall, NY, 1991.

[51] Kellerer H., Pferschy U., Pisinger D. Knapsack Problems. Springer. 2004.

APPENDIX I:

The technique used for sampling from a non-homogenous Poisson process, with rate $\lambda(t)$, $0 \le t \le T$, where S(I) will contain the arrival times which were samples:

Set λ=max $\{\lambda(t)\}$.
Set t=0, I=0
Generate U1 /* sample U1 from a Uniform[0,1] distribution */
Set t= t-log(U1)/ λ /* sample from a homogenous poisson process with rate λ */
While t < T
    Generate U2 /* sample U2 from a Uniform[0,1] distribution */
    If U2 $\le \lambda(t)/\lambda$
        Set I=I+1, S(I)=t
    Generate U1 /* sample U1 from a Uniform[0,1] distribution */
    Set t= t-log(U1)/ λ /* sample from a homogenous poisson process with rate λ */

APPENDIX II:

Proof of Theorem 1 Step 1:

When mortality rate of the two stations are equal ($\theta = \theta_1 = \theta_2$), the greedy solution is optimal.

Proof by induction:

Stage 1: Proof that when $\theta_1=\theta_2$, any two time units (n=2) and any initial conditions ($Q_1(0)$, $Q_2(0)$), the Greedy solution is the Optimal solution (e.g. the priority of stations is determined by the problem parameters and does not change throughout event: if $(1-p_{12})\mu_1 > \mu_2$ station 1 gets priority, if $(1-p_{12})\mu_1 < \mu_2$ station 2 gets priority and if $(1-p_{12})\mu_1 = \mu_2$).

The time unit we choose can be as small as we wish by doing that we guarantee that the optimal solution for the discrete problem is equal to the optimal solution of the continuous problem.

Stage 2: Assume that when $\theta_1=\theta_2$ and any n (or less) time units the Greedy solution is Optimal, and proof for n+1 time units.

Stage 1:

The problem for the first time unit (t=0) given $Q_1(0)$ and $Q_2(0)$ is :

$$\underset{N_1(0),\, N_2(0)}{\text{Max}} \quad \mu_1[1 - p_{12}]N_1(0) + \mu_2 N_2(0)$$

S.T.

$$N_1(0) \le Q_1(0)$$
$$N_2(0) \le Q_2(0)$$
$$N_1(0) + N_2(0) \le N$$
$$N_1(0),\ \ N_2(0) \ge 0.$$

The optimal solution is:

If $(1-p_{12})\mu_1 > \mu_2$ station 1 gets priority and

$$N_1(0) = \min(Q_1(0), N), \quad N_2(0) = \min(Q_2(0), N - N_1(0))$$

If $(1-p_{12})\mu_1 < \mu_2$ station 2 gets priority and

$$N_1(0) = \min(Q_1(0), N - N_2(0)), \quad N_2(0) = \min(Q_2(0), N)$$

The problem for the second time unit (t=1), given $Q_1(1)$ and $Q_2(1)$ is :

$$\underset{N_1(1),\, N_2(1)}{\text{Max}} \quad \mu_1(1- p_{12})N_1(1) + \mu_2 N_2(1)$$

S.T.

$$N_1(1) \le Q_1(1)$$
$$N_2(1) \le Q_2(1)$$
$$N_1(1) + N_2(1) \le N$$
$$N_1(1),\ \ N_2(1) \ge 0.$$

The optimal solution is:

If $(1-p_{12})\mu_1 > \mu_2$ station 1 gets priority and

$$N_1(1) = \min(Q_1(1), N), \quad N_2(1) = \min(Q_2(1), N - N_1(1))$$

If $(1-p_{12})\mu_1 < \mu_2$ station 2 gets priority and

$$N_1(1) = \min(Q_1(1), N - N_2(1)), \quad N_2(1) = \min(Q_2(1), N)$$

The problem for the two time units altogether is:

$$\underset{\substack{N_1(0),\, N_2(0)\\ N_1(1),\, N_2(1)}}{\text{Max}} \quad (1-p_{12})\mu_1(2\text{-}\theta)N_1(0) + \mu_2(2\text{-}\theta)N_2(0) + (1-p_{12})\mu_1 N_1(1) + \mu_2 N_2(1)$$

s.t.

$N_1(0) + N_2(0) \le N$

$N_1(1) + N_2(1) \le N$

$N_1(0) \le Q_1(0),\ N_2(0) \le Q_2(0)$

$N_1(1) \le Q_1(1),\ N_2(1) \le Q_2(1)$

$N_1(0),\ N_2(0) \ge 0$

$N_1(1),\ N_2(1) \ge 0.$

Eight cases have to be checked regarding the initial conditions:

If $(1-p_{12})\mu_1 \ge \mu_2$:

1. $N \le Q_1(0),\quad N \le Q_1(1),\quad \forall Q_2(0), Q_2(1)$
2. $N > Q_1(0),\quad N > Q_1(1),\quad \forall Q_2(0), Q_2(1)$
3. $N \le Q_1(0),\quad N > Q_1(1),\quad \forall Q_2(0), Q_2(1)$
4. $N > Q_1(0),\quad N \le Q_1(1),\quad \forall Q_2(0), Q_2(1)$

If $(1-p_{12})\mu_1 < \mu_2$:

5. $N \le Q_2(0),\quad N \le Q_2(1),\quad \forall Q_1(0), Q_1(1)$
6. $N > Q_2(0),\quad N > Q_2(1),\quad \forall Q_1(0), Q_1(1)$
7. $N \le Q_2(0),\quad N > Q_2(1),\quad \forall Q_1(0), Q_1(1)$
8. $N > Q_2(0),\quad N \le Q_2(1),\quad \forall Q_1(0), Q_1(1)$

Case 1:

The solution for solving each time unit separately is $N_1(0) = N$, $N_2(0) = 0$, $N_1(1) = N$, $N_2(1) = 0$.

The value of the objective function is:

$T = N\mu_1 (1\text{-}p_{12})(3\text{-}\theta)$

We assume that it is preferable not to assign all the resources to the first station at the first time unit

$N_1(0) < N$, $N_2(0) = \min(N\text{-}N_1(0), Q_2(0))$

We define: $N\text{-}N_1(0) = X > 0 \ \rightarrow\ N_1(0) = N\text{-}X$

Substituting the above in the objective function of the two time units altogether yields the following:

$$T' = N_1(0)(2-\theta)\mu_1(1-p_{12}) + N_2(0)(2-\theta)\mu_2 + N\,\mu_1(1-p_{12}) \le$$
$$N_1(0)(2-\theta)\mu_1(1-p_{12}) + (N-N_1(0))(2-\theta)\mu_2 + N\,\mu_1(1-p_{12}) =$$
$$N(3-\theta)\mu_1(1-p_{12}) - X(2-\theta)[\mu_1(1-p_{12}) - \mu_2] \le N(3-\theta)\mu_1(1-p_{12}) = T$$

Since we wish to maximize the objective function we get a contradiction to our assumption.

We assume that it is preferable not to assign all the resources to the first station at the second time unit

$N_1(1) < N$, $N_2(1) = \min(N-N_1(1), Q_2(1))$

We define: $N-N_1(1) = X > 0 \rightarrow N1(1) = N-X$

Substituting the above in the objective function of the two time units altogether yields the following:

$$T' = N(2-\theta)\,\mu_1(1-p_{12}) + N_1(1)\mu_1(1-p_{12}) + N_2(1)\mu_2 \le$$
$$N(2-\theta)\,\mu_1(1-p_{12}) + N_1(1)\mu_1(1-p_{12}) + N\mu_2 - N_1(1)\mu_2 =$$
$$N(2-\theta)\,\mu_1(1-p_{12}) + N\mu_1(1-p_{12}) - N\mu_2 + N\mu_2 - X[\mu_1(1-p_{12}) - \mu_2] =$$
$$N\mu_1(1-p_{12})(3-\theta) - X[\mu_1(1-p_{12}) - \mu_2] \le N\mu_1(1-p_{12})(3-\theta) = T$$

Again we get a contradiction to our assumption.

The other cases are proven in the same way.

Stage 2:

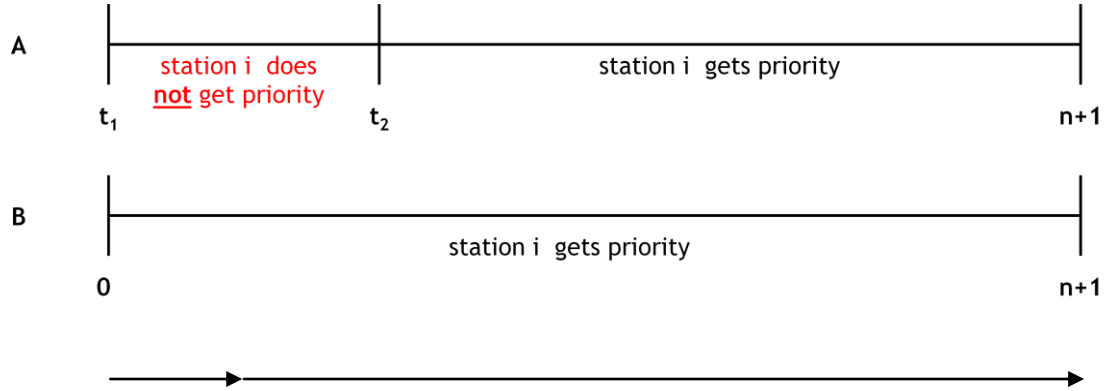
The induction assumption is that for n time units (or less) the Greedy solution is the optimal one and we wish to prove it for n+1 time units.

For n time units station i gets priority. We wish to prove that for n+1 station i will still get priority. We assume that in the optimal solution of n+1 time units exists a time interval $[t_1, t_2]$ of at least one time unit in which the priority is not for station i. There are four options regarding the location of the time interval $[t_1, t_2]$:

1. At the beginning of the event ($t_1=0$, $t_2 < n+1$).
2. At the middle of the event ($t_1 > 0$, $t_2 < n+1$).
3. At the end of the event ($t_1 > 0$, $t_2 = n+1$).
4. At the entire event ($t_1=0$, $t_2 = n+1$).

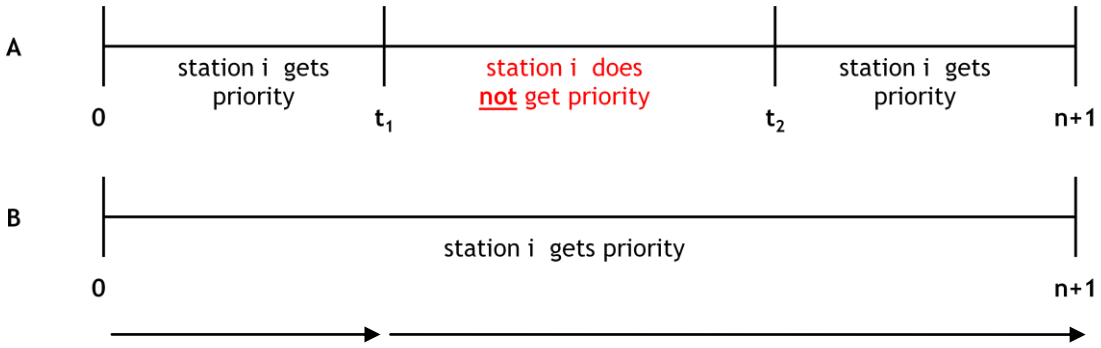
For the four cases we prove that option B is preferable in the following figures

Option 1: $t_1=0$, $t_2 < n+1$



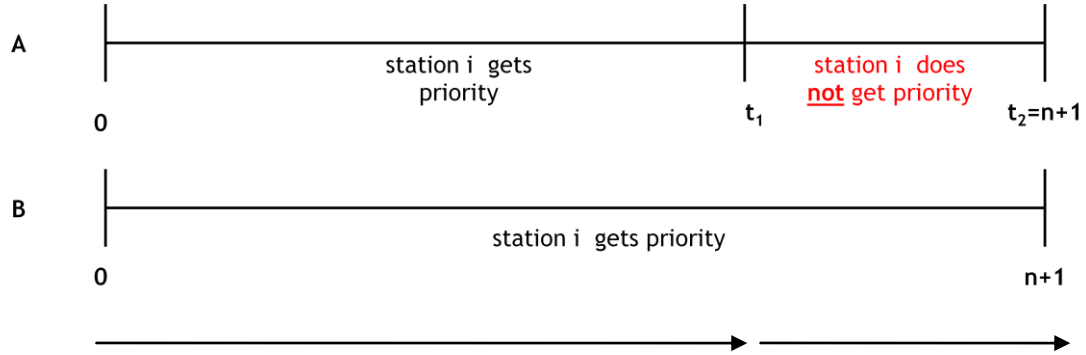
We divide the n+1 minutes into two intervals [0, t] and [t, n+1] where $t_1 < t < t_2$ both with length less than n time units. In the first interval, according to the Induction assumption, option B is preferable since it gives priority to the ith station. In addition, in the second interval the induction assumptions also holds and therefore option B is preferable. If option B is preferable in both intervals, it is also preferable for the entire interval.

Option 2: $t_1 \geq 0$, $t_2 \leq n+1$



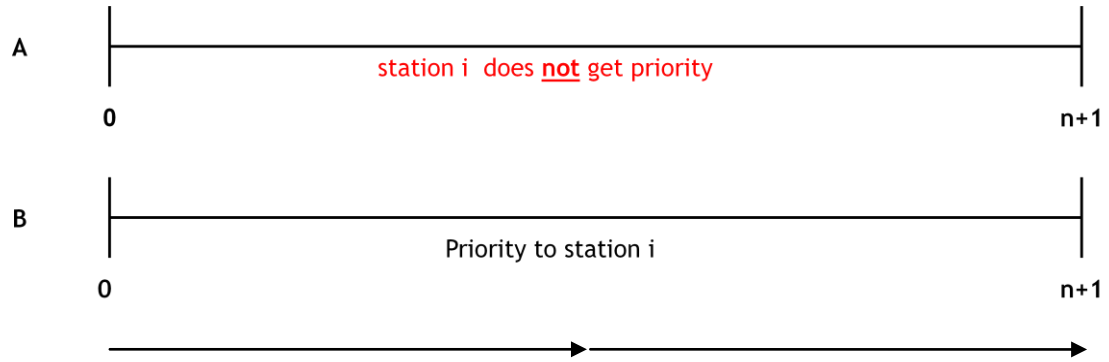
We divide the n+1 time units into two intervals [0, t_1] and [t_1, n+1] both with length less than n time units. In the first interval, the two options are the same. In the second interval, according to the induction assumptions option B is preferable, since it gives priority to station i through the entire interval. If option B is preferable in one interval and is the same in the second, it is also preferable for the entire interval.

Option 3: $t_1 > 0$, $t_2 = n+1$



We divide the n+1 time units into two intervals [0, t_1] and [t_1, n+1] both with length less than n time units. In the first interval, the two options are the same. In the second interval, according to the induction assumptions option B is preferable, since it gives priority to station i. If option B is preferable in one interval and is the same in the second, it is also preferable for the entire interval.

Option 4: t_1=0, $t_2 = n+1$



We divide the n+1 time units into two intervals [0, t] and [t, n+1] both with length less than n time units. In both intervals, according to Induction assumption option B is preferable since it gives priority to station i. If option B is preferable in both intervals it is also preferable for the entire interval.

APPENDIX III:

Proof of Theorem 1 Step 2:

If when $\theta_i=\theta_j$ no station or Station i gets priority, then this priority will also be given to Station i when $\theta_i > \theta_j$.

Proof by induction:

Stage 1: Proof that when $(1-p_{12})\mu_1 \geq \mu_2$ and $\theta_1 > \theta_2$ or when $(1-p_{12})\mu_1 \leq \mu_2$ and $\theta_1 < \theta_2$, any two time units (n=2) and any initial conditions ($Q_1(0)$, $Q_2(0)$), the Greedy solution is the Optimal solution (e.g. the priority is given to the first station in the first case and to the second station in the second case).

Stage 2: Assume the Greedy solution is Optimal under the above conditions for any n (or less) time units, and proof for n+1 time units. Stage 2 is identical to Stage 2 in Step 1.

Stage 1:

The problem for the first time unit (t=0) given $Q_1(0)$ and $Q_2(0)$, is :

$$\underset{N_1(0),\, N_2(0)}{\text{Max}} \quad \mu_1[\theta_1 - \theta_2 p_{12}]N_1(0) + \theta_2\mu_2 N_2(0)$$

S.T.

$$N_1(0) \leq Q_1(0)$$
$$N_2(0) \leq Q_2(0)$$
$$N_1(0) + N_2(0) \leq N$$
$$N_1(0),\ N_2(0) \geq 0.$$

The optimal solution is:

If $(\theta_1 - \theta_2 p_{12})\mu_1 \geq \theta_2\mu_2$ station 1 gets priority and

$$N_1(0) = \min(Q_1(0), N), \quad N_2(0) = \min(Q_2(0), N - N_1(0))$$

If $(\theta_1 - \theta_2 p_{12})\mu_1 < \theta_2\mu_2$ station 2 gets priority and

$$N_1(0) = \min(Q_1(0), N - N_2(0)), \quad N_2(0) = \min(Q_2(0), N)$$

The problem for the second time unit (t=1), given $Q_1(1)$ and $Q_2(1)$ is :

$$\underset{N_1(0),\,N_2(0)}{\text{Max}} \quad \mu_1[\theta_1 - \theta_2 p_{12}]N_1(1) + \theta_2\mu_2 N_2(1)$$

S.T.

$N_1(1) \le Q_1(1)$

$N_2(1) \le Q_2(1)$

$N_1(1) + N_2(1) \le N$

$N_1(1),\ N_2(1) \ge 0.$

The optimal solution is:

If $(\theta_1 - \theta_2 p_{12})\mu_1 \ge \theta_2\mu_2$ station 1 gets priority and

$$N_1(1) = \min(Q_1(1), N), \quad N_2(1) = \min(Q_2(1), N - N_1(1))$$

If $(\theta_1 - \theta_2 p_{12})\mu_1 < \theta_2\mu_2$ station 2 gets priority and

$$N_1(1) = \min(Q_1(1), N - N_2(1)), \quad N_2(1) = \min(Q_2(1), N)$$

The problem for the two time units altogether is:

$$\underset{\substack{N_1(0),\,N_2(0)\\ N_1(1),\,N_2(1)}}{\text{Max}} \quad [\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)]\mu_1 N_1(0) + \theta_2\mu_2(2-\theta_2)N_2(0) + [\theta_1 - \theta_2 p_{12}]\mu_1 N_1(1) + \theta_2\mu_2 N_2(1)$$

s.t.

$N_1(0) + N_2(0) \le N$

$N_1(1) + N_2(1) \le N$

$N_1(0) \le Q_1(0),\ N_2(0) \le Q_2(0)$

$N_1(1) \le Q_1(1),\ N_2(1) \le Q_2(1)$

$N_1(0),\ N_2(0) \ge 0$

$N_1(1),\ N_2(1) \ge 0.$

Eight cases have to be checked regarding the initial conditions:

If $(1-p_{12})\mu_1 \geq \mu_2$:

1. $N \leq Q_1(0), \quad N \leq Q_1(1), \quad \forall Q_2(0), Q_2(1)$
2. $N > Q_1(0), \quad N > Q_1(1), \quad \forall Q_2(0), Q_2(1)$
3. $N \leq Q_1(0), \quad N > Q_1(1), \quad \forall Q_2(0), Q_2(1)$
4. $N > Q_1(0), \quad N \leq Q_1(1), \quad \forall Q_2(0), Q_2(1)$

If $(1-p_{12})\mu_1 \leq \mu_2$:

5. $N \leq Q_2(0), \quad N \leq Q_2(1), \quad \forall Q_1(0), Q_1(1)$
6. $N > Q_2(0), \quad N > Q_2(1), \quad \forall Q_1(0), Q_1(1)$
7. $N \leq Q_2(0), \quad N > Q_2(1), \quad \forall Q_1(0), Q_1(1)$
8. $N > Q_2(0), \quad N \leq Q_2(1), \quad \forall Q_1(0), Q_1(1)$

Case 1:

The solution for solving each time unit separately is $N_1(0) = N$, $N_2(0) = 0$, $N_1(1) = N$, $N_2(1) = 0$.

The value of the objective function is:

$$T = N\mu_1 [\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] + N\mu_1 (\theta_1 - p_{12}\theta_2)$$

We assume that it is preferable not to assign all the resources to the first station at the first time unit

$N_1(0) < N$, $N_2(0) = \min(N-N_1(0), Q_2(0))$

We define: $N-N_1(0) = X > 0 \quad \rightarrow \quad N_1(0) = N-X$

Substituting the above in the objective function of the two time units altogether yields the following:

$$\begin{aligned}
T' &= N_1(0)\mu_1[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] + N_2(0)(2-\theta_2)\theta_2\mu_2 + N\,\mu_1(\theta_1 - \theta_2 p_{12}) \leq \\
&= N_1(0)\mu_1[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] + N(2-\theta_2)\theta_2\mu_2 - N_1(0)(2-\theta_2)\theta_2\mu_2 + \\
&\quad + N\,\mu_1(\theta_1 - \theta_2 p_{12}) = \\
&= N_1(0)\mu_1\{[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] - (2-\theta_2)\theta_2\mu_2\} + N(2-\theta_2)\theta_2\mu_2 + \\
&\quad + N\,\mu_1(\theta_1 - \theta_2 p_{12}) = \\
&= N\{\mu_1[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] - (2-\theta_2)\theta_2\mu_2\} + N(2-\theta_2)\theta_2\mu_2 + \\
&\quad + N\,\mu_1(\theta_1 - \theta_2 p_{12}) - X\{[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] - (2-\theta_2)\theta_2\mu_2\} = \\
&= N\mu_1[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] + N\,\mu_1(\theta_1 - \theta_2 p_{12}) - \\
&\quad - X\{[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] - (2-\theta_2)\theta_2\mu_2\} = \\
&= T - X\{\mu_1[\theta_1(2-\theta_1) - \theta_2 p_{12}(2-\theta_2)] - (2-\theta_2)\theta_2\mu_2\} = \\
&= T - X[\theta_1(2-\theta_1)\mu_1 - \theta_2(2-\theta_2)(\mu_2 + \mu_1 p_{12})] \leq T
\end{aligned}$$

The reason the coefficient of X is positive:

1. $\mu_1 - (\mu_2 + \mu_1 p_{12}) \geq 0$ since $\mu_1(1 - p_{12}) \geq \mu_2$

2. $\theta_1(2 - \theta_1) - \theta_2(2 - \theta_2) \geq 0$ when $\theta_1 > \theta_2$ and $(\theta_1 + \theta_2) \leq 2$ since
   $2\theta_1 - \theta_1^2 \geq 2\theta_2 - \theta_2^2$
   $2(\theta_1 - \theta_2) \geq \theta_1^2 - \theta_2^2 = (\theta_1 - \theta_2)(\theta_1 + \theta_2)$
   when $\theta_1 > \theta_2$
   $2 \geq (\theta_1 + \theta_2)$

Since we wish to maximize the objective function we get a contradiction to our assumption.

The other cases are proven in the same way.

Stage 2: Assume the Greedy solution is Optimal for any n (or less) time units, and proof for n+1 time units. Stage 2 is proven the same as Stage 2 in Step 1, detailed in Appendix II.

מידול וניתוח אירוע רב נפגעי בבית חולים – גישה תפעולית באמצעות מודל נוזלים

נועה ז'יכלינסקי

מידול וניתוח אירוע רב נפגעי בבית חולים - גישה תפעולית באמצעות מודל נוזלים

חיבור על מחקר
לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בהנדסת תעשיה

נועה ז'יכלינסקי

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

תמוז תשע"ב חיפה יולי 2012

המחקר נעשה בהנחיית

דר' יצחק כהן ופרופ' אבישי מנדלבאום

בפקולטה לתעשיה וניהול

אני מודה לטכניון על התמיכה הכספית הנדיבה

בהשתלמותי

תקציר

באירוע רב נפגעים (אר"ן) הצורך במשאבים לטיפול בנפגעים עולה על קיבולת המשאבים הקיימת. אירועים רבי נפגעים אינם נדירים והשפעתם יכולה להיות עולמית (למשל, הצונמי שהתרחש באוקיינוס ההודי ב-2004 וגרם למותם של 200,000 איש) או מקומית (כמו מתקפת טרור, תאונת רכבת או התרסקות מטוס). בכל אירוע כזה קיים חוסר איזון בין עומס העבודה שנוצר לכמות המשאבים הזמינים.

בעת אירוע רב נפגעים בית החולים נדרש ליישם תכנית חירום. הקשיים בהכנת תכניות חירום נובעים מחוסר היכולת לבצע תחזיות בנוגע להתפתחות האירוע ולכמות הפצועים בהם יצטרכו לטפל. בדרך כלל עבור כל תרחיש (אירוע קונבנציונלי, כימי, ביולוגי, אטומי, קרינה ורעידת אדמה) קיימת תכנית חירום ייעודית. תכניות אלו כוללות היבטים קליניים ותפעוליים. במחקר זה, אנו מתמקדים בהיבטים התפעוליים של אירועים קונבנציונליים.

פצועים המגיעים לחדר מיון באר"ן מסווגים לאחר מיון ראשוני (טריאג') לשתי קבוצות: מיידיים (המהווים כ-15% מכלל הפצועים) וממתינים (המהווים כ- 85% מכלל הפצועים). פצוע מיידי הוא פצוע הנמצא בסכנת חיים אלמלא יקבל טיפול רפואי מיידי. בנוסף, משך הטיפול בפצוע מיידי הוא ארוך יותר ודורש משאבים רבים יותר. מסיבות אלו הפצועים המיידים הם היוצרים את מירב עומס העבודה על בית החולים בכלל ועל הצוות הרפואי בפרט. על כן, במחקר זה אנו מתמקדים בסוג פצועים אלו. מייד בתום המיון הראשוני מועבר הפצוע המיידי לטיפול מציל חיים המבוצע בחדרי הלם (עמדות החייאה). בתום הטיפול בפצוע ובהתאם למצבו מחליט הרופא אם להעבירו לטיפול או נמרץ, להדמייה ממוחשבת (בדר"כ CT) או במקרה בו הפצוע אינו יציב לחדר ניתוח. ממחקרים קודמים ומניתוח אירועי אמת עולה כי משאב צוואר הבקבוק הוא הרופאים הכירורגיים, אשר פועלים בשתי תחנות (1) טיפול מציל חיים (2) חדרי ניתוח. מסיבה זו, בחרנו במחקר זה להתמקד בתחנות אלו וברופאים הכירורגיים.

הגישה הנפוצה בניתוח תפעולי של אירועים רבי נפגעים היא שימוש בסימולציה, המאפשרת למדל סביבות מורכבות, סטוכסטיות ודינמיות. החסרונות המרכזיים של סימולציה הם משך הריצה הארוך, העובדה שקשה לקבל תובנות מבניות לגבי התנהגות המערכת וחוסר היכולת לבצע אופטימיזציה. מטרות מחקר זה הן ראשית, למצוא מודל מתמטי שיתאר את המצב התפעולי בחדר המיון בכל רגע ושנית, למצוא מהי ההקצאה האופטימלית (דינמית) של רופאים כירורגיים בין שתי התחנות בהן הם פועלים, במטרה למזער את התמותה באירוע.

החלק הראשון של המחקר עוסק בבחירת מודל הנוזלים המתאים מבין שני מודלי נוזלים הקיימים בספרות. המודל הראשון מבוסס על מערכת של משוואות דיפרנציאליות, אחת עבור כל תחנה, המתארת את כמות הפצועים בכל תחנה בכל רגע. פתרון מערכת המשוואות הוא מספר הפצועים בכל תחנה בכל רגע כתלות בפונקצית ההגעה של הפצועים ובמספר הרופאים הכירורגים בכל תחנה. המודל השני מבוסס על פונקציות הגעה ועזיבה מצטברות של כל תחנה. באמצעות פונקציות אלו ניתן לחשב את כמות הפצועים בכל תחנה (הן בתור והן בטיפול) בכל רגע. המודלים עבור תחנה אחת הורחבו כך שיתאימו למידול רשת תורים.

כדי לתקף ולהשוות בין המודלים, נבנה (ב- SimEvent של Matlab) מודל סימולציה ממוחשב של רשת תורים בחדר מיון. תוצאות הסימולציה שהתקבלו הושוו לתוצאות משני המודלים עבור מספר תרחישים של אירועים. ההשוואה בוצעה באופן גרפי ובאמצעות מדדים סטטיסטיים והעלתה שהמודל הראשון, המבוסס על משוואות דיפרנציאליות מדויק יותר. על כן, בחרנו להתמקד בו בחלק השני של המחקר. בנוסף, מצאנו שהמודל השני כלל

לא מתאים לתיאור המצב התפעולי בכל תחנה אם הרגע בו מתחיל להיווצר תור באחת התחנות קורה לפני שהסתיים משך שירות אחד.

החלק השני של המחקר עוסק בבעיית אופטימיזציה שמטרתה היא למצוא הקצאה אופטימלית של רופאים כירורגים בין שתי התחנות בהן הם פועלים, כך שהתמותה תהיה מינימאלית. את התמותה אנחנו ממדלים כפי שממדלים נטישות במרכז שירות, באמצעות קצב תמותה שמוגדר עבור כל אחת מהתחנות. להבדיל ממרכז שירות רגיל בו הנטישות מתבצעות בזמן המתנה בתור, הנטישות במודל שלנו יכולות להתבצע גם במהלך המתנה וגם במהלך קבלת טיפול. פתרון בעיית האופטימיזציה הראה שבכל רגע ניתנת עדיפות לאחת משתי התחנות, כתלות בפרמטרים של הבעיה (קצבי שירות, קצבי תמותה והסתברות מעבר בין התחנות). בחלק מהתרחישים ניתנה עדיפות לתחנה מסוימת לאורך כל האירוע ובחלק מהמקרים העדיפות התחלפה במהלך האירוע: בהתחלה ניתנה לתחנה אחת ואז התחלפה וניתנה האחרת. כדי להבין את החוקיות, ניסחנו בעיה חמדנית (Greedy Problem) מקבילה. המטרה בבעיה זו היא לקבוע הקצאת משאבים בכל דקה, כך שבדקה הבאה תמוזער התמותה. כלומר, במקום לפתור בעיית אופטימיזציה אחת עבור משך הזמן הכולל, פותרים מספר בעיות אופטימיזציה (לפי משך האירוע בדקות), כך שפתרון הבעיה בכל דקה מהווה תנאי התחלה עבור הדקה הבאה. פתרון הבעיה החמדנית מראה כי לאורך כל האירוע ניתנת עדיפות לאחת מהתחנות, כתלות בפרמטרים של הבעיה, ולא מתבצעת החלפה בעדיפות כפי שהתבצעה בפתרון הבעיה המקורית. מצאנו כלל ברור, לפיו ניתן לחזות, עפ"י הפרמטרים של הבעיה, לאיזו מהתחנות תינתן עדיפות בבעיה החמדנית. כלל זה משמש אותנו בקביעת המדיניות האופטימלית לחלוקת משאבים בבעיה המקורית.

הוכחנו, באמצעות אינדוקציה מתמטית, שאם קצב התמותה בשתי התחנות זהה, הפתרון החמדני הוא הפתרון האופטימלי. במילים אחרות, לא תתבצע החלפה בעדיפות בין התחנות והעדיפות תינתן לתחנה באמצעות הכלל שמצאנו עבור הבעיה החמדנית. בנוסף, הוכחנו באמצעות אינדוקציה מתמטית, שכאשר קצב התמותה בשתי התחנות זהה ויש עדיפות לתחנה הראשונה, אזי גם אם קצב התמותה יהיה גבוה יותר בתחנה הראשונה, היא תמשיך לקבל את העדיפות במשך כל האירוע. וגם להפך, כאשר קצב התמותה בשתי התחנות זהה ויש עדיפות לתחנה השנייה, אזי גם אם קצב התמותה יהיה גבוה יותר בתחנה השנייה, היא תמשיך לקבל את העדיפות במשך כל האירוע . המקרה השלישי, אותו הראינו בצורה אמפירית, מתחלק לשני מקרים: 1. אם כאשר קצב התמותה שווה, העדיפות היא לתחנה הראשונה, אז כאשר קצב התמותה בתחנה השנייה גבוה יותר, תתבצע החלפה בעדיפות. כלומר, בתחילת האירוע תינתן עדיפות לתחנה הראשונה ובמהלך האירוע עדיפות זו תתחלף ותינתן לתחנה השנייה. 2. אם כאשר קצב התמותה שווה, העדיפות היא לתחנה השנייה, אז כאשר קצב התמותה בתחנה הראשונה גבוה יותר, תתבצע החלפה בעדיפות, כלומר בתחילת האירוע תינתן עדיפות לתחנה השנייה ובמהלך האירוע תתחלף העדיפות ותינתן לתחנה השנייה.

בנוסף, ניסחנו בעיית אופטימיזציה דומה, במטרה למצוא הקצאה אופטימלית של משאבים בין שתי התחנות, אך בתוספת אילוץ, המונע שינוי ההקצאה בכל רגע ומאפשר לשנותה רק כל פרק זמן קבוע (בו, בהתאם להערכת המצב יחליט מנהל האירוע כיצד לשנות את ההקצאות של הרופאים).

לסיכום, המודל שאנו מציעים מאפשר לחזות את מספר הפצועים שיהיו בכל תחנה בבית חולים בזמן אירוע רב נפגעי. גישת הפתרון שאנו מציעים מוצאת את ההקצאה הדינמית האופטימלית של רופאים כירורגים בין התחנות,

כך שסך התמותה תמוזער. פתרון בעיה חמדנית מקבילה שניסחנו מאפשר למצוא תנאים בהם פתרון הבעיה החמדנית הוא הפתרון האופיטמלי. עבור מקרים אחרים, ניסחנו את מבנה הפתרון האופטימלי.

על אף פשטותו היחסית של המודל, הוא גמיש ומאפשר לבצע ניתוח מהיר לתרחישים שונים של אירועים רבי נפגעים ובכך לסייע בהכנות לאירועים ובניהולם. המחקר מחזק את ההבנה של ההשפעות התפעוליות באר"ן ודרך המודל ופתרונו ניתן להסיק תובנות ניהוליות בנוגע להקצאה הדינמית של משאבים בין התחנות השונות בהם הם נדרשים.

קיימים מספר כיוונים אפשריים עבור מחקרי המשך. אחד מהכיוונים הוא ביצוע מחקר אנליטי שיאפשר למצוא את נקודת הזמן האופטימלית בה מתבצעת ההחלפה בעדיפויות בין התחנות, כמן כן, יאפשר למצוא חסמים עבור תוצאות הפתרון של הבעיה החמדנית ביחס לפתרון האופטימלי. התובנות ממחקר זה יכולות לשמש כבסיס לניתוח שינויים במבנה הפרמטרים, למשל הגדרת קצבי תמותה או סיכויי הישרדות תלויי זמן או הגדרת קצבים שונים בזמן טיפול ובזמן המתנה. כיוון נוסף הוא ביצוע הרחבה של המודל עבור תרחישי אר"ן שאינם קונבנציונליים (למשל אר"ן כימי, ביולוגי, אטומי וקרינה) אשר בהם תוכנית החירום של בית החולים, הפעילויות והמשאבים הם שונים. כמו כן, ניתן להרחיב את רשת התורים שניתחנו כך שתכלול את כל התחנות בחדר המיון. כיוון מחקר נוסף הוא ניתוח והשוואה של נתוני אמת שנאספו בבתי חולים בזמן אירועים לתוצאות המודל שפיתחנו, במטרה להסיק תובנות ניהוליות שיתרמו בהכנה ובניהול של אירועים עתידיים.