Impact of Patients Priority and Resource Availability in Ambulance Dispatching

Arun Mohan Gnanasekaran, Mohammad Moshref-Javadi, Hao Zhong, Mohsen Moghaddam, and Seokcheon Lee

School of Industrial Engineering, Purdue University, 315 N. Grant St., West Lafayette, IN, USA

Abstract

A main role of Emergency Medical Service (EMS) systems is to efficiently allocate and dispatch emergency medical resources to patients. Efficiency of EMS systems is subject to three critical real-time decisions: 1) prioritizing emergency calls, 2) dispatching appropriate ambulances, and 3) selecting appropriate hospitals to serve the patient. A new model is proposed for improving current ambulance dispatching and relocation models by investigating on these three assumptions. Emergency calls are classified into types A and B, according to their urgency (i.e. priority). After a call arrives, the best available ambulance is dispatched and then the best available hospital is selected for serving the patient. Availability of ambulances and hospitals depend on their real-time capacity. The best choices among ambulances and hospitals are made via two algorithms investigating preparedness of ambulances and hospitals over the time. Moreover, ambulances are relocated according to types of patients; type A must be carried to the best hospital according to area preparedness and distance, while type B is served at the call location and then the ambulance is relocated. Several scenarios are designed to investigate improvements in service level and area preparedness under capacity limitation.

Keywords

Ambulance dispatching, Preparedness, Hospital selection, Call priority, Capacity

1. Introduction

Many factors are involved in an Emergency Medical Service (EMS) to be accomplished successfully: the location of hospitals and ambulances, availability of ambulances, crew proficiency and quickness, etc. The most important performance measure in EMS is response time. All efforts should be done in such a way to minimize response time while maintaining quality of the service. Dispatching the best ambulance to call location is among the essential factors which affect response time of the service significantly. This problem is called “ambulance dispatching problem”. As the simplest dispatching rule, the closest ambulance can be dispatched to the patient location. But, this problem becomes more complicated considering many more factors involved such as capacity of hospitals, ambulances, number of hospitals. In this paper, concentration is on ambulance dispatching and hospital selection considering limited resources to design an effective emergency service in network. It is assumed that there are two categories of patients A and B, prioritized according to severity of situations. There are, however, different classifications of patients in literature; but the main point in the classification used in this study is relocation of ambulances after arriving at the call location and we will elaborate on this in section 3.

Several studies have been conducted to address the problem and improving reliability and quality of EMS responses under limitations in capacity, time, etc. A modular capacitated maximal covering location problem was proposed by Yin & Mu [1] to optimally allocate ambulances to different potential nodes considering several possible capacity levels for each ambulance at a specific node. Schmid [2] proposed an approximate dynamic programming approach to dynamic ambulance relocation and dispatching problem. The objective was to minimize the response time in a day, considering patient’s priority and stochasticity in patient’s pick-up and drop-off times and traveling times (to the scene, to the hospital, and to the waiting site). Iannoni and Morabito [3] analyzed EMS systems on highways considering various types of calls, different servers (i.e. ambulances), pre-allocated servers to specific regions, and multiple dispatching option for specific calls. Kergosien et al. [4] proposed a Tabu Search heuristic for dynamic
transportation of patients considering both static (forecast) and dynamic demand, transportation constraints, prioritizing urgent demands, dissimilar ambulances and different level of nurses’ expertise. A well-known approach for dealing with resource limitation in serving huge call volumes is to prioritize the patients each time with respect to a set of criteria. Lee [5] proposed nearest neighbor policy for prioritizing emergency calls based on centrality and closeness, in which weight on centrality and choice of centrality measure were defined as two parameters of centrality policy. Other priority measures can be defined according to severity of patient’s illness or injury, identified by EMS call takers. Kuisma et al. [6] defined four categories of medical priorities and monitored the resulting pre-hospital mortality measures through classifying patients in different categories of medical priorities. In a similar study Ek et al. [7] analyzed the impacts of assigning different priorities to the emergency calls based on specific medical indexes on the adequacy and quality of EMS. Thus, the following gaps have been recognized in the literature:

- Calls can have different priorities based on the severity of the injured patients. Thus, calls might have unequal urgency.
- It has been supposed that patients are cured in call location and are not taken to hospital. This assumption is not realistic since seriously injured/sick patients must be treated at hospital. Thus, hospitals should be considered in the problem.
- All calls cannot be serviced in accident location. Based on the severity of the accident, some patients should be taken to hospital to be served.
- Capacity constraints and resource level should be considered for ambulances and hospitals.

The hospital preparedness in terms of capacity for the emergency conditions in terms of capacity has not been considered as an important factor in the medical emergency system. The hospital capacity can be used effectively by dispatching the emergency patients to the hospital which are more prepared (less congested than the ones which are less prepared (more congested). This will have a direct impact on the survival rate of the patient. The ambulance capacity also plays a major role in the survival rate and the response time. Instead of sending an ambulance which is ill prepared to meet the needs of the patient, we could send an ambulance which is more prepared.

The objective of this research is effective utilization of the ambulance and hospital capacity and also to reduce the response time taken by the ambulance to reach the patients by effective dispatching of ambulances based on the priority and the capacity of ambulance and hospitals. The rest of the paper is organized as follows. Section 2 represents problem description. Methodology of the research is proposed in Section 3 in addition to the algorithms we designed for ambulance dispatching. The assumptions and parameters of the experiment we use are explained in section 4. Section 5 presents detailed experiment results and analysis. Finally, conclusions and discussion are elaborated in Section 6.

2. Problem Description and Assumptions

In this section, we describe the problem and explain assumptions.

2.1 Network Characteristics

- There are \( N \) hospitals (i.e. hubs) in the network and there are predefined numbers of ambulances in each hospital. The locations of hospitals are predetermined.
- Each hospital has capacity level which depends on the number of available beds, nurses, etc.
- This capacity is variable over time according to the number of patients being served.
- Each ambulance has capacity level which represents amount of medical resources in the ambulance. Resources in the hospital are variable over time. The ambulance is refilled whenever it goes to a hospital.
- Ambulances can be shared by different hospitals. In other words, ambulances can take patients to a hospital which they do not belong to.
- There are different types of calls which can be categorized into types A and B. Patients type A are more severe injured patients while type B includes less severe ones. Thus, priority of calls of type A is higher than the priority of calls of type B.

2.2 Ambulance Dispatching Assumptions

- An ambulance is selected with respect to the type of call, capacity of the ambulance, and preparedness of the area. If the ambulance resource level is not enough to serve, it returns to a hospital which is closest to the ambulance.
- Destination of the dispatched ambulance to the patient zone is defined as follows:
Patient type A must be taken to the hospital to be treated due to the fact that the severity of injury is significant. Thus, the ambulance which was dispatched to the patient zone for injury type A takes the patient back to a hospital which has enough capacity. Figure 1 shows this schematically.

- After both types of services A and B, medical resources in the ambulance are reduced by specific amounts.
- Call arrival time and ambulance travel time are stochastic.

### 3. Methodology

#### 3.1. Priority Assignment

Assignment of priority to calls is one of the most important decisions in EMS. Lindstrom et al. [8] studied the new Emergency Medical Communication Center (EMCC) priority assessment implemented in Finland using performance indicators. In the EMCC the emergency calls were categorized into four types of priority based on the response time needed for each call. In this study, we categorize calls into two types: Type A and Type B.

- Type A: The patient has a life threatening situation or has been exposed to a high energy incident or there is a suspicion of failure of vital functions.
- Type B: No suspicion of failure of vital functions and the patients who need assessment by an emergency medical team.

Victor et al. [9] analyzed the emergency calls and found out that of the total number of calls arriving to the Emergency Medical Service only 60% requires medical service and the rest don’t need a medical service based on the feedback from the emergency medical crew.

In this paper we use this information for assigning the number of calls to type A and type B. Therefore, 60% of the calls arriving to the system will be assigned as Type A and the rest will be assigned as type B.

#### 3.2. Ambulance Dispatching Algorithm

According to problem definition in section 2.2, our problem is composed of two stages:

- Ambulance dispatching
- Hospital selection

Thus, separate algorithms are used at each stage. We will describe each algorithm in this section.

Our proposed algorithm is based on preparedness algorithm by Andersson and Varbrand [10] and Lee [11]. Their algorithm did not consider ambulance capacity and types of calls. Thus, we modify preparedness algorithm to incorporate ambulance capacity and call types (call priority) in the algorithm. The modified algorithm is as follows:

i) When a call arrives, available ambulances which are in distance \( \delta \) (Euclidean distance) from the call location form set \( T \). Since we have two types of calls we use different threshold for different types of calls. Patients type A need more urgent service, thus \( \delta_A < \delta_B \).

ii) If \( T = \emptyset \), then the closest ambulance outside the region is dispatched; Otherwise, go to step iii).

iii) For all \( j \in Z \) (set of all zones), calculate \( Z_j(i) \) which is zone preparedness if ambulance \( i \in T \) is dispatched.

\[
Z_j = \frac{1}{\lambda_j} \sum_{i \in A} \frac{C_i}{t_{ij}}, \quad \forall j
\]  

(1)

where \( \lambda_j \) represents the call rate in zone \( j \), \( A \) is the set of available ambulances, \( C_i \) current available
ambulance resource, and \( t_{ij} \) is travel time between ambulance \( i \) and location \( j \).

iv) Aggregate zone preparedness to calculate area preparedness \( a(j) \). Different methods can be used to aggregate zone preparedness into area preparedness. One method is average. Other methods are social welfare functions and \( \min_j \{ z_j \} \).

v) Select ambulance \( i \) which has the greatest \( a(i) / t_{ic} \) value, where \( t_{ic} \) is travel time between call location and ambulance \( i \).

Note that if a call arrives and no ambulance is available, it is supposed that calls are queued. In this queue all call types are queued. Once an ambulance is released, calls of Type A have more priority to be served.

3.3 Hospital Selection Algorithm

There are two types of calls in this problem, call types A and B. After an ambulance is dispatched to a call location for patient type B, patient is served and the ambulance will wait for a pre-specified time in that location. If no call arrives during this time, the ambulance returns to a hospital to refill. In the second case, if the ambulance is dispatched to a call location to serve patient type A, patient should be taken to a hospital to be served. Thus, we have two types of hospital selection: 1- selection of hospital for ambulance which is taking patient type A to a hospital, 2- selection of hospital for ambulance which is returning to a hospital after serving patient type B.

1) **Hospital selection algorithm for an ambulance which is taking patient type A to hospital:**

i) All of the hospitals which are in distance \( \tau \) of the ambulance and have minimum amount of required resources (\( \beta \)) are identified and create set \( U \).

ii) For all hospitals in set \( U \) (\( h \in U \)), the below formula is calculated:

\[
S_h = \frac{c_h}{t_{ih}} , \quad \forall h \in U ,
\]

where \( c_h \) is current available resource level at the hospital and \( t_{ih} \) is travel time between the ambulance taking patient type A and hospital \( h \). By using this formula, more priority is assigned to hospitals which are closer to the call location and has more current resource.

iii) Hospital which has the greatest \( S_h \) is selected.

2) **Hospital selection algorithm for an ambulance after serving patient type B and waiting for a pre-specified time:**

Ambulances after serving patient type B should go back to hospital if no calls arrive after a pre-specified time. The closest hospital to the ambulance is selected and the ambulance returns back to the hospital to refill.

4 Experiment Design

4.1 Parameters Setup

In order for us to investigate the effect of capacity and call priority on ambulance dispatching and hospital selection problems, we designed an experiment with different scenarios. In the context of this research, we are modeling EMS for urban area. A city (service area) can be abstracted as a square with side \( L \). To simulate a model as close to real life as possible, we use some of the base parameter set reported in Alanis et al. [12]. Edmonton can be estimated as a square with side length \( L = 26 \) km. We use four hospitals in the urban area and their locations (coordinates) are known.

Generally, the arrival rate of calls varies from time to time and from city to city [13]. A Poisson process can be used to model the arrival of calls in the urban area. We also use four different call patterns: uniform, centered, cornered, and bipartite. Their probability density functions and discretized rates are shown in Table 1.

When the EMS receives a call, the caller will report the coordinates to the call operator. During the conversation, the operator will ask several predefined questions to find which type of patient is requesting the service and determine the priority. In the simulation, the emergency calls arrive in a Poisson process with arrival rate (\( \lambda \)).

Call and hospital locations are shown in Figure 2. The call arrival rate is averagely 4 per hour. Besides these available data, we assume there are infinite dispatchers/call operators in the system to answer calls, so every call will be answered immediately. The parameters are summarized in Tables 2 and 3.

4.2 Scenarios

The results of the proposed ambulance dispatching and hospital selection algorithms will be examined under 3 scenarios. Scenario 1 (With Priority) uses capacity values presented in Table 3 with prioritized patients, i.e. patient types A and B. Scenario 2 is similar to Scenario 2 (Without Priority) except that all calls are of type A and no call
type B arrives. Patients type A are treated as we explained in section 3. Scenario 3 has been designed to investigate the effect of capacity on performance measures. In Scenario 3, we use prioritized patients where capacity of hospital 2 is 50 and capacity of hospital 3 is 30. The hospital preparedness accommodates the incoming patients so that it can provide adequate treatment facilities for the patients and also avoid congestion for emergency cases. The scenarios are simulated for different sets of the values and the corresponding response time and the overall time for the Type A patient to reach the hospital are studied. Each scenario is simulated under four different call pattern. The results are compared with the existing models to get an insight on the performance of our algorithms.

![Figure 2: Network used in the example problem. Circles are call locations (12 call locations). There are 4 hospitals and 3 ambulances in the network.](image)

<table>
<thead>
<tr>
<th>Call Pattern</th>
<th>Call rate (discretized for 3 by 4 zones)</th>
<th>Call Pattern</th>
<th>Call rate (discretized for 3 by 4 zones)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>0.08333 for all zones</td>
<td>Cornered</td>
<td>0.2106 0.0301 0.0301 0.2106</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0162 0.0023 0.0023 0.0162</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2106 0.0301 0.0301 0.2106</td>
</tr>
<tr>
<td>Centered</td>
<td>0.0405 0.0891 0.0891 0.0405</td>
<td>Bipartite</td>
<td>0.0278 0.0278 0.0278 0.0278</td>
</tr>
<tr>
<td></td>
<td>0.0752 0.1655 0.1655 0.0752</td>
<td></td>
<td>0.0833 0.0833 0.0833 0.0833</td>
</tr>
<tr>
<td></td>
<td>0.0405 0.0891 0.0891 0.0405</td>
<td></td>
<td>0.1389 0.1389 0.1389 0.1389</td>
</tr>
</tbody>
</table>

Table 2: Simulation input parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>length of service square</td>
<td>26 km</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Emergency call arrival rate</td>
<td>Exp(1/15) (mean = 15 minutes)</td>
</tr>
<tr>
<td>$N$</td>
<td>Base number of ambulances</td>
<td>3</td>
</tr>
<tr>
<td>$H$</td>
<td>Base number of hospitals</td>
<td>4</td>
</tr>
<tr>
<td>$Z$</td>
<td>Number of call zones</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Proportion of call types ($A%$, $B%$)</td>
<td>(60%, 40%)</td>
</tr>
</tbody>
</table>
Table 3: Input data for ambulances and hospitals

<table>
<thead>
<tr>
<th>Item</th>
<th>Value/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1 location/ capacity</td>
<td>(10.9, 10.4) / 40</td>
</tr>
<tr>
<td>Hospital 2 location/ capacity</td>
<td>(12.5, 18.6) / 30</td>
</tr>
<tr>
<td>Hospital 3 location/ capacity</td>
<td>(13.6, 6) / 50</td>
</tr>
<tr>
<td>Hospital 4 location/ capacity</td>
<td>(16.9, 4) / 40</td>
</tr>
<tr>
<td>Ambulance 1 starting location/ capacity</td>
<td>Hospital 1 / 10</td>
</tr>
<tr>
<td>Ambulance 2 starting location/ capacity</td>
<td>Hospital 2 / 10</td>
</tr>
<tr>
<td>Ambulance 3 starting location/ capacity</td>
<td>Hospital 3 / 10</td>
</tr>
<tr>
<td>( \delta_A )</td>
<td>8.7</td>
</tr>
<tr>
<td>( \delta_B )</td>
<td>1.2 ( \delta_A )</td>
</tr>
<tr>
<td>Coordinates of call locations</td>
<td>1-(0,0) 2- (8.6,0) 3-(17.2,0) 4-(26,0) 5-(0,13) 6- (8.6,13) 7-(17.2,13) 8-(26,13) 9-(0.26) 10- (8.6,26) 11-(17.2,26) 12-(26,26)</td>
</tr>
<tr>
<td>Ambulance resource reduction</td>
<td>3</td>
</tr>
<tr>
<td>Travel times</td>
<td>Exponential of distance</td>
</tr>
<tr>
<td>Service time for patient type A at hospital</td>
<td>Exponential- mean: 5 hours</td>
</tr>
<tr>
<td>Service time for patient type B at call location</td>
<td>3 minutes</td>
</tr>
</tbody>
</table>

5. Computational Results

A test problem is designed and applied to investigate on performance of the proposed model and algorithms in improvement of EMS network efficiency under different scenarios. The test problem (see Figure 2) is designed for a small-sized EMS network with 3 ambulances and 4 hospitals and the input parameters are set as mentioned in Table 4. Four call patterns uniform, centered, cornered, and bipartite are examined for each scenario. The model is run under different scenarios as discussed in Table 5.

5.1 Impact of Priority

5.1.1 Average Travel & Response Time

The average travel time to the hospitals as well as response time for patients type A and B are shown in Figure 3 under call patterns Uniform, Centered, Cornered and Bipartite.

As shown in Figure 3, average response time for patients and average hospital travel time are simulated under Scenarios 1 (i.e. priority for calls patients) and 2 (i.e. no priority for calls). Results show that except Uniform call pattern, the average response and travel time is considerably decreased in Scenario 1 compared to Scenario 2. That is, considering different priorities for the calls in Scenario 1, the ambulances will get free after serving patients type B in the call location and so will be able to serve other incoming calls in a more timely-efficient manner, compared to Scenario 2, where all ambulances have to get back to the hospital after serving every call, regardless of type of service requested (i.e. patient). A significant observation in Figure 3 is that under uniform distribution of calls, effect of call priority on the performance of the EMS network will be decreased (i.e. Scenarios 1 and 2 yield
almost the same results). That is the result of having equal preparedness levels at all zones due to uniform distribution calls. The results are also shown in a different form in Figure 4, where impact of call patterns is analyzed on the average service and response times, under Scenarios 1 and 2. An observation from Figure 4(c) and 5(d) is higher service time in the cornered pattern. The reason behind this fact is higher call rates in the corners of the cornered pattern and on the other hand central locations of hospitals in the network. Therefore, ambulances take longer distances from corners of the network to the hospitals which are located in the center of the network. Thus, service time, especially for patient type A increases.

![Figure 4](image)

**Figure 4**: Comparison of average hospital travel time and response time to the patients and effects of different call patterns: a) Scenario 1; call pattern comparison, b) Scenario 2; call pattern comparison, c) Scenario 1; response & travel time comparison, d) Scenario 2; response & travel time comparison

### 5.1.2 Resource Utilization

Resource utilization refers to the portion of time that a given resource (in this case hospitals and ambulances) is busy, i.e. serving entities (in this case patients), over total available time. The average utilization of hospitals and ambulances under both scenarios of with and without call priority with four call patterns are simulated and the results are shown in Figures 5 (hospitals) and 7 (ambulances).

![Figure 5](image)

**Figure 5**: Average utilization of hospitals under four call patterns: a) Scenario 1; b) Scenario 2

![Figure 6](image)

**Figure 6**: Average ambulance utilization under four call patterns: a) Scenario 1; b) Scenario 2

As shown in Figure 5, the average utilization of hospitals is significantly decreased under Scenario 1, i.e., assigning priorities to the calls, compared to Scenario 2. An important observation here is that under Scenario 1, patients are
classified as Type A and Type B and then patients Type B are served in the call location without being carried to the hospitals. Accordingly, utilization of hospitals will be decreased compared to Scenario 2, where all patients have to be served at hospitals. This will, in turn, lead to lower congestion in hospitals under Scenario 1 compared to Scenario 2, as shown in Figure 5. This implies that through classification of patients, not only more severe cases are prioritized, but also congestion in hospitals will be decreased through providing the EMS network with this option to perform medical services on site, avoiding overloading of hospitals. Another meaningful observation in these figures is that hospital 3 is considerably more utilized than other hospitals. The reason is the fact that based on higher capacity of this hospital in comparison with other hospitals this hospital has more chance of being selected in the proposed hospital selection algorithm. Thus, the utilization of this hospital is higher than other hospitals.

Another important observation is in average utilization of ambulances (Figure 6). Under Scenario 1, based on definition, ambulances serve patients Type B on site, in contrast with Scenario 2, where all patients have to be carried to a hospital. As a result, under Scenario 1, ambulances travel less distances compared to Scenario 2, since not all patients must be necessarily carried to the hospitals; therefore, average utilization of ambulances (i.e. busy time) under Scenario 1 will be less than Scenario 2, as shown in Figure 6. In a nutshell, it is also shown that Scenario 1 (i.e. prioritizing patients) outperforms Scenario 2, in terms of utilization and availability of resources.

Figure 6 shows that ambulances are more utilized in the cornered pattern. This result is reasonable due to the fact that more calls are generated in the corners of the network. On the other hand, hospitals are mostly located in the center of the network. As a result, ambulances need to travel longer distances to refill or take patient type A to the hospital for further treatment. Thus, utilization of ambulances as well as the service time (Figure 4) increases.

5.1.3 Patients’ Waiting Times
The average waiting times of patients to be served in hospitals are calculated for Scenarios 1 and 2 under four call patterns. Results are shown in Figure 7.
As shown in Figure 7, the average waiting time of patients is almost equal in all hospitals under Scenarios 1 and 2, except in Hospital 3. As mentioned in the previous section, average hospital utilization is very high in Hospital 3 compared to other hospitals, which is due to its location and available capacity as well. However, under Scenario 1, average waiting time in the hospitals, which can be interpreted as congestion in hospitals, is significantly decreased compared to Scenario 2. That is because of classification and prioritization of patients in Scenario 1 under which some patients are served in the call location (i.e. patients type B), less patient are carried to the hospitals and so congestion level is decreased. Hence, it also implies superiority of Scenario 1 to Scenario 2 in terms of minimizing waiting time and congestion in hospitals and thus increasing service and customer satisfaction level as well.

5.2 Impact of Capacity
In this section, we compare Scenario 1 and Scenario 3 on hospital capacities. In Scenario 1 and 3 we use all parameters in Table 3 except that in Scenario 3 hospital 2 and 3 capacities are 50 and 30 respectively. Figure 8 illustrates the results of simulation for each call pattern. As it can be seen, results are almost the same in the uniform and bipartite call patterns. In the cornered and centered types, there are some differences in response time of patient type A while there is not major effect on patient B response time. Results show that when more capacity is allocated to hospital 2 than hospital 3, response time has reduced. These results indicate that the capacities of hospitals as well as the locations of hospitals have effect on the response time of patients type A.
Figure 8: Average hospital travel time and response time to the patients: a) Uniform, b) Centered, c) Cornered, and d) Bipartite call patterns

Figure 9, 10 illustrates resources (hospitals and ambulances) utilizations. There are no major changes in ambulances’ utilizations since we kept the capacities of all ambulances unchanged in scenario 3. However, as Figure 10 shows, utilization of hospital 2 has increased as we have increased hospital 2 capacity from 30 to 50 and reduced hospital 3 capacity from 50 to 30. According to our proposed hospital selection algorithm, capacity plays an important role in hospital utilization. This fact can also be concluded from Figure 11 as the average waiting time of patients increased in Scenario 3 in comparison with Scenario 1. This is mainly due to hospital 2 location along with the capacity factor.

Figure 9: Average ambulance utilization for each scenario: a) Uniform, b) Centered, c) Cornered, and d) Bipartite

Figure 10: Average hospital utilization for each scenario: a) Uniform, b) Centered, c) Cornered, and d) Bipartite

Figure 11: Average waiting time of patients for each scenario: a) Uniform; b) Cornered; c) Centered; d) Bipartite
6. Conclusions
In this study, a simulation inspired from two preparedness algorithms for ambulances and hospitals is proposed for improving performance of EMS networks in terms of efficient allocation and dispatching of emergency medical resources to the patients. Patients are classified as type A and B based on severity and also type of requested service. That is, patients Type ‘A’ have more priority and must be taken to a hospital, while patient type B can be served at the call location. Results show that classifying patient in the two aforementioned categories will result in less travel and service time and less congestion in the hospitals in terms of utilization and availability of resources as well as patients’ waiting times in the hospitals. Accordingly, we showed that prioritizing emergency calls, dispatching appropriate prepared (or available in terms of capacity) ambulances, and selecting appropriate prepared (based on available capacity) hospitals to serve the patients are three key decisions in improving the performance of EMS networks. Besides, providing ambulances with the option that they can go to another call location after serving non-critical cases plays a strategic role in balancing capacity with the dynamic demands (calls). The results are also investigated based on different call patterns, uniform, centered, cornered and bipartite. Results show that under uniform call pattern, Scenarios 1 and 2 perform the same, which may be due to uniform distribution of calls in all zones with no weights for different zones. However, this issue is worthy of attention in future studies and also further development of this work as well. We also investigated the effects of capacity of three different performance measures we focused on. According to our hospital selection algorithm, results show that capacity of hospitals play important role in average hospital utilization and waiting of patients at the hospitals. As an area for future research, we highly recommend that more hospital selection algorithms be proposed and the results be compared on service time as well as other performance measures such as average waiting time at hospital and hospital utilization.

Acknowledgment
The authors would like to acknowledge the constructive comments of anonymous reviewers.

References