Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds.

COMPARISON OF AMBULANCE DIVERSION POLICIES VIA SIMULATION

Adrian Ramirez-Nafarrate

Instituto Tecnologico Autonomo de Mexico
Rio Hondo 1
Mexico City, 01080, MEXICO

A. Baykal Hafizoglu
Esma S. Gel
John W. Fowler

Arizona State University
699 S. Mill Avenue
Tempe, AZ 85281, USA

ABSTRACT

Ambulance diversion (AD) is often used by emergency departments (EDs) to relieve congestion. When an ED is on diversion status, the ED requests ambulances to bypass the facility; therefore ambulance patients are transported to another ED. This paper studies the effect of AD policies on the average waiting time of patients. The AD policies analyzed include (i) a policy that initiates diversion when all the beds are occupied; (ii) a policy obtained by using a Markov Decision Process (MDP) formulation, and (iii) a policy that does not allow diverting at all. The analysis is based on an ED that comprises two treatment areas. The diverted patients are assumed to be transported to a neighboring ED, whose average waiting time is known. The results show significant improvement in the average waiting time spent by patients in the ED with the policy obtained by MDP formulation. In addition, other heuristics are identified to work well compared with not diverting at all.

1 INTRODUCTION

Overcrowding of Emergency Departments (EDs) is a problem present in many countries around the world. This problem has been highlighted in the United States (US) in the last decade and is very likely to continue since the arrivals of patients has an increasing trend in recent years (Associated Press 2006; Centers for Disease Control and Prevention 2010).

There are several problems caused by overcrowding of EDs, one of the most important is the long waiting time before a patient is treated. Several papers have studied this problem using queuing formulations and simulation models to re-design patient flow, identify bottlenecks and increase the throughput of EDs (McConnell et al. 2005; Cochran and Roche 2009). Another alternative to relieve congestion is diverting ambulances to less crowded facilities (United States General Accounting Office 2003; United States General Accounting Office 2009). This paper compares different ambulance diversion (AD) policies using simulation.

AD policies refer to the rules or guidelines that must be met in order to trigger the diversion status. While on diversion, an ED requests that all ambulances bypass the facility and transport the patients to another ED. Practitioners consider AD as an inefficient and risky method; therefore, they recommend to avoid AD. Papers from the medical perspective have the objective of designing AD policies that minimize diversion (Vilke et al. 2004; Asamoah et al. 2008; Patel et al. 2006). On the other hand, prohibiting AD might lead to increase in waiting time and stressing the operations of EDs (Massachusetts Nurses Association 2009).

Few papers have analyzed the impact of AD policies using analytical methods. For example, papers by Deo and Gurvich (2011) and Ramirez et al. (2011) analyzed the impact of AD policies on a network of EDs, the former using game theory and queuing formulations, and the later using discrete-event simulation. In addition, Hagtvedt et al. (2009) also studied the diversion decisions using agent-based simulation. These papers suggest the existence of an agent that coordinates AD decisions on a network of EDs. Furthermore,

these papers analyzed AD heuristics with a threshold structure on the ED occupancy, but they do not explore the optimality of the policies.

In this paper, we propose an AD policy obtained by a Markov Decision Process (MDP) formulation, and compare the average waiting time of patients using that policy with that of simple policies that are likely to be used in practice. The remaining sections of the paper are organized as follows: Section 2 describes the simulation model used in the experimentation, Section 3 introduces the AD policies used in the experimentation and presents the formulation of the MDP model, Section 4 presents the results and the comparison of the effectiveness of different AD policies and finally Section 5 presents the concluding remarks.

2 SIMULATION MODEL OF AN ED

2.1 Overview of the Patient Flow

Figure 1 shows an overview of the patient flow in the simulation model. The model consists of two arrival streams: ambulance arrivals and walk-in arrivals. We assume non-homogeneous Poisson process for the arrivals given that the total arrivals to the ED exhibit a pattern seen in several places around the US (Centers for Disease Control and Prevention 2008; Green 2006; Cochran and Roche 2009).

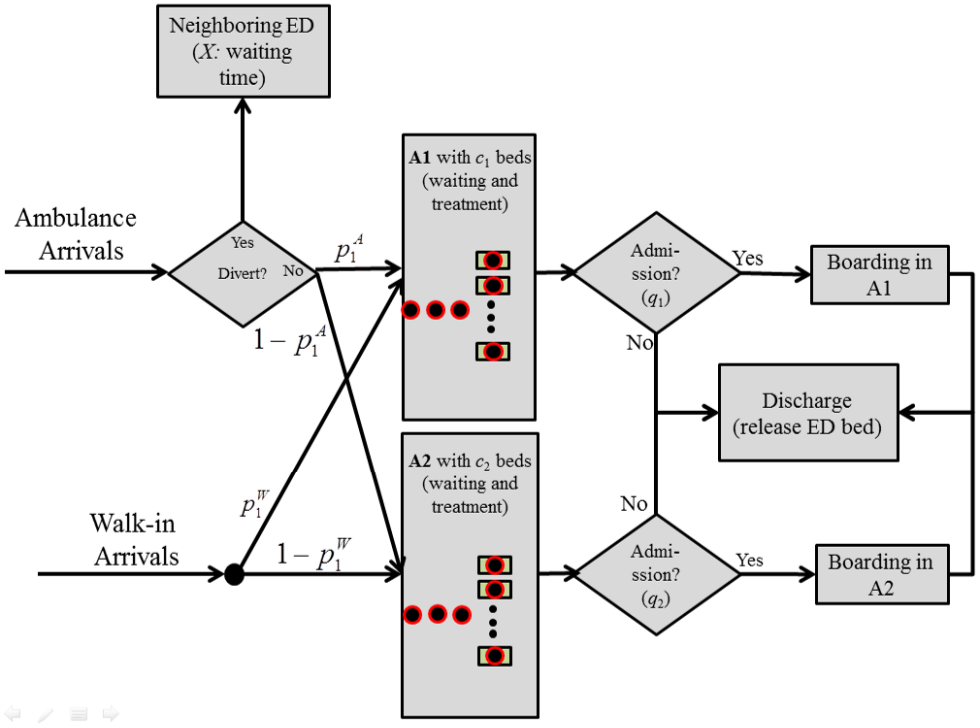


Figure 1: Overview of the patient flow in the simulation model.

Arriving patients are classified in one out of two severity levels: most critical patients belong to level 1 and less emergent patients belong to level 2. There are two treatment areas: A1 and A2 where patients with level 1 and level 2 are treated, respectively. This structure is similar to layouts of many EDs across the US, where a fast-track area is included (Cochran and Roche 2009). Furthermore, severity level 1 represents patients with immediate and emergent need of care, and severity level 2 includes patients with urgent and semi/non urgent needs. We assume that ambulance patients are classified as level 1 with probability p_1^A and level 2 with probability $1-p_1^A$. Similarly, walk-in patients are classified as level 1 with probability

p_1^W and level 2 with probability $1 - p_1^W$. We refer to the tuple (p_1^A, p_1^W) as the patient mix. Patients are served according to the FCFS discipline in each treatment area. The number of beds where patients receive treatment is c_1 and c_2 for areas A1 and A2, respectively. The only resource modeled in the simulation is the bed. The doctors, nurses and lab equipment are discarded from the model because the CDC found that other resources have a low impact on the diversion decisions (Centers for Disease Control and Prevention 2006a).

Once a bed is assigned to a patient, the length of stay is divided in treatment time and boarding time. The treatment time in an ED bed is assumed to have a lognormal distribution, which is identified as a reasonable assumption to model treatment times (Hoot et al. 2008). After ending treatment, the patient with severity level i might require to be admitted to an inpatient unit of the hospital with probability q_i. If the patient requires admission, then the patient remains in the ED bed for an additional amount of time, which is referred to as boarding time. While a patient is boarding in the ED, he/she blocks access to that bed in the ED to other patients.

The AD policies require to observe if the conditions to divert a patients are met given the current state of the system upon an ambulance arrival. If the diversion condition is satisfied, then the ambulance patient is assumed to be transported to a neighboring ED, where the waiting time is a random variable X. Similar to the arrival pattern observed for the ED, it is reasonable to assume that if a patient is diverted, the expected waiting time in a neighboring ED also changes during a day. Therefore, we change the parameters of the distribution of X throughout the day.

The simulation model is built in Arena (Kelton et al. 2010). Pilot runs are used to determine a warm-up period of two months and the replication length after warm-up is set to one year. Thirty replications for each policy and scenario are executed, obtaining an average relative precision of 2.82% using 95% confidence intervals on the average patient waiting time. In addition, common random numbers are applied in order to reduce the noise when comparing the policies (Banks et al. 2010).

2.2 Input Data for the Simulation Model

In order to build a realistic model, we use information from published papers and reports from US agencies to obtain the input data. The input requirements are set as follows:

- **Severity mix (p_1^A, p_1^W):** It is reasonable to assume that p_1^A is a value close to 1. For this experimentation, we set $p_1^A = 0.9$ and $p_1^W \in \{0.3, 0.5\}$. If the severity mix is (0.9, 0.3), then the area A2 is congested, meaning that utilization of beds in A2 is higher than in A1. If the severity mix is (0.9, 0.5), then the congested area is A1.
- **Number of beds:** We assume that $c_1 = 15$ beds and $c_2 = 5$ beds. These values are close to the average number of standard treatment spaces and other treatment spaces found by the CDC in the US (Centers for Disease Control and Prevention 2006b).
- **Arrival rates:** Ambulance and walk-in arrival rates are assumed to follow non-homogeneous Poisson processes. Green (2006) finds reasonable to use a Poisson process to model arrivals to EDs. Cochran and Roche (2009) presents multiplicative indices for the seasonality of arrivals to an ED throughout the day. In order to mimic that pattern, we first find λ^W such that the utilization due to walk-in arrivals is 90% for the peak hour between 7pm and 8pm. Then, we scale the arrivals across the day using the multiplicative indices. We find λ^A assuming that ambulance arrivals represent 15% of all the arrivals to the ED. This percentage is close to the average found by Centers for Disease Control and Prevention (2010). The resulting arrival pattern can be observed in Figure 2 for a severity mix of (0.9, 0.3).
- **Treatment time:** We assume that the treatment time in the ED is lognormally distributed with mean of 240 minutes and 60 minutes for patients with levels 1 and 2, respectively. The standard deviation of the distribution was adjusted to obtain the coefficients of variations found in Cochran and Roche (2009), which are 0.72 and 0.102 for treatments in A1 and A2, respectively.

- **Boarding time:** We assume that $q_1 = 0.24$ and $q_2 = 0.045$, which imply that 12.1% of all the ED patients require admission to an inpatient unit. This percentage is the same found by Singer et al. (2011) and very similar to the findings in United States General Accounting Office (2003). Furthermore, based on Singer et al. (2011) we assume that boarding time is uniformly distributed in between [0,2], [2,6], [6,12] and [12,24] hours with probabilities 0.5022, 0.3705, 0.0763 and 0.051, respectively. This scheme produces an average waiting time of 3.58 hours, which is very similar to the averages found in other references (United States General Accounting Office 2003).
- **Waiting time in the neighboring hospital, X:** We define three settings for the distribution of X. In all the settings, X follows a triangular distribution with parameters as shown in Table 1. In addition, the values of the parameters also depend on the traffic in the ED under study. The traffic is classified as low, medium and high as also shown in Figure 2. The purpose of this scheme is to have a positive correlation between the traffic intensity in the ED under study and the expected waiting time in the neighboring ED. Given the behavior of arrivals observed in Figure 2, it is very likely that the traffic intensity follows the same pattern for neighboring EDs.

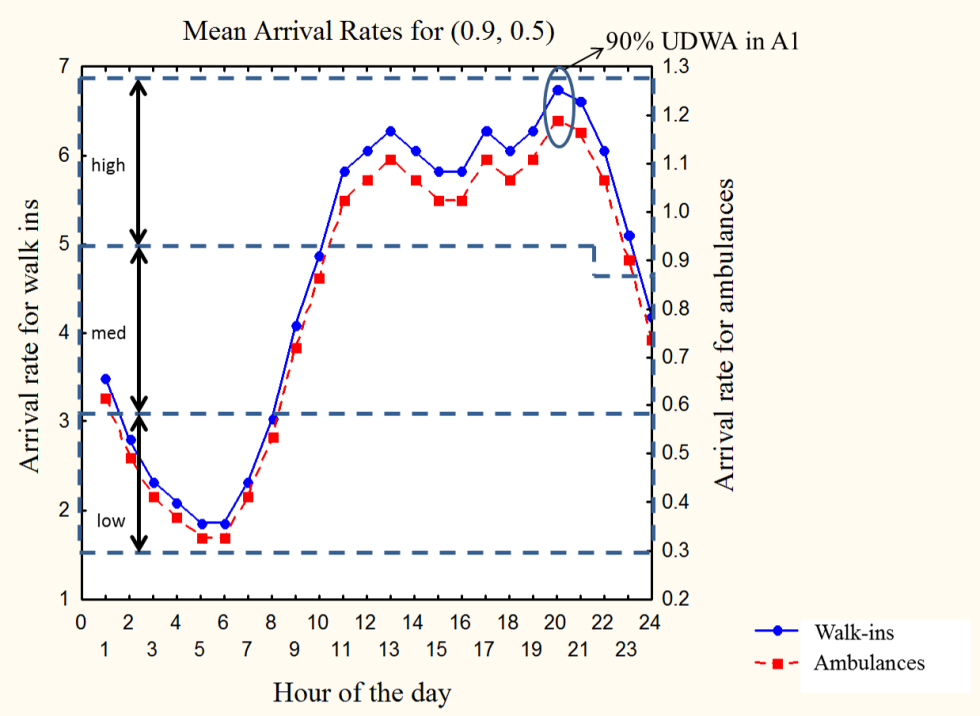


Figure 2: Mean arrival rates to the ED for a severity mix of (0.9, 0.3).

Table 1: Settings of X used in the simulation model.

	Parameters of Triangular Distribution (mins)		
Traffic in main ED	Setting 1	Setting 2	Setting 3
Low	(5, 15, 25)	(5, 15, 25)	(10, 30, 50)
Medium	(10, 30, 50)	(15, 45, 75)	(25, 75, 125)
High	(15, 45, 75)	(25, 75, 125)	(40, 120, 200)

3 AD POLICIES EVALUATED

The simulation model described in Section 2 is used to compare several AD policies, including one obtained by using a Markov Decision Process model and other simple policies found in the literature, some of them applied in practice. The complete list of AD policies is the following:

1. MDP Policy (MDP). AD policy obtained by an MDP model over a simplified version of the ED presented in Section 2.
2. Full Beds in A1 (FB A1). This policy diverts patients to a neighboring hospital when all the beds in A1 are occupied.
3. 14 Beds in A1 (14 A1). This policy diverts patients when there are 14 beds occupied in A1, which implies that there is only 1 bed available in A1.
4. 13 Beds in A1 (13 A1). This policy diverts patients when there are 13 beds occupied in A1, which implies that there are 2 beds available in A1.
5. 12 Beds in A1 (12 A1). This policy diverts patients when there are 12 beds occupied in A1, which implies that there are 3 beds available in A1.
6. Full Beds in A1 or in A2 (FB A1/A2). This policy diverts patients when all the beds in A1 or in A2 are occupied.
7. Full Beds (FB). This policy diverts patients when all the beds in A1 and in A2 are occupied.
8. Myopic policy (Myopic): This policy diverts an arriving ambulance only if the expected waiting time for the current ambulance patient at the neighboring hospital is smaller than the expected waiting time if he/she is accepted. Thus, under the myopic policy, the ambulance is diverted only when $p_1^A W_1^D + (1-p_1^A)W_2^D \leq p_1^A W_1(n_1) + (1-p_1^A)W_2(n_2)$. Note that this heuristic evaluates $W_1(n_1)$ and $W_2(n_2)$ under the assumption that length of stay of a patient in the ED is exponentially distributed.
9. No AD Policy (No AD). This policy never diverts patients.

3.1 MDP Model Formulation

A simplified model of the ED presented in Section 2 is used to obtain an AD policy via MDP. For analytical tractability, the model assumes two arrival streams following homogeneous Poisson processes: ambulance arrivals with rate λ^A and walk-in arrivals with rate λ^W. Similar to the simulation model, ambulance patients are classified as level 1 with probability p_1^A and level 2 with probability $1-p_1^A$. Walk-in patients are classified as level 1 with probability p_1^W and level 2 with probability $1-p_1^W$. Patients with level 1 receive treatment in A1 where there are c_1 beds, and patients with level 2 receive treatment in A2 where there are c_2 beds. After waiting for an ED bed in the corresponding area, the patient stays in the ED for an amount of time referred to as length of stay, which is the sum of treatment time and boarding time. For tractability purposes of the MDP model, we assume that the length of stay in the ED is exponentially distributed with rates μ_1 and μ_2 for areas A1 and A2, respectively. After the stay in the ED, the patients are discharged.

The state of the system is represented by the total number of patients in A1 and A2, denoted as the tuple (n_1, n_2). Hence, the state space is given by $S = \{(n_1, n_2) : n_1 \geq 0, n_2 \geq 0\}$. The objective of the MDP model is to obtain a state-dependent policy to divert patients to a neighboring hospital that minimizes the long-run average expected waiting time over an infinite horizon.

Let $W_i(n_i)$, for $i \in \{1,2\}$, denote the expected waiting time of an arriving patient with level i given that there are n_i patients in Ai upon his arrival. Considering that the average length of stay in Ai is $1/c_i\mu_i$, we have that

$$W_i(n_i) = \begin{cases} \frac{n_i - c_i + 1}{c_i \mu_i}, & \text{if } n_i \geq c_i, \\ 0, & \text{if } n_i < c_i. \end{cases} \tag{1}$$

The MDP model also assumes that if an ambulance is diverted, the patient is transported to a neighboring hospital to receive appropriate treatment. The waiting time in the neighboring hospital is modeled as a

random variable X with a probability density function of $f(x)$. Let W_i^D be the expected waiting time of a diverted patient with level i for $i \in \{1,2\}$, we have that

$$W_i^D = \int_0^\infty x f(x) dx, \tag{2}$$

The continuous-time MDP model is converted to an equivalent discrete time model using uniformization with rate $\nu = \lambda^A + \lambda^W + c_1\mu_1 + c_2\mu_2$. Let v^* be the optimal long-run average expected waiting time and $h^*(n_1,n_2)$ denote the optimal relative effect of starting in state (n_1,n_2). The Bellman equation is given as

$$\begin{aligned} v^*\frac{\lambda^W+\lambda^A}{\nu} + h^*(n_1,n_2) = {} & \frac{\lambda^W p_1^W}{\nu}[W_1(n_1)+h^*(n_1+1,n_2)] + \frac{\lambda^W(1-p_1^W)}{\nu}[W_2(n_2)+h^*(n_1,n_2+1)] \\ & + \frac{\widetilde{c}_1\mu_1}{\nu}h^*(n_1-1,n_2) + \frac{\widetilde{c}_2\mu_2}{\nu}h^*(n_1,n_2-1) \\ & + \min\Big\{\frac{\lambda^A p_1^A}{\nu}[W_1^D+h^*(n_1,n_2)] + \frac{\lambda^A(1-p_1^A)}{\nu}[W_2^D+h^*(n_1,n_2)], \\ & \frac{\lambda^A p_1^A}{\nu}[W_1(n_1)+h^*(n_1+1,n_2)] + \frac{\lambda^A(1-p_1^A)}{\nu}[W_2(n_2)+h^*(n_1,n_2+1)]\Big\} \\ & + \Big(1 - \frac{\lambda^W+\lambda^A+\widetilde{c}_1\mu_1+\widetilde{c}_2\mu_2}{\nu}\Big)h^*(n_1,n_2), \end{aligned} \tag{3}$$

where

$$\widetilde{c}_i = \begin{cases} n_i & \text{if } n_i \le c_i, \\ c_i, & \text{if } n_i > c_i, \end{cases}$$

for $i \in \{1,2\}$.

The first two terms on the right hand side of Equation 3 refer to walk-in arrivals of patients with severity level 1 and 2, respectively. The third and fourth term refer to discharge of patients from A1 and A2, respectively. The terms inside the minimum expression refers to the potential actions upon an arriving ambulance. The first term inside the minimum statement represents the average waiting time if an ambulance arriving patient is diverted to a neighboring hospital; while the second term represents the average waiting time if the patient is accepted to the ED. The last term is a self loop due to uniformization.

A working paper presents theoretical properties of the optimal policy (Ramirez-Nafarrate et al. 2012). The optimal diversion policy is characterized by a monotonic threshold curve as illustrated in Figure 3. Above the threshold curve, the optimal action is to divert the patients; below the threshold curve, the optimal action is to accept the patients.

The AD policy based on the MDP model is obtained using the relative value iteration algorithm coded in C++ (Puterman 2005). The input parameters for the MDP model are obtained from the assumptions made in Section 2 in the following way:

- The arrival rates λ^W and λ^A are computed as the average arrival rates across the day.
- The length of stay assumes that the component regarding treatment time has rates of 0.25 and 1, for A1 and A2, respectively, as assumed in the simulation model. In addition, as assumed also in the simulation model, we considered that the average boarding time is 3.58 hours and the probabilities of admissions are 0.24 and 0.045 for patients with level 1 and level 2, respectively. Hence, the length of stay is assumed to be exponentially distributed with rates of 0.20 and 0.86 in A1 and A2, respectively.
- Regarding the waiting time in the neighboring ED, X is assumed to have a triangular distribution with parameters resulting from the average of the values of parameters across the day. This assumption resulted from a pilot study evaluating different options.

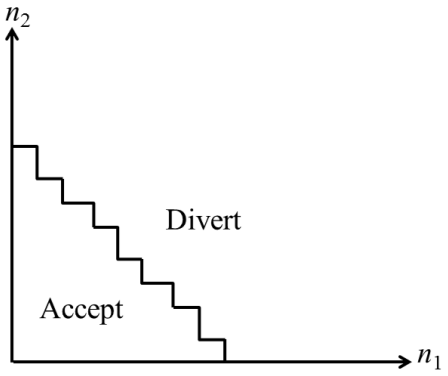


Figure 3: Illustration of the optimal diversion policy.

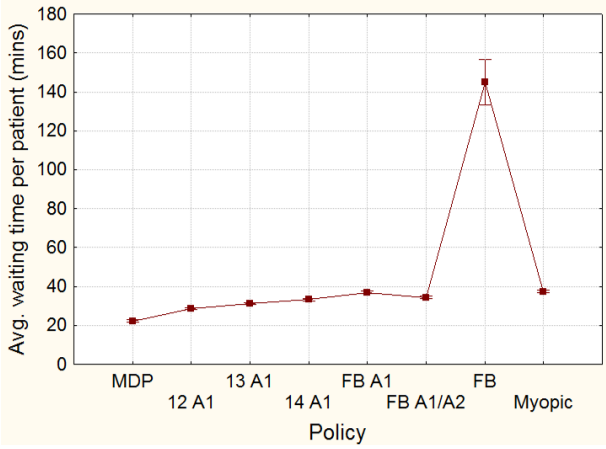
4 RESULTS

The results obtained during the experimentation show that not diverting at all, which is the No AD policy, produces average waiting time per patient of 1048.27 ± 215.11 minutes when the congested area is A1; and 28.85 ± 0.64 minutes when the congested area is A2. Note that the model does not include features such as patients leaving without being seen or time-dependent service rates, which may yield smaller average waiting times. Figure 4 presents the confidence intervals for settings 1 and 3 of X, given in Table 1. The results for setting 2 are not shown in this table because they fall somewhere between the results from settings 1 and 3.

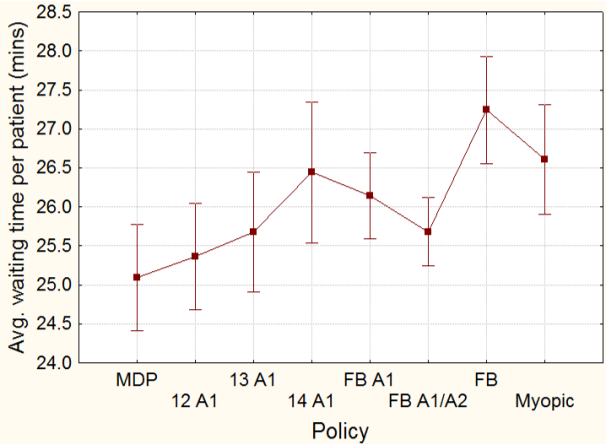
The results presented in Figure 4 show that the policy obtained by the MDP model is, at least, as good as other heuristics and in some cases it is significantly better. Since the MDP policy is obtained through a model that is a simplified version of the actual simulation mode, it does not guarantee to be the optimal policy. However, the performance of the MDP is superior to other heuristics. On the other hand, most of the heuristics perform significantly better than No AD, contradicting recommendations of the medical community. However, the benefits of ambulance diversion is subject to several conditions, including health conditions of ambulance patients, distance and traffic conditions for traveling to another ED.

The AD policies based on available capacity in A1 (i.e. FB A1, 14 A1, 13 A1, and 12 A1) also perform reasonable well compared with the MDP policy. This is due to the fact that the threshold policy, as illustrated in Figure 3, usually recommends saving few beds in A1 before going on diversion, especially if the congested area is A1 and the expected waiting time in the neighboring hospital is relatively small. As the expected waiting time in the neighboring hospital increases and congestion is observed in A2, the threshold policy recommends to observe full occupancy in A1 before diverting. In addition, the threshold policy usually allows some queueing in A2. This aspect also explains why the policy FB A1/A2 performs better if the congested area is A1 than if it is A2. If the congested area is A2, the policy FB A1/A2 diverts patients unnecessarily, increasing the average waiting time per patient. In a similar way, the policy FB performs much worse than the other policies if congested area is A1, because this policy delays going on diversion significantly. The Myopic policy performs significantly worse than the MDP, which implies that a good AD policy must observe the impact of the current decision on future patients.

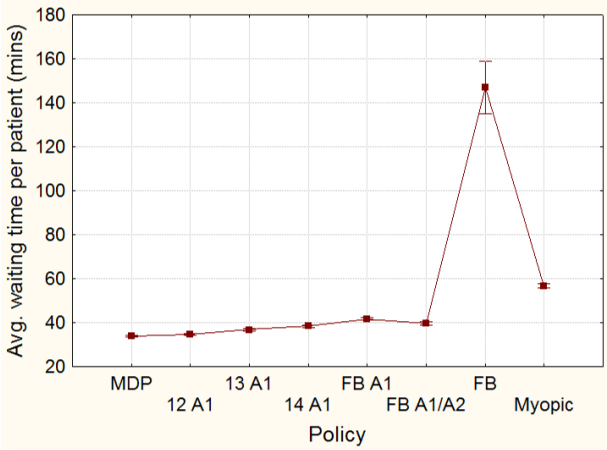
Another important result to highlight is the magnitude of the improvement on the average waiting time using AD, depending on which the congested area is. Table 2 shows the relative performance of the AD policies compared with the performance obtained by the policy obtained through the MDP formulation. It it is clear how AD improves the performance regarding the average waiting time per patient when the congested area is A1. This is due to the fact that most ambulance patients have severity level 1; therefore, AD is more effective handling congestion when A1 has a high utilization. In addition, AD policies that perform as good as the suggested by the MDP model can be identified when congested area is A2.



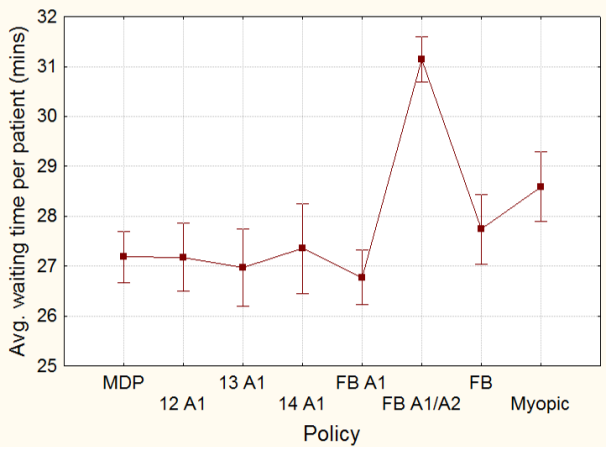
(a) Setting 1, Congested area: A1



(b) Setting 1, Congested area: A2



(c) Setting 3, Congested area: A1



(d) Setting 3, Congested area: A2

Figure 4: 95% confidence intervals on ETP.

A factor that is critical for obtaining an effective policy via MDP is to identify which the congested area is. The congested area depends not only on the arrival rate, but also on the severity mix. A wrong analysis of the congestion in the ED may lead to a highly ineffective policy. For instance, observe the relative performance of the FB policy. This policy performs almost as good as the MDP policy when congested area is A2, but if congested area is A1, then the performance of the policy is very poor.

In summary, the results show that the AD policy obtained using the MDP model performs better than other policies in many situations. However, given that the objective function in the MDP model does not penalize diverting ambulances, the MDP policy might require to spend a large fraction of time on diversion in order to minimize the average waiting time per patient. The fraction of time spent on diversion status is an important performance measure used by EDs. Table 3 shows the average fraction of time spent on diversion for the MDP policy.

Note that the average fraction of time spent on diversion does not depend on the setting for all the policies but for the MDP. This happens because the values of the parameters of the distribution of X affect the optimal policy; whereas for the other policies, the settings do not affect the time spent on diversion.

Table 2: Relative performance of AD policies compared with MDP policy (%). Bold numbers show statistical significance using 95% confidence level.

	Setting 1		Setting 2		Setting 3	
Policy	A1	A2	A1	A2	A1	A2
FB A1	**65.99**	**4.18**	**45.41**	-1.36	**23.36**	-1.50
14 A1	**49.01**	**5.37**	**31.86**	0.17	**13.40**	0.63
13 A1	**39.78**	2.32	**24.87**	-2.13	**8.77**	-0.79
12 A1	**27.65**	1.08	**15.55**	-2.51	**2.41**	-0.01
FB A1/A2	**54.24**	2.35	**36.69**	**3.98**	**17.76**	**14.57**
FB	**551.21**	**8.56**	**446.12**	2.55	**334.42**	2.05
Myopic	**67.64**	**6.03**	**69.18**	2.13	**67.80**	**5.19**
No AD	**4608.41**	**15.00**	**3827.98**	**7.86**	**2999.75**	**6.17**

It can be observed that the MDP policy may spend a high fraction of time on diversion, especially if the congested area is A1. However, as the waiting time in the neighboring hospital decreases, the percentage of time spent on diversion also decreases. Furthermore, the policies prescribed by the MDP seem to be on diversion less time than the other policies when the area A2 is congested. When the area A1 is congested, the MDP policy spends more time on diversion for low and moderate values of X.

Table 3: Average fraction of time spent on diversion.

Policy	Setting 1		Setting 2		Setting 3	
	A1	A2	A1	A2	A1	A2
MDP	0.9863	0.1590	0.7824	0.0944	0.6032	0.0477
FB A1	0.5915	0.1351	0.5915	0.1351	0.5915	0.1351
14 A1	0.6286	0.1790	0.6286	0.1790	0.6286	0.1790
13 A1	0.6727	0.2296	0.6727	0.2296	0.6727	0.2296
12 A1	0.7132	0.2941	0.7132	0.2941	0.7132	0.2941
FB A1/A2	0.3227	0.2353	0.3227	0.2353	0.3227	0.2353
FB	0.7732	0.2480	0.7732	0.2480	0.7732	0.2480

In order to reduce the fraction of time on diversion and at the same time reduce the average waiting time per patient, EDs must address the problem of capacity. The lack of capacity might be an issue upstream (in the ED) and/or downstream (in the inpatient units). In any case, a strategic plan is recommended in order to design a system (the whole hospital) that is capable of satisfying the demand in a reasonable time. In order to observe the impact of capacity, Figure 5 shows the average fraction of time spent on diversion and the average waiting time per patient varying the number of beds in A1 (c_1). This analysis is conducted for Setting 2 of X according to Table 1 when congested area is A1.

Results presented in Figure 5 show that a small change in the capacity of the ED for our model makes a significant impact in both performance measures. Therefore, for a given desired performance, an analyst could obtain the appropriate capacity given a limited budget. In addition, the analysis should be extended to inpatient units to reduce the impact of boarding patients in the congestion of the ED.

The policy obtained by the MDP model is characterized by a single threshold that recommends when to accept or divert patients. The single threshold may cause that the ED changes the diversion status very frequently. However, in practice, when EDs go on diversion they stay on that status for some amount of time. In order to ensure that the ED does not go on and off diversion often, the Bellman's equation shown in 3 should include a penalization for changing the diversion status.

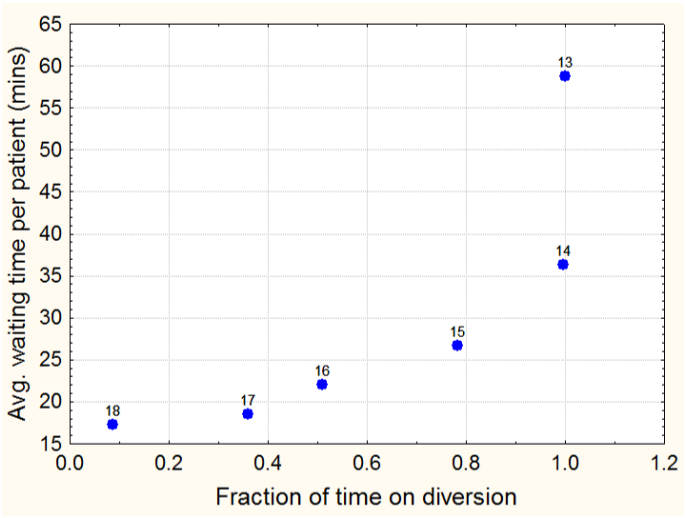


Figure 5: Average fraction of time on diversion vs average waiting time per patient for different number of beds in A1 and considering the MDP prescribed policy.

5 CONCLUSIONS

We compare the performance regarding average waiting time per patient of several AD policies and a simulation model with realistic input data is used to mimic the patterns and behaviors of several EDs around the US. The AD policies included in the study comprise several simple heuristics and a policy obtained by a Markov Decision Process model of a simplified version of the ED.

The results show that the policy prescribed by the MDP model performs significantly better than most of the heuristics. In particular, the MDP policy performs better when the emergent care area of the ED is more congested than the fast-track care area. Furthermore, most of the AD policies show a significant improvement on the average waiting time than not diverting at all. However, the results show than an intelligent design of the AD policies may lead to effective results. In order to identify the appropriate AD policies, it is required to identify which the congested area of the ED is. In addition, the model assumes that the waiting time in the neighboring hospital is known. Therefore, information sharing amongst hospitals that serve a common area is needed to ensure that the AD policies work effectively.

Nevertheless, the MDP formulation proposed in this paper does not penalize diverting ambulances, which causes that a large fraction of time might be spent on diversion. In order to improve the timeliness and accessibility of the emergency care system, measured with the average waiting time per patient and fraction of time on diversion, respectively, decision makers must address the issue of capacity. The analysis of capacity should include not only the number of resources required in the ED, but also in the inpatient units.

Future research about this topic includes redesigning the MDP formulation to include a penalization for being on diversion. In addition, other MDP objective functions will be explored to obtain an MDP policy that does not allow going on and off of diversion frequently, as it may happen with a single AD threshold. Other models that include the dynamics of neighboring hospitals will also be explored.

REFERENCES

Asamoah, O. K., S. J. Weiss, A. A. Ernst, M. Richards, and D. P. Sklar. 2008. "A novel diversion protocol dramatically reduces diversion hours". *American Journal of Emergency Medicine* 26 (6): 670–675.

Associated Press 2006. "Report: ER care in U.S. at 'breaking point"'. Available at: "http://www.msnbc.msn.com/id/13320317/ns/health-health_care/t/report-er-care-us-breaking-point/".

Banks, J., J. Carson II, B. Nelson, and D. Nicol. 2010. *Discrete-Event System Simulation*. Upper Saddle River, NJ.: Pearson Education, Inc.

Centers for Disease Control and Prevention 2006a. "Advanced Data from Vital and Health Statistics Saffing, Capacity, and Ambulance Diversion in Emergency Departments: United States, 2003-04". Available at: "http://www.cdc.gov/nchs/data/ad/ad376.pdf".

Centers for Disease Control and Prevention 2006b. "National Hospital Ambulatory Medical Care Survey: 2004 Emergency Department Summary". Available at: "http://www.cdc.gov/nchs/data/ad/ad372.pdf".

Centers for Disease Control and Prevention 2008. "National Hospital Ambulatory Medical Care Survey: 2006 Emergency Department Summary". Available at: "http://www.cdc.gov/nchs/data/nhsr/nhsr007.pdf".

Centers for Disease Control and Prevention 2010. "National Hospital Ambulatory Medical Care Survey: 2007 Emergency Department Summary". Available at: "http://www.cdc.gov/nchs/data/nhsr/nhsr026.pdf".

Cochran, J. K., and K. T. Roche. 2009. "A multi-class queuing network analysis methodology for improving hospital emergency department performance". *Computers & Operations Research* 36 (5): 1497–1512.

Deo, S., and I. Gurvich. 2011. "Centralized vs. Decentralized Ambulance Diversion: A Network Perspective". *Management Science* 57 (7): 1300–1319.

Green, L. 2006. "Queueing analysis in healthcare". *Patient Flow: Reducing delay in healthcare delivery*:281–307.

Hagtvedt, R., P. Griffin, P. Keskinocak, M. Ferguson, and F. Jones. 2009, December. "Cooperative strategies to reduce ambulance diversion". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1861–1874. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hoot, N., L. LeBlanc, I. Jones, S. Levin, C. Zhou, C. Gadd, and D. Aronsky. 2008. "Forecasting emergency department crowding: A discrete event simulation". *Annals of Emergency Medicine* 52 (2): 116–125.

Kelton, W., R. Sadowski, and N. Swets. 2010. *Simulation with Arena*. Columbus, OH.: McGraw-Hill.

Massachusetts Nurses Association 2009. "State's no diversion policy is putting strain on Massachusetts hospitals".

McConnell, K. J., C. F. Richards, M. Daya, S. L. Bernell, C. C. Weathers, and R. A. Lowe. 2005. "Effect of increased ICU capacity on emergency department length of stay and ambulance diversion". *Annals of Emergency Medicine* 45 (5): 471–478.

Patel, P. B., R. W. Derlet, D. R. Vinson, M. Williams, and J. Wills. 2006. "Ambulance diversion reduction: the Sacramento solution". *American Journal of Emergency Medicine* 24 (2): 206–213.

Puterman, M. 2005. *Markov Decision Processes Discrete Stochastic Dynamic Programming*. Hoboken, NJ: Wiley.

Ramirez, A., J. Fowler, and T. Wu. 2011, December. "Design of centralized ambulance diversion policies using simulation-optimization". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 1251–1262. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ramirez-Nafarrate, A., A. B. Hafizoglu, E. S. Gel, and J. W. Fowler. 2012. "Optimal Ambulance Diversion Control Policies". *Working Paper*.

Singer, A. J., H. C. Thode, P. Viccellio, and J. M. Pines. 2011. "The Association Between Length of Emergency Department Boarding and Mortality". *Academic Emergency Medicine* 18 (12): 1324–1329.

United States General Accounting Office 2003, March. "Hospital Emergency Departments: Crowded Conditions Vary among Hospitals and Communities". Available at: "http://www.gao.gov/new.items/d03460.pdf".

United States General Accounting Office 2009, April. "Hospital Emergency Departments: Crowding Continues to Occur, and Some Patients Wait Longer than Recommended Time Frames". Available at: "http://www.gao.gov/products/GAO-09-347".

Vilke, G. M., E. M. Castillo, M. A. Metz, L. U. Ray, P. A. Murrin, R. Lev, and T. C. Chan. 2004. "Community trial to decrease ambulance diversion hours: The San Diego County Patient Destination Trial". *Annals of Emergency Medicine* 44 (4): 295–303.

AUTHOR BIOGRAPHIES

ADRIAN RAMIREZ-NAFARRATE is a Professor in the Department of Industrial & Operations Engineering at Instituto Tecnologico Autonomo de Mexico. His research interests include modeling, simulation and analysis of healthcare delivery systems. He received a PhD degree in Industrial Engineering from Arizona State University, an MS in Manufacturing Systems at ITESM and a BS in Industrial Engineering at Universidad de Sonora. His email address is adrian.ramirez@itam.mx.

A. BAYKAL HAFIZOGLU earned his Ph.D. from the Department of Industrial Engineering at Arizona State University. His research focuses on stochastic modeling and optimal control of manufacturing and service systems. He earned his B.S. and M.S. degrees from Department of Industrial Engineering at Middle East Technical University in 2005 and 2007, respectively. His email address is baykal@asu.edu.

ESMA S. GEL is Graduate Program Chair and Professor of the Industrial Engineering program at the School of Computing, Informatics and Decision Systems Engineering at Arizona State University. Her re-search interests include applied probability, stochastic processes, queuing theory and stochastic modeling. Her email address is esma.gel@asu.edu.

JOHN W. FOWLER is Chair and Professor of the W.P. Carey Supply Chain Management Department at Arizona State University. His research interests include modeling, analysis, and control of manufacturing and service systems. He is a Fellow of the Institute of Industrial Engineers and is the SCS representative on the Board of Directors of the Winter Simulation Conference. He is an Area Editor of *Transactions of the Society for Computer Simulation International*, an Associate Editor of *IEEE Transactions on Semiconductor Manufacturing*, and Editor of *IIE Transactions on Healthcare Systems Engineering*. His email address is john.fowler@asu.edu.