Design and Analysis of Ambulance Diversion Policies

by

Adrian Ramirez Nafarrate

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2011 by the
Graduate Supervisory Committee:

John W. Fowler, Co-Chair
Tong Wu, Co-Chair
Esma S. Gel
Jorge Limon Robles

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

Overcrowding of Emergency Departments (EDs) put the safety of patients at risk. Decision makers implement Ambulance Diversion (AD) as a way to relieve congestion and ensure timely treatment delivery. However, ineffective design of AD policies reduces the accessibility to emergency care and adverse events may arise. The objective of this dissertation is to propose methods to design and analyze effective AD policies that consider performance measures that are related to patient safety.

First, a simulation-based methodology is proposed to evaluate the mean performance and variability of single-factor AD policies in a single hospital environment considering the trade-off between average waiting time and percentage of time spent on diversion. Regression equations are proposed to obtain parameters of AD policies that yield desired performance level. The results suggest that policies based on the total number of patients waiting are more consistent and provide a high precision in predicting policy performance.

Then, a Markov Decision Process model is proposed to obtain the optimal AD policy assuming that information to start treatment in a neighboring hospital is available. The model is designed to minimize the average tardiness per patient in the long run. Tardiness is defined as the time that patients have to wait beyond a safety time threshold to start receiving treatment. Theoretical and computational analyses show that there exists an optimal policy that is of threshold type, and diversion can be a good alternative to decrease tardiness when ambulance patients cause excessive congestion in the ED. Furthermore, implementation of AD

policies in a simulation model that accounts for several relaxations of the assumptions suggests that the model provides consistent policies under multiple scenarios.

Finally, a genetic algorithm is combined with simulation to design effective policies for multiple hospitals simultaneously. The model has the objective of minimizing the time that patients spend in non-value added activities, including transportation, waiting and boarding in the ED. Moreover, the AD policies are combined with simple ambulance destination policies to create ambulance flow control mechanisms. Results show that effective ambulance management can significantly reduce the time that patients have to wait to receive appropriate level of care.

DEDICATION

I dedicate this dissertation to:

My wife, Lupita for her love and support. Thanks also for taking this goal as her own.

My parents, David and Maria de los Angeles for loving me and supporting me in each step I have given through my life and for being an example of honesty and hard work.

My sister for looking after me, and my brothers for giving me their hands when I need them and allowing me to be part of their families.

ACKNOWLEDGMENTS

I want to thank all the people that supported me through these years at ASU. First, I am very grateful to Dr. John Fowler, not only for all his support and guidance, but also for trusting in me and giving me the opportunity to be involved in multiple projects and activities during my stay at ASU. I also want to thank my co-advisor Dr. Teresa Wu for her feedback and for looking after my progress. Many thanks to Dr. Esma Gel for supporting my research and for giving me challenges that made me learn a lot from her. Thanks to Dr. Jorge Limon for his advice and participation in this dissertation.

I also want to express my gratitude for those that provided financial support to accomplish this objective. Thanks to the Mexican Council of Science and Technology (CONACYT) for allowing me to study a PhD and work for the development of my country. Thanks to Dr. Askin and the School of Computing, Informatics and Decision Systems Engineering for allowing me to teach and contribute to our school. Thanks to Dr. Tim Lant and the Decision Theater for inviting me to collaborate in their projects.

Thanks to my friends in Arizona that supported me and who made my life more pleasant. Also for those who I worked and I learned with. Thanks to Baykal for his friendship and for his valuable contribution to the second chapter of this dissertation. To my friends: Billibaldo, Raul and his family, Abraham and his family, Marco and his family, Luis, Gerardo, Yasser, Jose, Giovanna, Hugo, Jimmy and his family, Serhat, Ozgur, Shanshan, Shao-Jen, Tao and Josh. Thanks to all for supporting me.

Thanks to my friends in Mexico for their friendship and supporting me through many years: Tomas, Jaime, Vacio, Sebastian, Karime, Maestro Fonseca, Memo and Jorge. Thanks to my family in law of taking care of me and my wife.

Finally, I want to sincerely thank to my family, whom I dedicate this dissertation. I am very proud to be part of their lives. Thanks to my wife, my parents and my siblings.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

National expenditures in healthcare in the United States reached 16.2 % of Gross Domestic Product in 2008, and it is projected to be 19.3% by 2019 according to the Centers for Medicare and Medicaid Services of the United States Department of Health and Human Services (2010). Despite increasing expenditures and advances in medical service and technology, several areas of the healthcare delivery system face major issues regarding the effectiveness, quality and safety of service (National Academy of Engineering and Institute of Medicine 2005). One of the areas with problematic performance is the Emergency Department (ED).

EDs are the most common place where unforeseen illness or injury are diagnosed and treated (American College of Emergency Physician 2007). The performance of EDs in the United States has long been discussed due mainly to the overcrowding and the resultant consequences, such as long waiting times, long periods on ambulance diversion, long boarding time periods, high rate of patients leaving without treatment and adverse events occurring on patients requiring emergency assistance (Fatovich and Hirsch 2003). A report from the United States General Accounting Office (2003) highlights congestion of EDs in the United States and relates the congestion to three main indicators: ambulance diversion (AD), patients leaving without treatment (LWOT) and high number of patients boarding. Regarding the first indicator, the definition found in the

document implies that hospitals request that ambulances bypass their facilities, transporting the patients whose original destination was that ED to another emergency facility.

AD statistics shown in the report suggests that nearly 70% of the EDs that took part in the study went on diversion status at some point during fiscal year 2001 and about 10% were on diversion at least 20% of the time (United States General Accounting Office 2003). The regions with longer AD periods correspond to highly populate metropolitan statistical areas (MSA). These conclusions are very similar to the analysis shown by Burt et al. (2006) where they suggest that there was an ambulance diverted every minute in the United States in 2003. The Centers for Disease Control and Prevention through their Advance Data from Vital and Health Statistics (now called National Health Statistics Report) reveals that in 2003-04 the mean annual hours on diversion was 242.7 per hospital. However, the mean annual hours on diversion among the EDs that reported any period of AD was 403.9; which is the equivalent of 16 entire days on AD status (Centers for Disease Control and Prevention 2006b). This same report indicates that 2.7% of the surveyed hospitals were on AD more than 20% of operating time and also suggests that large-size hospitals (large number of beds and high occupancy rate) tend to spend more hours on AD.

Some common factors that influence the decision to go on AD are the lack of inpatients beds, the high numbers of ED patients waiting, the complexity of ED cases and the high number of boarding patients (Centers for Disease Control and Prevention 2006b; Asplin 2003). This decision is usually made by nursing staff, a

hospital administrator or a medical director (Centers for Disease Control and Prevention 2006b) and it varies across different providers and even among different hospitals of the same stakeholder. Despite the impact of the AD decision in the public health context, there is a lack of quantitative assessment showing that decisions are made with effectiveness, quality and safety of service all. Therefore, it is necessary to analyze the pros and cons of diversion in order to take advantage of the benefits and reduce the risk caused by overcrowded EDs.

1.2 Literature Review

1.2.1 Analysis of AD from the medical perspective

The medical literature has a great number of publications discussing the causes of AD and its impact not only on the performance of the ED, but also on the health condition of diverted patients. A comprehensive review of AD and its effects show that AD is tightly related to ED crowding and its contributing factors, such as increased patient complexity and acuity, increased patient volume, inpatient bed unavailability, delays in the use of supporting equipment and even patient language and cultural barriers (Pham et al. 2006). This same review relates AD with other secondary causes, which includes the lack of specialty services, facilities and patient preferences. On the other hand, among the consequences of AD, Green (2008) indicates that for each additional hour of ambulance diversion, there is an increased mortality of about 3% of patients suffering from acute myocardial infarction. In general, a great proportion of the medical literature criticizes the use of AD as a solution to ED congestion because

of the risks incurred by longer transportation (American College of Emergency Physician 2008).

It is reasonable to think that diverting ambulances from EDs might not be a safe decision for the diverted patient, but under certain conditions, longer transportation time could be compensated by a shorter waiting time inside the facility of another ED. In this case, the patient might be seen by a doctor sooner in the new hospital than by being accepted into a saturated ED. Nevertheless, policy makers in some regions have passed laws that prohibit the use of AD. There is evidence that these "no ambulance diversion" policies adopted in some areas across the United States have put strain on the operations of crowded EDs. For instance, hospitals in Massachusetts have seen a rise in the waiting times of ED patients and a greater number of patients boarding in inappropriate areas after this type of policy was implemented in that state (Massachusetts Nurses Association 2009).

Therefore, AD deserves a discussion about its potential benefits and under what conditions they can be met. Interestingly, some researchers have already highlighted the importance of AD and concluded that AD deserves to be studied in a deeper manner in order to be taken into account by policy makers. Specifically, they suggest linking AD with clinical outcomes, patient and provider satisfaction, quality-of-life measures, economic measures and quality management initiatives (Asplin 2003; Williams 2006).

Currently, empirical studies analyzing AD on specific healthcare providers exists and the results and impact vary depending on their characteristics. For

instance, a 1999 study on a major hospital in Toronto, Canada revealed that AD increased with the number of patients boarding, the treatment time and the boarding time; however, authors did not find a relation between AD and staffing levels (Schull et al. 2003). Another study was conducted on an ED Level I trauma center that is part of a 400-bed acute care hospital; the findings include a reduction of 66% on hours of complete AD after an expansion of 67 beds in the ICU unit, implying the importance of inpatient bed availability in diversion performance (McConnell et al. 2005). This is particularly important given that ambulance patients are about three times more likely to need admission to the hospital than other types of ED patients (Burt et al. 2006).

Other studies have been made to reduce AD in systems comprising multiple hospitals. Vilke et al. (2004a) show a project for minimizing AD hours in a system comprised by five hospitals located in San Diego County. The AD guidelines for these hospitals were redesigned and healthcare providers were asked to avoid the use of ambulance bypass. Previous the application of the project, the total number of hours on diversion in the system per week was 112.2 hours. During the application phase, the hospitals were asked to avoid diversion; as a result, the total number of hours on diversion in the system decreased to 0.3 hours/week. An extended project in this area during a longer period reinforced the results, i.e. there as a significant decrease in the average monthly hours on diversion (Vilke et al. 2004b). A similar study was applied and a new AD protocol was introduced in a county of 600,000 people and 10 hospitals in Wisconsin. This protocol limited the hours on AD to only 1 out of every 8

operational hours. As a result, monthly AD hours decreased by 251 hours (Asamoah et al. 2008). These studies strongly suggest that AD can be reduced if the guidelines to go on AD are designed so that there is an incentive to first take other types of actions to reduce ED saturation. The literature does not provide precise information about how these healthcare organizations achieved those levels; however, it can be inferred that significant changes in the system were applied. Thus, providers are required to search for solutions or changes that relieve congestion by limiting the number of hours to spend on diversion or being stricter in the guidelines to implement it.

Another reason that could explain the substantial reduction in the multi-hospital systems is the reciprocating effect of AD among EDs in the same geographic region. It has been observed that if one hospital goes on diversion, the traffic to neighboring facilities increases, often causing other hospitals to go on diversion as well. Therefore, enforcing minimizing diversion episodes in one ED is expected to reduce the AD periods in the surrounding EDs (Vilke et al. 2004a).

In general, the medical community recommends avoiding diversion by using other mechanisms to relieve congestion; however, AD is still a practice used by a lot of emergency care providers. Therefore, it is important to analyze the settings where AD can achieve the best benefits possible; but having in mind that it is not a long-term solution and that changes in the health characteristics in the population demand more collaboration among providers (Lagoe and Jastremski 1990).

1.2.2 Analysis of AD from systems engineering perspective

EDs have been often subject of research in the last decade, especially because of the problems described earlier; therefore, the number of publications that models the ED system and looks at its performance is very extensive. Most of this literature relies on the analysis of waiting time, LWOT, and capacity or staffing planning; furthermore, the methods usually applied include queuing theory and simulation. However, AD still is a relative unexplored area. Reasons for this may include the complexity of the problem, the local characteristics of the system and the priority given to other types of problems and solutions. The analytical work done on AD includes the application of logistic regression to compare a designed work score based on the number of patients waiting and boarding to predict ambulance diversion (Epstein and Tian 2006). Kolker (2008) applies discrete-event simulation to analyze the relationship between AD and patients waiting for treatment in the ED with the length of stay, where it was found that reducing the length of stay (LOS) could significantly reduce AD percentage.

Queuing theory has been applied to model AD when diversion policies are based on the number of patients boarding (Allon et al. 20011). This research models the hospital as a 2 station process (ED and an inpatient unit) to develop two approximations (heavy traffic and fluid) that will explain the important structural characteristics of the hospital related to diversion performance. The observations made in this research include a method to identify the bottleneck in

the system, which can be the ED or the inpatient unit depending on the structural properties.

On the other hand, the issues and complexity of modeling the ED and the diversion policy using analytical methods, such as birth-death process and the potential of game theory, have been highlighted. These methods have been used to suggest the need of a regulator agent to incentivize and penalize to the hospitals, allowing the cooperation strategies among different emergency care providers (Hagtvedt et al. 2009).

In addition, Deo and Gurvich (2011) propose a queuing network approach to analyze the average waiting time for two hospitals in an emergency system. The authors found that a centralized design of diversion policies is Pareto improving compared to not diverting at all.

In summary, the existing literature that provides quantitative assessment to AD suggests that this action could bring benefits to the system and improve overall performance. However, the structure of optimal AD policies has not been deeply explored. Furthermore, methodologies for the appropriate design of AD policies and analysis of the effects of AD on ED performance are scarce and do not consider many unique characteristics of emergency care delivery systems.

This dissertation proposes methods based on simulation and optimization to design and evaluate the effectiveness of AD policies. Moreover, the models proposed capture important complexities and relations in emergency care systems and the performance metrics are directly related to patient safety and satisfaction.

1.3 Description of the Chapters and Contributions

Several methodologies for the design and analysis of AD policies are presented in the following chapters. Chapter 2 presents a methodology based on simulation for the analysis of mean performance and variability of AD policies. This methodology includes graphical and quantitative methods to analyze policies with periodic and continuous review of the state of the system. The policies are compared in terms of patient average waiting time and percentage of time spent on diversion. The analysis of mean performance is analyzed through bi-criteria plots and Integrated Preference Functional (IPF) measures, which assess the identification of the policies with best performance. The analysis of variability includes simultaneous confidence ellipses and computation of coefficient of determination to observe the consistency of different policies. This chapter contributes to existing literature in proposing a structure of single-threshold AD policies and analyzing the tradeoff between service and accessibility to emergency care. In addition, the chapter proposes an equation based on regression to determine the appropriate threshold on one state variable to go on diversion.

Chapter 3 introduces a Markov Decision Process model to optimize the long run average expected tardiness per patient using AD. Opposed to the manufacturing setting, the expected tardiness is defined as the time that patients wait beyond a recommended safety time threshold. The model includes important characteristics in patient mix, service times and the knowledge of the time to be seen in another hospital. Theorems and analysis of the structural properties of optimal policies are also presented. In addition, the chapter presents a simulation

model with relaxed assumptions and compares different AD heuristics. The results show that AD can decrease significantly the average time to start treatment, which may be translated to higher safety. The methods presented in this chapter are one of the first studies that compute and analyzes the structural properties of optimal AD policies. Therefore, this chapter contributes to the existing literature by proposing a model that determines the optimal policy of a hospital and that is robust enough to handle the main characteristics in EDs.

Chapter 4 proposes the combination of simulation and genetic algorithm to design the AD policies of multiple hospitals simultaneously. These methods overcome the scalability issue of the methods presented in the previous chapter. The chromosome of the genetic algorithm represents the parameters of the AD policies for each hospital in an emergency care delivery system. Two types of policies are explored: single-factor and multiple-factor AD policies. In addition, the AD policies are combined with ambulance destination policies that determine which hospital a patient should be transported to. The combination of diversion and destination policies acts like an ambulance flow control mechanism that allocates patient to appropriate ED. The objective is to minimize the time that patient spend in activities that delay receiving the appropriate level of care. These activities include transportation, waiting in the ED and boarding in the ED. This chapter contributes to existing literature by proposing methods that allows the simultaneous design of effective AD policies for multiple hospitals. Moreover, the performance metric referred as average-patient non-value added time and the

structure of the chromosomes for the genetic algorithm represent a different approach to traditional methods and metrics used in the evaluation of AD policies.

Finally, Chapter 5 presents the overall conclusions and the most significant findings are remarked. Additionally, future research opportunities are identified.

CHAPTER 2

BI-CRITERIA ANALYSIS OF SINGLE-FACTOR AMBULANCE DIVERSION POLICIES

2.1 Introduction

As stated in the previous chapter, AD has been highlighted as a concerning issues in emergency care in the United States and in other countries around the world. Several papers and articles identify the main causes that contribute to trigger the diversion status. Some of these causes include the lack of inpatient beds, the high numbers of ED patients, the complexity of ED cases and the number of boarding patients (Asplin 2003; Centers for Disease Control and Prevention 2006b). The decision of going on AD is usually made by nursing staff, a hospital administrator or the medical director of a hospital (Centers for Disease Control and Prevention 2006b) and it varies across different providers and even among different hospitals of the same stakeholder. An interesting question becomes how to quantitatively assess the effectiveness, quality and safety of the decisions. This chapter uses a methodology based on simulation and analyzes the performance of different diversion decision policies in bicriteria space. The criteria chosen represent two of the main performance indicators of EDs: the percentage of time the ED is on diversion and the patient average waiting time. These criteria imply a trade-off between the timeliness of the service and accessibility to emergency care. The rest of the chapter is structured as follows: Section 2.2 provides aspects found in the literature regarding AD, Section 2.3 presents the proposed study, starting by defining the problem and scope and

introducing the framework utilized to analyze AD policies, Section 2.4 describes the simulation model developed, the definition of the diversion policies and the experiments designed, Section 2.5 presents the results, Section 2.6 discusses the implications of AD and the potential application of the methodology and finally Section 2.7 provides conclusions and future research directions.

2.2 Literature Review

EDs have often been the subject of research in the last decade, especially because of the problems described earlier; therefore, the number of publications that model the ED system and look at its performance is quite extensive. Most of that literature has focused on patient waiting times, the number of patients leaving without treatment (LWOT), and capacity or staff planning. Some common methods applied include queuing analysis, systems dynamics and discrete-event simulation. For instance, queuing networks are used along with simulation to balance bed allocation in a large-size hospital (Cochran and Bharti 2006). In addition, queuing analysis is applied to redesign the service of EDs and to predict LWOT percentage based on traffic intensity (Roche and Cochran 2007; Cochran and Roche 2009; Broyles and Cochran 2007), as well to study the impact of a 4-hour discharge rule in EDs (Mayhew and Smith 2008) and to define required staffing levels (Green et al. 2006). Other types of mathematical models, such as integer programming have been used to analyze staffing problems in the ED (Carter and Lapierre 2001). On the other hand, simulation is widely used due to its flexibility to handle the complex dynamics of EDs. For example, discrete-event simulation is combined with data mining tools to analyze patient flows

given the variety in process requirements of the patients and to identify bottlenecks in the system (Ceglowski et al. 2007). Simulation is also employed as a tool to forecast overcrowding situations in the ED (Hoot et al. 2008), to evaluate modifications in the ED operations to avoid congestion given the prediction of arrivals (Meng and Spedding 2008), and to plan the implementation of changes in patient flow and buffer utilization to reduce waiting times (Wilhelm et al. 2008; Medeiros et al. 2008).

Nonetheless, quantitative assessment of AD has not been well studied. This is probably due to the complexity, subjectivity and localness of these decisions. However, there have been some initial attempts to address this issue. In (Kolker 2008), the length of stay of patients in AD is studied using simulation. It is concluded that the duration of the treatment of ED patients has a significant effect on the probability of going on AD. In (Ramirez et al. 2009a), distributed simulation is proposed to analyze the implementation of AD strategies in a regional healthcare delivery network. Simulation and design of experiments are integrated to analyze the performance of a large-size hospital, where it is found that the number of patients boarding and the ED configuration have a significant impact on the time spent on diversion (Ramirez et al. 2009b). Hagtvedt et al. (2009) highlights the complexity of modeling the ED and the diversion policy using birth-death processes and the potential of game theory to define an external agent that enables regulation of AD strategies between providers.

The medical literature includes publications discussing the impact of AD not only on the performance of the ED, but also on the health condition of the

patients and the capability of the emergency system to respond to emergencies. A review of AD and its effects shows that AD is tightly related to ED crowding and its contributing factors, such as increased patient complexity and acuity, increased patient volume, inpatient bed unavailability, delays in the use of supporting equipment and even patient language and cultural barriers (Pham et al. 2006). This same review relates AD with other secondary causes, which include the lack of specialty services, facilities and patient preferences.

Studies analyzing AD for specific healthcare providers show that results and impact vary depending on their characteristics. For instance, a study of a major hospital in Toronto, Canada during 1999 reveals that AD increased with the number of patients boarding, the treatment time and the boarding time; however, the association between AD and the staffing levels was not identified (Schull et al. 2003). The findings from another study conducted in an ED Level I trauma center that is part of a 400-bed acute care hospital include a reduction of 66% in the hours on diversion after an expansion of 67 beds in the ICU unit, implying the importance of inpatient bed availability in the diversion performance (McConnell et al. 2005). This is particularly important given that ambulance patients are about three times more likely to need admission to the hospital than patients arriving by other modes (Burt et al. 2006).

In summary, the importance of AD has been well positioned in the literature. However, there is a lack of guidelines for making AD decisions. Though it is known that decision makers consider factors such as patients waiting, boarding and inpatient bed unavailability, the efficacy of the AD decisions is not

known, i.e., the impact of the decisions on the quality of service. This chapter proposes a methodology using discrete-event simulation to design AD policies based on the important factors and studies ED performance using a bicriteria approach. The criteria chosen are two of the main indicators seen in EDs evaluations: patient average waiting time, which is related to quality and safety of service; and percentage of time spent on diversion, which is related to accessibility.

The average waiting time in the EDs of the United States was 55.8 minutes in 2006 (Centers for Disease Control and Prevention 2008). This statistic changes drastically depending on the region. For instance, one of the states that have suffered problems recently with the waiting time in EDs is Arizona, which is ranked 48 in the average time spent in the ED with 355 minutes (Press Ganey Report 2009). The percentage of time on diversion is the most common parameter to measure diversion performance. In the GAO report, the two most populous counties in Arizona were classified in the worst category regarding AD, having more than 25% of their hospitals on diversion more than 10% of the time (United States General Accounting Office 2003). Hence, both performance measures chosen for analysis are representative of common problems found in EDs across the United States, including the state of Arizona. Furthermore, waiting time is directly related to satisfaction of patients (Press Ganey Report 2009), which is an important consideration for the analysis of diversion policies (Asplin 2003).

2.3 Proposed Study

In this chapter, a framework involving three phases is proposed: (1) model development and design of experiments, whose objective is to construct the model and define the AD policies that will be studied, (2) the experimentation phase, which consists of executing the simulation model that collects information about the performance of the ED in the two criteria of interest and (3) analysis of results, which is divided in two parts: analysis of the mean performance and analysis of the variability. Figure 2.1 depicts the framework of the analysis and Sections 2.4 and 2.5 show the process in detail used in each phase.

2.3.1 Model development and design of experiments

This chapter constructs a model of a fictitious hospital that includes an ED and an Inpatient Unit that captures the dynamics and complexity of the emergency care system across the United States. The product of this part is a model whose data is a realistic example that allows the virtual implementation of AD policies.

The AD policies of interest in this chapter are those that look at a single factor to decide whether or not to go on diversion. Hence, the parameters of the policies included in this study are related to a threshold that triggers the diversion status and a threshold or time-window that enables the re-evaluation of the system and/or removal of the AD status. The factors considered in this chapter are commonly mentioned in the literature as causes to go on diversion in practice (Centers for Disease Control and Prevention 2006b; Pham et al. 2006). They are: the total number of patients waiting in the ED, the total number of patients boarding and the number of beds available in the Inpatient Unit. Once the AD

policies to be analyzed are defined, the next step is to design the experimentation process.

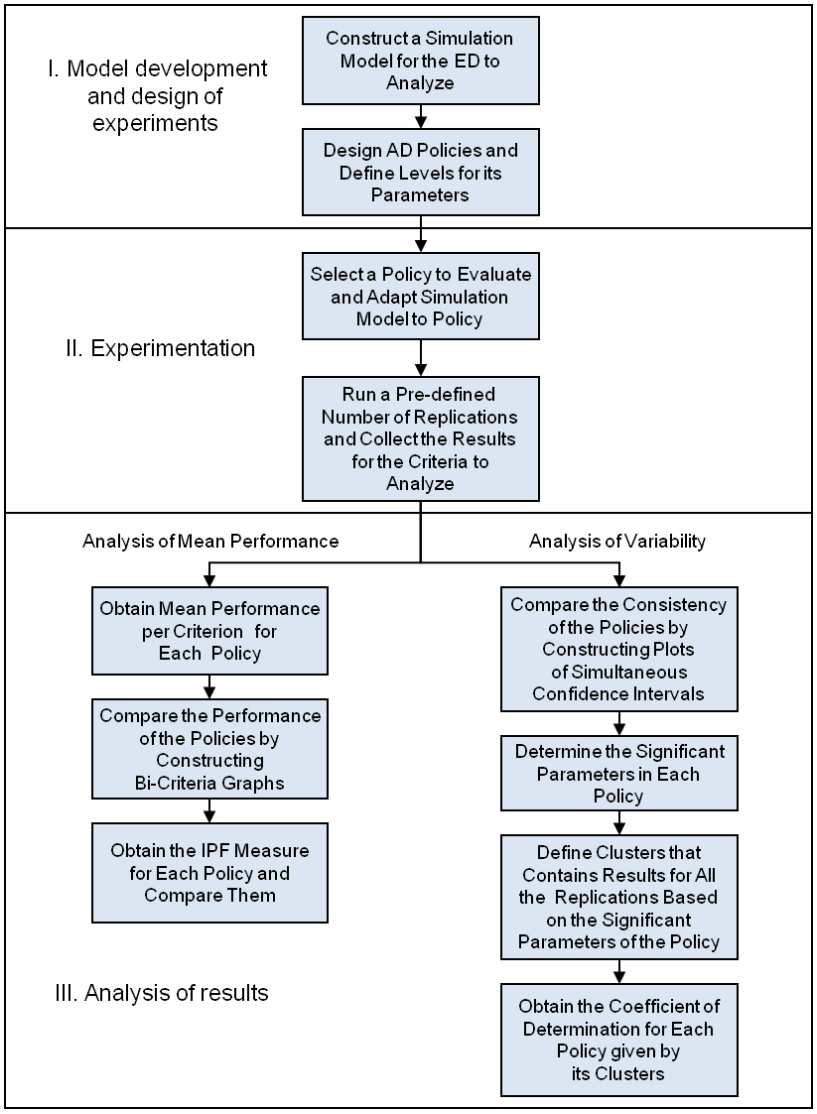

Figure 2.1. Framework proposed to analyze AD policies in bicriteria setting.

2.3.2 Experimentation

This phase consists of adapting the simulation model to the AD policy and the instances of the policies defined. The simulation model collects statistics regarding the average waiting time of the patients in the ED and the percentage of time spent on diversion.

2.3.3 Analysis of Results

This phase consists of the analysis of the results obtained from the execution of the simulation model under the treatments defined for each policy. The purpose is to compare and find the main differences across the policies and identify the policies that may offer more benefits or that are better aligned with the objectives of the provider. Therefore, the policies are evaluated in two aspects: the mean performance and the variability.

Analysis of Mean Performance

This part utilizes Pareto Analysis of the solutions obtained by each policy to compare the mean performance. The methods proposed include the construction of bicriteria graphs that enable visualization of the mean performance for each policy considered. In addition, it suggests the computation of a metric that allows the numeric comparison of solution sets. The process of the analysis of mean performance is given by the following steps:

1. Obtain the mean performance for the two criteria for every treatment in each policy. This implies obtaining the average for both criteria over all the replications.

2. Plot the mean performance for each treatment in bi-criteria space. This plot shows the behavior of the system under different policies.
3. Compare policies by plotting the set of non-dominated solutions per policy and computing the value of the Integrated Preference Functional (IPF). The IPF is a quantitative measure for comparison of the quality of the policies in terms of the distance of the non-dominated solutions from the ideal point and the impact of those solutions on the set of Pareto solutions of the policy (Carlyle et al. 2003; Bozkurt et al. 2010).

The application of these steps facilitates to observe the differences in the mean performance of the different policies in graphical and numerical ways. However, decision makers are also likely to be interested in analyzing the consistency of a policy in the long term. The second part of the analysis proposes methods to analyze the variability of the results of different policies.

Analysis of Variability

Similar to the analysis of the mean performance, the analysis of variability proposed allows graphical and numerical comparison of the variability obtained by the replications of the simulation model. The process of the analysis of variability is given by the following steps:

1. Plot simultaneous confidence ellipses for each policy, considering the results of all the replications. The shape and density of the ellipses show the consistency of the performance for each policy.
2. Apply regression analysis for each policy to determine the parameters of the policy that are significant for each criterion. The significant parameters should

be consistent with the observations made in the plot of the simultaneous confidence ellipses.

3. Define clusters that contain the results for all the replications depending on the significant parameters.

3. Obtain the coefficient of determination (R^2) (Montgomery 2005) for the clusters of every policy considering two types of centroids: the mean of the results of the cluster and the predicted response obtained from the regression equation. This process allows one to determine the consistency, but also to evaluate the prediction capability of the regression equation.

These steps enable the decision maker to observe graphically and numerically the consistency in the long-term of different policies. In addition, it also evaluates the accuracy of regression equations to predict the performance of a given policy. The application of the methodology is explained in more detail in Section 2.5.

2.4 Model Development and Implementation

The analysis of the impact of AD policies on waiting time requires building a robust model that considers the complexity of the system. This research proposes discrete-event simulation to perform this analysis given its flexibility to introduce arrival patterns, differences in acuity and length of stay and other factors contributing to the complex dynamics of EDs (Banks et al. 2010).

2.4.1 Simulation model

A model of a fictitious hospital, that contains the main elements of complexity, relations and flow of an ED, was created using discrete-event simulation. The simulation model was built in Arena version 12 (Kelton et al. 2007). Information regarding national averages and literature of healthcare providers from Arizona was used for the inputs (Cochran and Bharti 2006; Roche and Cochran 2007; Cochran and Roche 2009).

This hospital comprises an ED with 20 beds, and an Inpatient Unit with 78 beds. There are two arrival streams to the ED, which depend on the arrival mode. One stream belongs to ambulance arrivals and another to walk-ins. Before the ambulance patient enters the ED, the diversion status is observed. If AD is on, then the ambulance will be diverted, which is modeled by destroying the entity; otherwise, the patient enters the ED. The patients arriving by any mode are classified in one of five acuity levels (1-5), Level 1 being the most acute patients and Level 5 the least ill. If all the ED beds are occupied upon the arrival of a new patient, he/she will wait in a queue for being placed in a bed. The service discipline considered is based on priority given the acuity level. Therefore, patients of Level 1 receive the highest priority to be placed in a bed while patients of Level 5 have the lowest priority. If there is more than one patient of the same level, first come - first served is considered to assign beds. After concluding treatment time that depends on the acuity level, the patients can be admitted to the Inpatient Unit or be discharged.

The Inpatient Unit receives patients from direct admission arrivals and transfers from the ED. The treatment time depends on the source of the patient (external vs. ED). If an ED patient requires admission to the Inpatient Unit, but there is not any available bed, the patient will wait in the ED bed until a bed of the Inpatient Unit is released; this is defined as the boarding situation. After receiving treatment in the Inpatient Unit, the patients are discharged. Figure 2.2 depicts the logic of this model.

2.4.2 Data

Input data for this fictitious hospital was taken from national averages and from literature that models EDs. Since the arrivals to the ED represent an important factor for congestion, it is important to capture the dynamic nature of the arrivals usually seen in EDs. Several sources have highlighted the arrival pattern to EDs across the United States (Centers for Disease Control and Prevention 2008; Cochran and Roche 2009; Green 2006; Miller et al. 2009).

Therefore, the mean arrival rate to the ED being modeled depends on the time of the day and on the arrival mode. In this chapter, it is assumed that arrivals behave according to Poisson processes whose rates are based on a pattern seen in a real ED in Arizona (Cochran and Roche 2009).

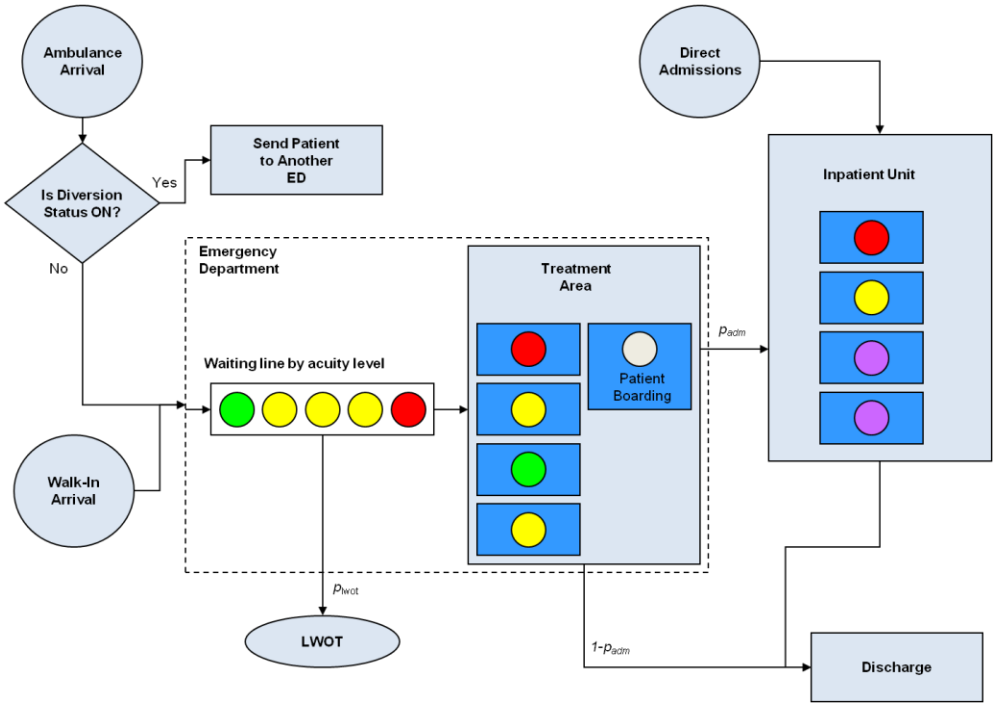

Figure 2.2. Logic of the model.

Figure 2.3 shows the mean arrival rate to the ED. It can be implied from this figure that average arrival rate to the ED is 6.4 patients/hour and that ambulance arrivals represent 15% of all the ED arrivals, which is consistent with national average of 15.4% (Centers for Disease Control and Prevention 2008). The use of Poisson process to represent the arrivals to EDs is a reasonable assumption given its properties, according to a discussion provided by Green (2006).

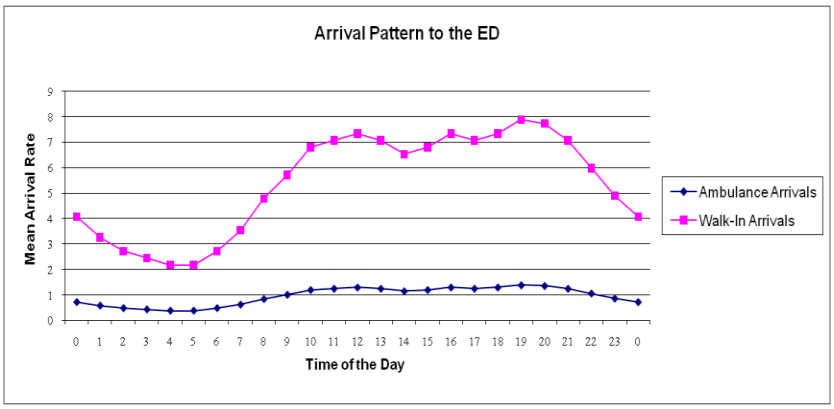

Figure 2.3. Mean arrival rate to the ED.

Upon arrival, patients are classified according to their acuity level based on percentages shown in Table 2.1. This data is also based on information published by an analysis of a local provider (Cochran and Roche 2009).

Table 2.1. Percentages of acuity mix.

	Arrival Mode		
Acuity Level	Ambulance	Walk-Ins	Overall
1	15	2	3.95
2	42	16	19.90
3	30	40	38.50
4	10	30	27.00
5	3	12	10.65
Overall	15	85	

Treatment time in the ED beds depends on the acuity level and it is assumed that it has an exponential distribution with mean shown in Table 2.2. These times are based on the same paper used to obtain arrival pattern (Cochran and Roche 2009). In addition, the average treatment time is similar to data found in other literature (Centers for Disease Control and Prevention 2006a).

Table 2.2. Mean treatment times in the ED.

Acuity Level	Mean Treatment Time (min)
1	261
2	261
3	162
4	90
5	30

After receiving treatment in the ED, patients can be admitted to the hospital or be discharged. Probability of admission to the Inpatient Unit for an ED patient is assumed to be 15%, which is in the range of admissions commonly seen in literature (Centers for Disease Control and Prevention 2006a; Center for Disease Control and Prevention 2008). Patients requiring admission to the Inpatient Unit are transferred only if there is an available inpatient bed; otherwise they keep occupying the ED bed until transfer is made, thus the patient is boarding until a bed in the Inpatient Unit is released.

Information regarding the Inpatient Unit considers data from another paper of a provider in Arizona (Cochran and Bharti 2006), which models a whole hospital with thirteen units. The Inpatient Unit receives admissions from the ED but also direct admissions, whose time between arrivals are exponentially distributed with a mean of 10 hours. Treatment time in the Inpatient Unit also has an exponential distribution with a mean that depends on the source of the patient as shown in Table 2.3. Patients are discharged after receiving treatment in the Inpatient Unit.

Table 2.3. Mean treatment times in the Inpatient Unit.

Patient Source	Mean Treatment Time (hrs)
ED	80
Direct Admission	70

Patients leaving without treatment are a feature difficult to model. Papers that have studied LWOT found that it is difficult to determine the time that the patient leaves because ED administration is not informed of this decision; furthermore, methods to predict LWOT percentage are quite complex (Broyles and Cochran 2007). Nevertheless, it is important to include LWOT since those patients contribute to congestion while they are in the ED. This chapter uses the approach used in (Miller et al. 2009) where patients leave if they have not been placed on a bed after 24 hours upon their arrivals. Patients with acuity Level 5 are most likely to be affected under this scheme because they have the lowest priority to receive a bed if there are other patients in the system. Actually, the standards regarding patient classification recommend that a patient Level 5 should be treated between 2 and 24 hours after arrival (Centers for Disease Control and Prevention 2008), so this approach can be used also to measure the compliance with this guideline.

2.4.3 Design of Ambulance Diversion Policies

The Advance Data from Vital and Health Statistics report from September 2007 highlights the main reasons for going on diversion during 2003-04; the first place was the lack of inpatient beds and the second place was the (high) number of ED patients (Centers for Disease Control and Prevention 2006b). On the other hand, causes which have less importance include the complexity of the ED cases,

shortage of hospital and ED staffing and equipment failure. Other sources agree that these factors are the main contributors of AD, however, the number of patients boarding is also highlighted in some studies (Asplin 2003; Ramirez et al. 2009b; Pham et al. 2006). Based on this information, diversion policies are considered based on one of these single factors. For this purpose, the following variables are defined:

x: total number of patients waiting for a bed in the ED, $x = 0, 1, 2, 3, \ldots$

y: total number of patients boarding in the ED, $y = 0, 1, 2, 3, \ldots, B_{ED}$.

z: number of beds available in the Inpatient Unit, $z = 0, 1, 2, 3, \ldots, B_{IP}$.

where,

B_{ED}: number of beds in the ED. In this case $B_{ED} = 20$.

B_{IP}: number of beds in the Inpatient Unit. In this case $B_{IP} = 78$.

The AD policies considered have the form: (*Don*, *Doff*), where *Don* is the threshold to set the diversion status on and *Doff* is the criterion to reevaluate or remove the AD status. The basic forms of the six policies studied in this chapter are presented in the following list:

a) Form of policies P1, P3 and P5: ($U.$, Δt).

Here, $U.$ is a threshold on a state variable of interest to go on diversion. The state variable is x for P1, y for P3 and z for P5. Hence, diversion status is set on if at some point $x > U_x$ for P1, $y > U_y$ for P3 or $z > U_z$ for P5. Once the ED has gone on diversion, the state of the system will be evaluated every Δt time units, until the decision to go off diversion is made. Diversion status will be removed when the state variable is smaller than the $U.$ threshold.

b) Form of policies P2 and P4: ($U.$, $L.$)

Similar to the previously defined policies, $U.$ is the upper threshold on the state variable of interest to go on diversion (x for P2 and y for P4). On the other hand, $L.$ is the lower threshold on state variable of interest to remove the diversion status. Hence, diversion is set on in a similar way to P1, P3 and P5; and it is removed as soon as the state variable is smaller than the $L.$ threshold.

c) Form of the policy P6: (Lz, U_z).

Since policy P6 is based on the number of available beds in the Inpatient Unit, the lower and upper thresholds defined for policies P2 and P4 are inverted for policy P6. Thus, diversion is set on as soon as $z < L_z$ and will be removed when $z > U_z$.

It can be seen that P1, P3 and P5 imply a periodic review of the state of the system after the decision of going on diversion is made. On the other hand, P2, P4 and P6 imply a continuous review to remove the diversion state. In addition, these policies require that $U. > L.$ Complete diversion is considered in this chapter. Thus, all ambulances that were supposed to arrive to the hospital will be diverted if the ED has the AD status on, regardless of the acuity level of the patient being transported.

2.4.4 Experimentation Design

The model described in Section 2.4.1 is used to run experiments based on different levels of the *Don* and *Doff* parameters of the six policies defined in Section 2.4.2. For every policy, (*Don*, *Doff*) levels are set, based on the scale of the model while trying to cover a large range of possible values of the parameters.

Then, the simulation model is adapted to set and remove the AD status depending on the policy. The chosen levels of the policy parameters are shown in Table 2.4.

Table 2.4. Levels of the policy parameters used in experimentation.

Policy	*Don*	*Doff*
P1	10, 20, 30, 40, 50, 60, 70 patients	15, 30, 45, 60 minutes
P2	10, 20, 30, 40, 50, 60, 70 patients	0, 10, 20, 30, 40, 50, 60 patients
P3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10 patients	15, 30, 45, 60 minutes
P4	1, 2, 3, 4, 5, 6, 7, 8, 9, 10 patients	0, 1, 2, 3, 4, 5, 6, 7, 8, 9 patients
P5	0, 1, 2, 3, 4, 5, 6, 7, 8, 9 beds	15, 30, 45, 60 minutes
P6	0, 1, 2, 3, 4, 5, 6, 7, 8, 9 beds	1, 2, 3, 4, 5, 6, 7, 8, 9, 10 beds

The levels include very conservative policies and policies that are not so conservative. For instance, *Don* levels for P1 and P2 include scenarios where the ED will go on diversion if there are only ten patients waiting and others where there are at least seventy. P3 and P4 consider scenarios which trigger the AD status as soon as one patient is boarding or wait to see up to ten patients. Similarly, P5 and P6 comprise scenarios of going on AD if there are nine beds available in the Inpatient Unit or going on AD only when there is not any inpatient bed available. The experimentation is based on a factorial design where each permissible combination of *Don* and *Doff* levels is used as a treatment (Montgomery 2005).

Forty replications are run for each treatment using antithetic random numbers (Law 2007), which produces 20 observations per treatment. Each replication collects information for ten thousand processed ED patients and it includes a warm up period of three weeks.

2.5 Analysis of Results

This phase analyzes the performance of the ED under the different policies. Of special interest is the study of the effect of the three factors on the responses by constructing bi-criteria graphs, which show the percentage of time spent on diversion on the horizontal axis and the average waiting time of the patients admitted to the ED on the vertical axis. In order to evaluate the trade offs of AD, the average waiting time in the ED is obtained if AD is not implemented is 1.8 ± 0.1 hours (95% confidence interval).

2.5.1 Mean Performance of Policies

This section shows the mean performance of the six policies in two ways. First, a brief analysis of the performance across the six policies is made by locating the solutions of each policy in a bi-criteria space. Then, pair-wise comparisons are made to study the performance of policies with a common factor in the *Don* parameter. Thus, policy P1 is compared to P2, P3 is compared to P4, and P5 is compared to P6. The analysis of every policy pair includes a plot that allows the visualization of the policy performance and a metric that is used to compare numerically the quality of solutions sets. This metric is called Integrated Preference Functional (IPF).

The IPF measure was first proposed by Carlyle et al. (2003) and extended in 2010 (Bozkurt et al. 2010). IPF provides a robust quantitative measure of the quality of a solution set. In addition, IPF takes into account several characteristics of the set in a single value, such as coverage, uniformity and cardinality.

Therefore, AD policy makers are able to observe the form of the best policies for their facilities through IPF comparison of policies.

The computation of the IPF(*P.*) value for policy *P.* used in this chapter utilizes a weighted Tchebycheff function of its set of non-dominated solutions. IPF for each policy is calculated as follows:

$$IPF(P.) = \sum_{i \in I} \left(\int_{\alpha_L^i}^{\alpha_b^i} h(\alpha)(1-\alpha) f_2^i d\alpha + \int_{\alpha_b^i}^{\alpha_U^i} h(\alpha)\alpha f_1^i d\alpha \right) \qquad (2.1)$$

where,

I: Set of non-dominated solution of policy *P*.

f_1^i : Value of criterion Z_1 (percentage of time on diversion) for solution *i*

f_2^i : Value of criterion Z_2 (average waiting time) for solution *i*

α : Weight given to criterion Z_1

$h(\alpha)$: Density function of the weight

IPF formulations exist for convex combination of criteria (Carlyle et al. 2003) and also for Tchebycheff function (Bozkurt et al. 2010). The second option has been chosen given that it enables including unsupported points in the comparison of nondominated solutions. Thus, IPF computation of a policy includes all the Pareto solutions, not only those that define the efficient frontier.

The density function of the weights can be seen as the probability of the preference of the policy maker for the weight of each criterion (i.e. uniform, triangular, etc.) (Carlyle et al. 2003). This chapter assumes a uniform density function for the weights, i.e. decision maker cares equally about the weights across the range of α values. However, IPF can be adapted to the preference of

the decision maker. Properties of IPF values state that a set that dominates another set will have a smaller IPF value than the dominated set. In addition, adding nondominated solutions to a set will never increase the IPF value (Carlyle et al. 2003). These properties imply that the smaller the IPF value is among different policies, the better that policy is. Readers are referred to (Bozkurt et al. 2010) for a full description of the steps to compute IPF.

One issue regarding the computation of IPF is its sensitivity to large differences in scale. This is due to the potential nullification of one criterion by another. The results of this chapter show a significant difference between the scales of the two criteria chosen to study. Hence, the computation of IPF was actually applied to scaled data obtained by the application of Equation (2.2).

$$\left(g_1^i, g_2^i\right)=\left(\frac{f_1^i-\min_{i\in I}\left(f_1^i\right)}{\max_{i\in I}\left(f_1^i\right)-\min_{i\in I}\left(f_1^i\right)}\right),\left(\frac{f_2^i-\min_{i\in I}\left(f_2^i\right)}{\max_{i\in I}\left(f_2^i\right)-\min_{i\in I}\left(f_2^i\right)}\right) \qquad (2.2)$$

where,

(f_1^i, f_2^i): Non-scaled criteria values of non-dominated solution i.

(g_1^i, g_2^i): Scaled criteria values of non-dominated solution i.

$\min_{i\in I}(f_j^i)$ [$\max_{i\in I}(f_j^i)$]: minimum [maximum] $f_\cdot^i$ among all $f_\cdot^i$`s in competitive sets of Pareto optimal solutions used to scale the data.

2.5.1.1 Overall comparison

The mean performance considering all treatments across the six policies is shown in Figure 2.4. It can be seen that policies considering the different factors are located along a band whose characteristics resembles a disjointed convex line. This band exhibits solutions of the six policies, but the factor that is used in the policy determines the location range in the band. For instance, policies P1, P2, P3 and P4 have solutions in the first half of the band; their average waiting time vary from 1.75 hours to about 0.75 hours, while the percentage of time spent on diversion varies between 1% and 25%. On the other hand, policies P5 and P6 are located in the second half of the band, producing average waiting time that can vary between 0.5 and 0.75 hours, but also causing a large proportion of time spent on diversion that can go from 25% up to nearly 60%.

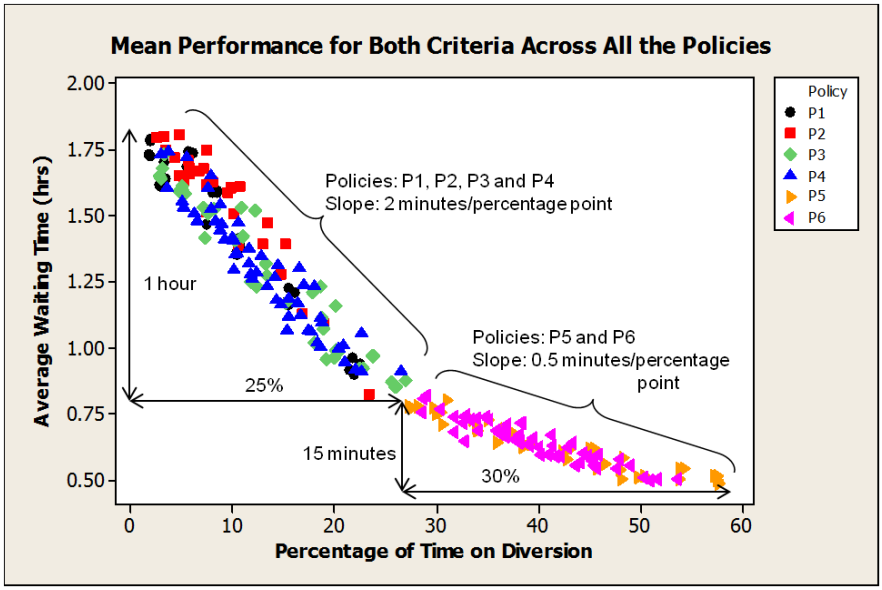

Figure 2.4. Union of solutions of all policies.

The convexity and change in the slope of the band suggest that reduction of average waiting time is more significant for small percentages of time on diversion. Therefore, the first half of the efficient band implies that using P1, P2, P3 or P4 can achieve a reduction of about 2 minutes in the average waiting time per every percent point increased in the time spent on diversion; on the other hand using P5 or P6, the reduction is only about half a minute.

Therefore, it is important to see the difference of the results that each policy can achieve. For instance, policies P1 to P4 produce results that are in the range that decision makers are most likely interested in, because they typically do not compromise accessibility as much as policies P5 and P6. In addition, the reduction of waiting time is more significant as stated previously; however, it is necessary to do a deeper analysis to look at the differences among these four policies, which is done in the following sections.

The global comparison of policies enables one to explain why some healthcare providers fail to reduce their time spent on diversion or what could happen if an AD policy is designed such that one of the three main factors studied in this chapter have a larger weight to decide when to go on diversion. For instance, the Advance Data from Vital and Health Statistics report from September 2006 mentioned that near 12% of hospitals located in metropolitan areas spent between 5 and 19% of time on diversion status and about 2.7% spent more than 20% of their time on diversion. Furthermore, the most frequent reason to go on diversion was the lack of inpatient beds (Centers for Disease Control and Prevention 2006b).

2.5.1.2 Policies based on number of patients waiting in the ED

The first two policies defined trigger the diversion status based on the total number of patients waiting in the ED. Figure 2.5 (*top*) shows the mean performance of these policies across all the treatments designed with the appropriate thresholds. The figure shows the results grouped by the *Don* level. Note that the number of treatments per group in P2 varies because of the condition that *Don* > *Doff*. Hence, according to the values defined in Table 2.5, if $U_x = 10$ then there is only one option for L_x, which is $L_x = 0$.

Interesting observations can be made from this figure. First, it is evident in P1 that the results are clusters for the same level of the *Don* parameter. This effect implies that the percentage of time spent on diversion and the reduction of the waiting time depends primarily on the threshold chosen to trigger the diversion state, at least in the range of reasonable values of *Doff* like those set in the experimentation.

Comparing the performance of periodic review of the state once AD has been set (P1) versus continuous review (P2), a similar performance on the average waiting time can be seen, especially in the lower range of the time on diversion. However, the percentage of time spent on diversion is greater in P2 policies than their counterpart in P1 and the variation of the performance for the policies with the same *Don* level is smaller in P1.

Note that the most conservative policy instance in these graphs is when $U_x = 10$, where a percentage of time on diversion between 20% and 25% is observed. However, in reality the most conservative policy would be to set the diversion

status when one patient is seen waiting or as soon as all the ED beds are occupied. In that case, it is expected that the percentage of time on diversion will reach a larger fraction.

The IPF value was obtained to numerically analyze the different policies. The first process to obtain the IPF value was to scale the data because the method is sensitive to large differences in the scale of the criteria. Since policies P1 to P4 have results that are very similar to each other, these results were used to scale and compare the IPF among them.

Figure 2.5 (*bottom*) shows the nondominated solutions for both policies and their respective IPF values. It can be observed graphically and numerically that the set of solutions produced by P1 has better characteristics than solutions of policy P2. The set P1 is never intersected by the set P2 in the range of the first, which implies that P1 can be used to obtain a desired level of average waiting time with a lower percentage of time spent on diversion than using P2. Consequently IPF(P1) < IPF(P2), which reinforces the superiority of P1 over P2.

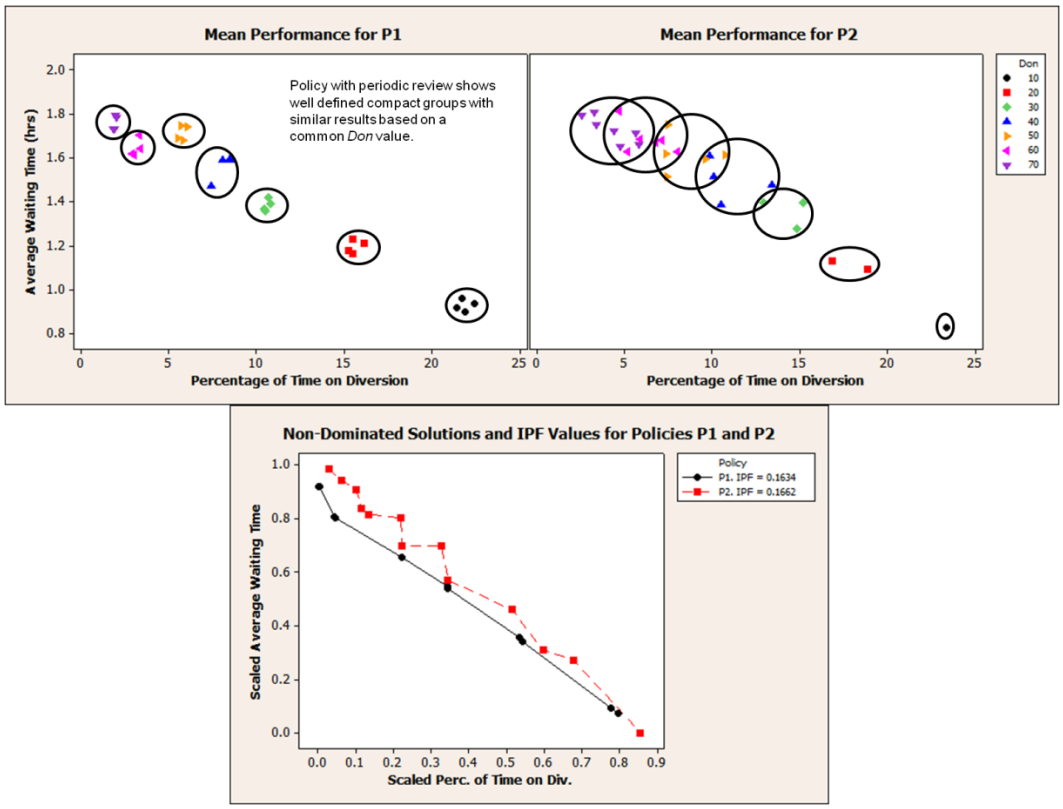

Figure 2.5. Mean performance of P1 and P2 (*top*) and nondominated solutions and IPF value per policy (*bottom*).

2.5.1.3 Policies based on number of patients boarding in the ED

The mean performance of policies P3 and P4 are shown in Figure 2.6 (*top*). The clustering of results depending on the *Don* parameter is observed again in P3. However, results in P4 have greater variation causing the clusters to overlap.

Note that the most conservative policy of these types is observed when diversion state is triggered as soon as one patient is seen boarding. This implies that the maximum percentage of time on diversion that P3 and P4 can achieve is about 26% and a minimum average waiting time of 0.87 hours. On the other hand,

the maximum U_y threshold considered is 10, obtaining average waiting time of 1.68 hours and 3% of the percentage of time on diversion.

Figure 2.6 (*bottom*) shows the nondominated solutions and the IPF values for policies P3 and P4 after scaling the solutions considering policies P1 to P4. It can be seen that both policy sets intersect each other, making difficult to observe what type of review configuration produces better results. However, the smaller IPF value of P4 suggests that this policy produces results with better characteristics than P3; furthermore, IPF also suggests that solutions belonging to P4 have the best characteristics of the first four policies. This is because the solutions produced by P4 have larger cardinality and coverage than the other policies.

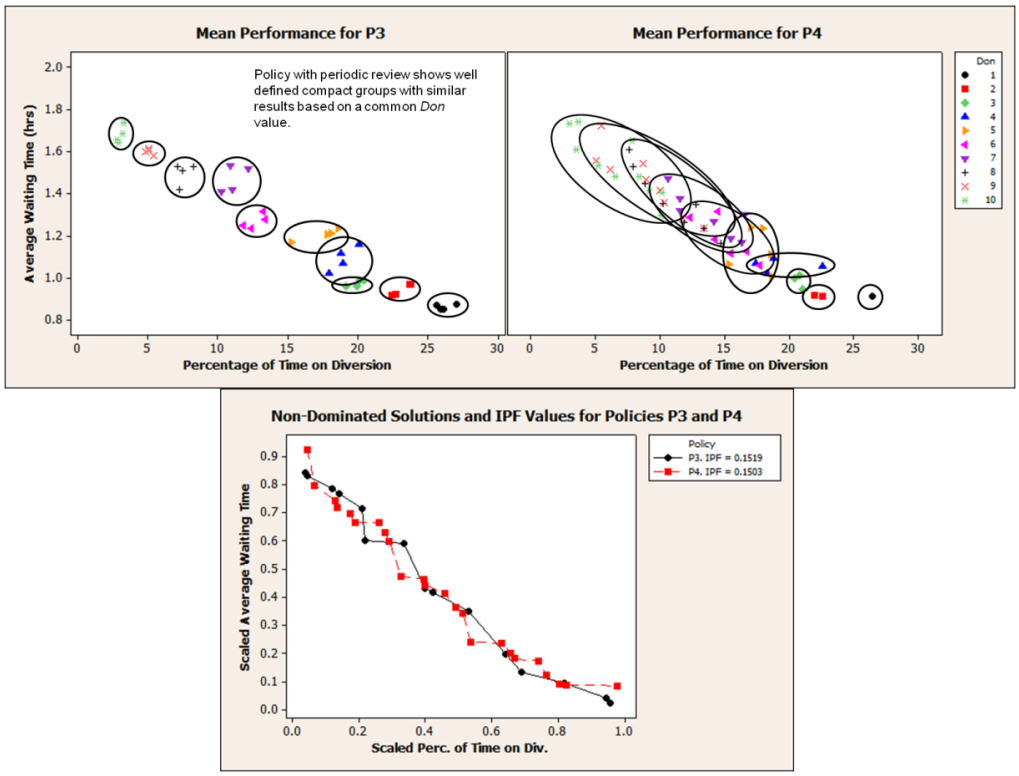

Figure 2.6. Mean performance of P3 and P4 (*top*) and nondominated solutions and IPF value per policy (*bottom*).

2.5.1.4 Policies based on number of beds available in the Inpatient Unit

The last set of policies to analyze are P5 and P6, whose mean performance are shown in Figure 2.7 (*top*). Clustering according the *Don* parameter can be observed again, especially in the case of periodic review (P5). These policies are more conservative since they are based on the number of beds available in the Inpatient Unit. For the instances being analyzed, the most conservative policy triggers the diversion status when there are nine beds available in the Inpatient Unit. On the other hand, the least conservative policy is obtained when diversion is set if all the beds are occupied. For the ED modeled, these policies achieve a minimum of percentage of time of diversion of almost 30%. These policies can reduce the average waiting time to less than 45 minutes, but the accessibility is much compromised.

IPF values for these policies were obtained scaling the data for the six policies so the metric could capture the increased proportion of the time spent on diversion. IPF shows that set of solutions belonging to P5 have better characteristics than solutions of P6. Both sets intersect each other, but P5 has a greater coverage.

The IPF metric suggests that policies P1 to P4 are very competitive as shown in Table 2.5, having better characteristics than policies P5 and P6. Therefore, decision makers should be careful when implementing AD policies primarily based on the inpatient available capacity. Furthermore, IPF also allows observing that policies based on boarding patients might produce solutions with higher quality than policies based on number of patients waiting; however, the

variability of the results should be also considered when designing an AD policy. The next section studies the difference in variability across the policies.

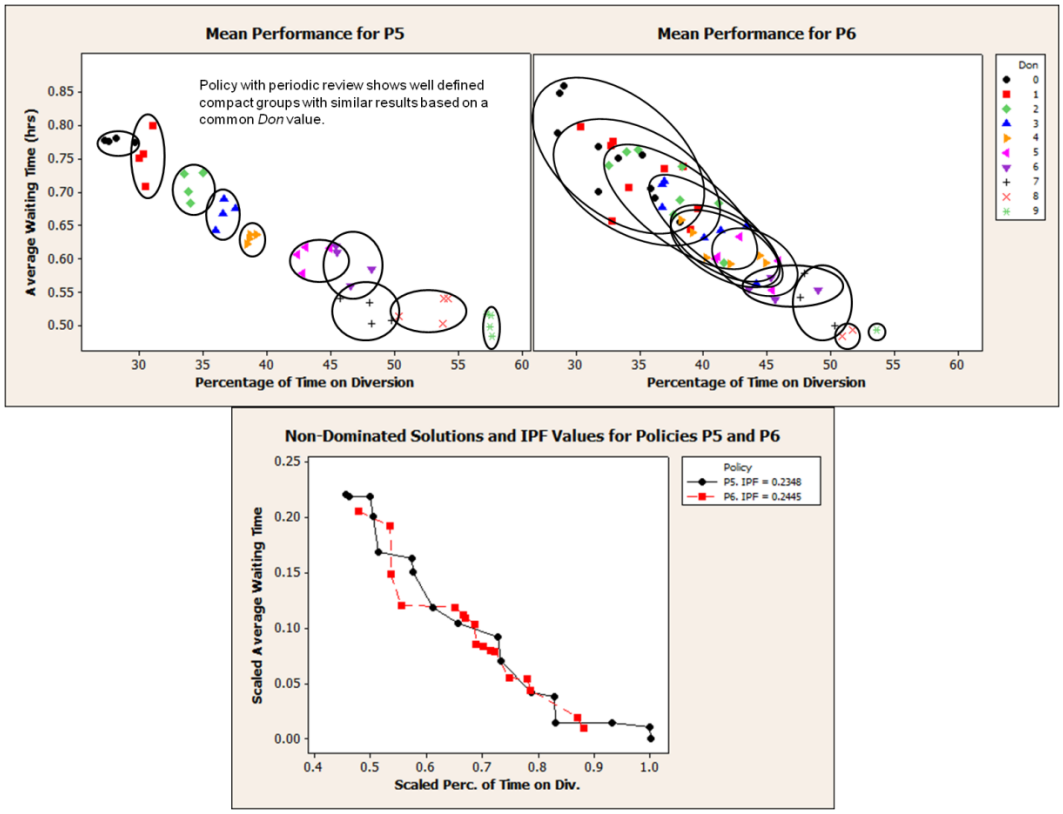

Figure 2.7. Mean performance of P5 and P6 (*top*) and nondominated solutions and IPF value per policy (*bottom*).

Table 2.5. Summary of IPF values for AD policies.

Policy	IPF
P1	0.1634
P2	0.1662
P3	0.1519
P4	0.1503
P5	0.2348
P6	0.2445

2.5.2 Variability of the Policies

Mean performance across the six policies show clustering patterns depending mainly on the Don level. This clustering is more evident in periodic

review. However, the graphs presented above showing the mean performance do not provide information regarding variability.

Similarly to the analysis of mean performance, the assessment of variability across policies use methods that allow a graphical observation of the variability by constructing simultaneous confidence ellipses (CE), as well as a quantitative measure by computing the R^2 of clusters of policy instances.

The simultaneous confidence ellipses are constructed taking advantage of the potential correlation between criteria. The plots of simultaneous confidence ellipses for a given instance of a policy yields ellipses that contains data points from all the replications executed for that particular case. The larger the correlation between criteria is, the better defined the ellipse is; thus, the data from replications is concentrated in an ellipse with smaller area, whose large axes has a positive slope. On the other hand, a small correlation yields an ellipse similar to a circle with a slope for the large axes close to zero. Overlapping ellipses for two or more instances imply that there is not a significant difference in the performance among the instances.

The quantitative assessment of variability for a policy relies on the formation of clusters and application of *K*-means concepts to analyze the spread of solutions from all the replications included in the cluster. Each cluster is a group of solutions of the form: (percentage of time on diversion, average waiting time).

The clusters defined for a policy depends on the number of significant parameters in the policy. Thus, the policies proposed in this chapter comprise two

parameters: *Don* and *Doff*. For some policies, *Don* is the only significant parameter to explain the variability in the two criteria, while for other policies both parameters are significant. Regression analysis comprising all the replications for all the instances of a policy is used to find the significant parameters of the policy. Then, the clusters for the policy are defined.

Each cluster contains the results for all the replications run with common level of the significant parameter. Thus, for a policy whose *Don* parameter is the only significant factor, the number of clusters formed is the number of levels for that parameter used in the experimentation. The computation of the R^2 per policy requires the computation of the total sum of squares (SS_{Total}) and the sum of squared error (SS_{Error}) for each cluster of the policy, which are obtained using Equations (2.3), (2.4) and (2.5) respectively.

$$SS_{Total} = \sum_{i=1}^{K} \sum_{s \in S_i} dist(\mu, s)^2 \tag{2.3}$$

$$SS_{Error} = \sum_{i=1}^{K} \sum_{s \in S_i} dist(c_i, s)^2 \tag{2.4}$$

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}} \tag{2.5}$$

where,

dist(*a*, *b*): Euclidean distance between *a* and *b*

s: response of the form (percentage of time on diversion, average waiting time)

S_i: the *i*th cluster

K: number of clusters

c_i: centroid of *i*th cluster

μ: centroid obtained from the grand average of the responses of all the clusters (all treatments, all replications). Thus,

$$\mu = \left(\frac{\sum_{j \in P.} DIV_j}{\tau(P.)}, \frac{\sum_{j \in P.} WT_j}{\tau(P.)} \right) \tag{2.6}$$

where,

$P.$: type of policy being analyzed $P. \in$ {P1, P2, P3, P4, P5, P6}

DIV_j: jth response "percentage of time on diversion" of policy P.

WT_j: jth response "average waiting time" of policy P.

$\tau(P.)$: total number of responses of policy P. (number of treatments x number of replications per treatment)

This chapter considers two types of centroids (c_i), one is given by the mean response of the cluster (Equation (2.7)) and another is given by the predicted response obtained from the simultaneous application of regression equations per criteria (Equation (2.8)). It is expected that choosing the mean as a centroid for every cluster will produce higher R^2 values than using regression equations, because choosing the mean as the centroid minimizes the SS_{Error} (Tan et al. 2006). Nevertheless, if both R^2 values are very similar, it is convenient to use the regression equations to predict the performance of new instances or find the policy parameters that could yield performance in a desired range. Hence, the centroids used are given by:

$$c_i = \left(\frac{\sum_{j \in S_i} DIV_j}{\tau(S_i)}, \frac{\sum_{j \in S_i} WT_j}{\tau(S_i)} \right) \tag{2.7}$$

$$c_i = \left(DIV(P.,i), WT(P.,i)\right) \tag{2.8}$$

where,

DIV(*P.*, *i*): response regarding percentage of time on diversion from regression equation depending on policy type and policy parameters of cluster *i*.

WT(*P.*, *i*): response regarding average waiting time from regression equation depending on policy type and policy parameters of cluster *i*.

2.5.2.1 Simultaneous confidence ellipses

The precision of individual confidence intervals is measured by computing the relative precision (half width of a 95% confidence interval / average) for all the treatments used for the six policies. The findings show that the performance measure was consistent across replications. Average relative precision of the average waiting time across the six policies is 7.10%, 6.41%, 5.8%, 5.49%, 6.4% and 6.43% for policies P1 to P6, respectively. On the other hand, the average relative precision of the percentage of time spent on diversion is 10.85%, 12.04%, 12.15%, 12.88%, 4.57% and 5.09% for policies P1 to P6, respectively. Therefore, looking at the precision of individual confidence intervals, the performance across policies behaves consistently, especially for P5 and P6 that provide better precision on the time spent on diversion, mainly due to the scale of their solutions. However, differences across the policies are observed by looking at 95% simultaneous confidence ellipses in Figure 2.8. For instance, there exist significant differences in variability depending on the factor considered in the policy. P1 exhibits very well defined ellipses that allow one to discriminate between instances with different *Don* level. It is also evident for P1 that

confidence ellipses of the policies with the same *Don* level overlap, regardless of the *Doff* level, confirming that the amount of time to re-evaluate the diversion status does not have a significant effect on the responses. P2 is similar to P1, but the ellipses have a wider area and more ellipses overlap as the percentage of time on diversion goes to zero.

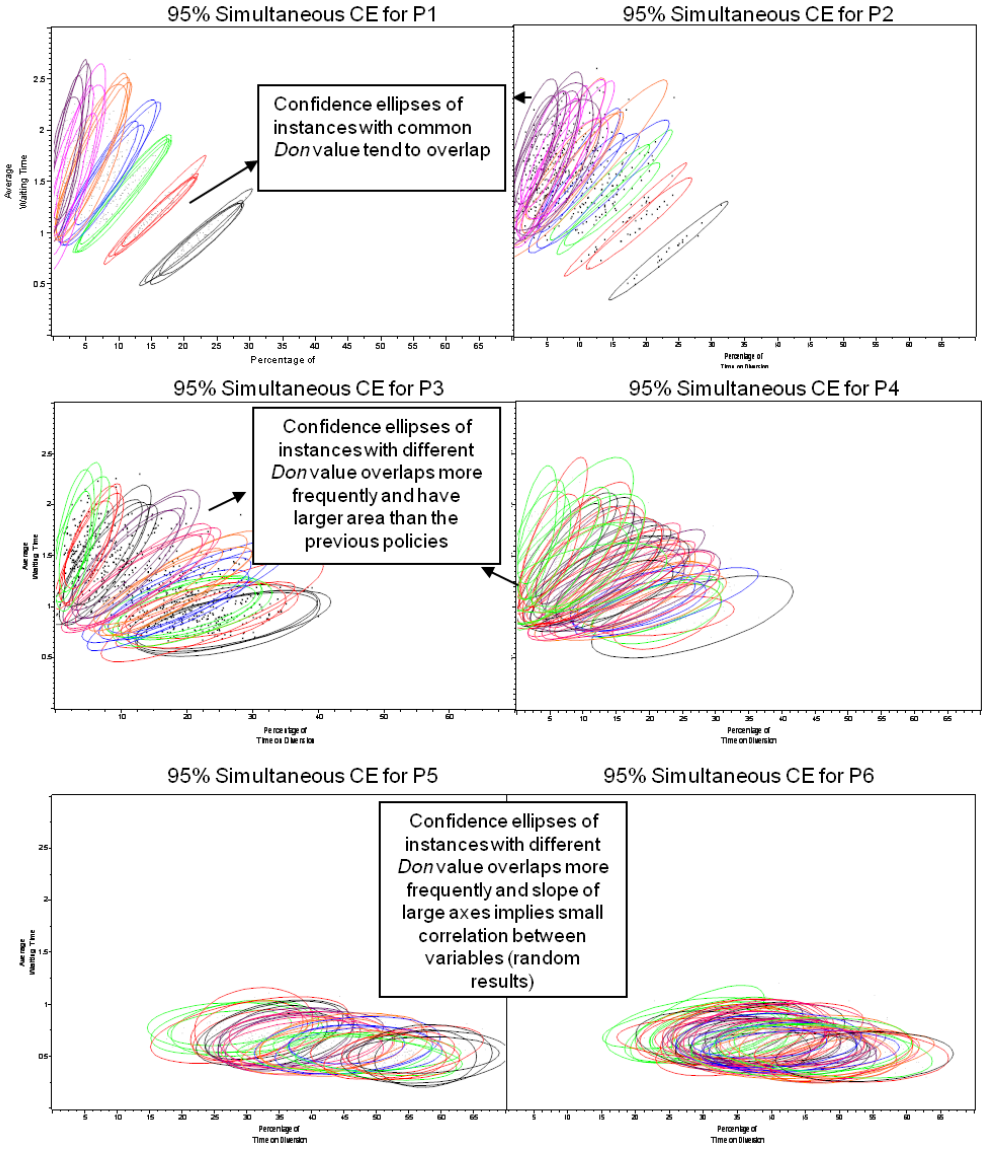

Figure 2.8. 95% simultaneous confidence ellipses for P1 and P2 (*top*); P3 and P4 (*middle*); and P5 and P6 (*bottom*).

Policies P3 and P4 exhibit ellipses that are larger than those in P1 and P2, causing more confidence ellipses to overlap. However, it is still possible to observe the clustering of the ellipses with a common *Don* value in policy P3. Policies P5 and P6 have a larger amount of ellipses overlapping, making difficult the discrimination among treatments. Nevertheless, there are two aspects to highlight; first, some of the ellipses for the last set of policies are smaller due to the small values in the average waiting time. Second, the change in the slope of the large axis of the ellipses is evident depending on the factor that the policy is based on. Thus, for policies depending on the number of patients waiting in the ED the slope and definition of the ellipses allow one to infer that there is a high correlation between the average waiting time and percentage of time on diversion. On the other hand, policies P5 and P6 exhibit ellipses whose large axis is almost parallel to the horizontal axis, meaning that the correlation between the two performance measures is very small; hence, results of these policies look random.

2.5.2.2 Clustering and computation of R^2

The behavior seen in the simultaneous confidence ellipses are useful to interpret the analysis of regression applied to both responses, whose results are shown in Table 2.6. The table shows the regression equation for each response, which in all cases is significant. In addition, the R^2, the R^2 prediction and the significance level of policy parameters are shown.

The analysis of regression confirms that the *Doff* parameter is not significant in policies P1 and P3, but it is for P2, P4, P5 and P6, considering a significance level of 0.05 for both responses. The influence of the parameters

included in the regression equation on the variability of the response is given by R^2 and the capability to predict future responses is given by R^2 prediction. It can be seen that the R^2 for the average waiting is moderate, which means that the parameters of the AD policies have a moderate effect on the variability observed in the average waiting time. Thus, there could be other controllable factors that affect this variable. On the other hand, the parameters of the AD policies have greater impact on explaining the variability of the percentage of time spent on diversion; therefore, the R^2 is larger for this variable. From these results, it can be concluded that regression equations considering the significant policy parameters are able to make better predictions for the percentage of time spent on diversion than for the average waiting time. Furthermore, the policies that can better explain the variability on the individual responses are the policies with periodic review than their counterparts with continuous review (P1, P3 and P5 over P2, P4 and P6).

Table 2.6. Regression analysis applied to each metric independently.

Average Waiting Time					
Policy	Regression Equation	R^2	R^2 (pred)	*p*-value *Don*	*p*-value *Doff*
P1	0.922 + 0.013 *Don*	0.479	0.475	< 0.0001	0.7434
P2	0.950 + 0.010 *Don* + 0.0028 *Doff*	0.392	0.385	< 0.0001	0.0002
P3	0.734 + 0.095 *Don*	0.657	0.656	< 0.0001	0.1180
P4	0.775 + 0.052 *Don* + 0.0500 *Doff*	0.561	0.558	< 0.0001	< 0.0001
P5	0.790 – 0.031 *Don* – 0.0005 *Doff*	0.436	0.432	< 0.0001	0.0261
P6	0.833 – 0.0198 *Don* – 0.015 *Doff*	0.319	0.315	< 0.0001	< 0.0001

Percentage of Time on Diversion					
Policy	Regression Equation	R^2	R^2 (pred)	*p*-value *Don*	*p*-value *Doff*
P1	22.37 - 0.319 *Don*	0.828	0.827	< 0.0001	0.3983
P2	22.68 – 0.242 *Don* – 0.070 *Doff*	0.725	0.722	< 0.0001	< 0.0001
P3	28.67 – 2.570 *Don*	0.742	0.741	< 0.0001	0.3791
P4	27.03 – 1.570 *Don* – 0.950 *Doff*	0.636	0.634	< 0.0001	< 0.0001
P5	26.31 + 3.180 *Don* + 0.026 *Doff*	0.782	0.781	< 0.0001	0.0090
P6	26.36 + 1.630 *Don* + 1.180 *Doff*	0.572	0.570	< 0.0001	< 0.0001

Nevertheless, the existing correlation of the responses suggests that a joint analysis should be used instead of looking at the individual performance. For this purpose, the clustering analysis previously proposed is applied and the R^2 for the two types of centroids are obtained. Table 2.7 shows the results of this analysis. There are two important aspects to highlight: the R^2 of the joint analysis is, at least, as good as the R^2 obtained from the individual analysis of the percentage of time on diversion, which means that variability of the pair average waiting time and percentage of time on diversion is better explained by the AD policies when these variables are analyzed together. In addition, it can be seen that the centroids obtained by the regression equations produce R^2 very close to those obtained by the mean.

Table 2.7. R^2 of joint analysis using *K*-means clustering.

				Centroid: Mean		Centroid: Regression Equations	
Policy	Number of Clusters	Total Number of Observations	SS_{Total}	SS_{Error}	R^2	SS_{Error}	R^2
P1	7	560	27740.33	3304.14	0.881	4778.46	0.828
P2	28	560	18779.82	4607.47	0.755	5171.13	0.725
P3	10	800	59010.41	14797.36	0.749	15203.04	0.742
P4	55	1100	51177.21	17732.29	0.654	18635.20	0.636
P5	40	800	85521.43	17432.26	0.796	18579.34	0.783
P6	55	1100	69250.74	28179.30	0.593	29628.96	0.572

Despite the mean taken as centroid is more accurate than regression equation, the latter can be used to define policies that have a specific objective in terms of waiting time and/or percentage of time on diversion. Since regression equations of P1 and P3 have only one parameter (*Don*), it is possible to obtain the value of the *Don* parameter that can produce the desired value of waiting time,

and then use the *Don* value obtained to estimate the expected percentage of time on diversion. For example, if the decision maker is interested in P1 to reduce the patient average waiting time to 1.3 hours, then *Don* parameter should be set to 30. This application of the regression equations as a search algorithm can reduce the number of simulation runs to obtain a policy that satisfies the desired objectives of the decision maker.

In addition, the changes of the R^2 across the policies as shown in Table 2.7 suggest that the variability and the accuracy of a prediction depend on the configuration of the review of the system. Hence, the policies with a periodic review produce results whose performance is more consistent than those produced by its counterpart with continuous review. For example, the R^2 of P1 is greater than the R^2 of P2. The value of P3 is greater than the value of P4 and the value of P5 is greater than the value of P6.

Combining the analysis of the mean performance and the variability across different policies can help in the design of robust policies that might be able to achieve consistent performance level in a desired range. Therefore, besides analyzing the trade-off between the time spent on diversion and the reduction of the waiting time, the decision maker should leverage the quality of the mean performance produced by the policy and its variability. For example, policy P5 shows good consistency given that its R^2 is the second highest across all the policies, but its IPF value and the range of solutions produced imply that the mean performance seriously affects the accessibility to the system.

For the case of an ED with similar characteristics to those analyzed in this chapter, the design of an effective policy in terms of accuracy and precision would consider the number of patients waiting in the ED or the number of patients boarding with a periodic review.

2.6 Discussion

The increasing trend of time spent on diversion in regional healthcare delivery networks has caused some local governments to prohibit the use of this practice. The Center for Disease Control found that about 7.5% of hospitals surveyed for the Staffing, Capacity and Ambulance Diversion report of 2003-04 have prohibited using AD because of state or local regulations (Centers for Disease Control and Prevention 2006b).

However, the recent "no ambulance diversion" policies adopted in some areas across the United States have put a strain on the operations of crowded EDs. For instance, hospitals in Massachusetts have seen a rise on the waiting times of ED patients and a greater number of patients boarding in inappropriate areas after this type of policy was implemented in that state (Massachusetts Nurses Association 2009). On the other hand, this type of law can be interpreted as an incentive for healthcare providers to look for other solutions by investing in research, engineering analyses or resources to relieve congestion from their EDs.

For instance, adding more beds to the system modeled in this chapter would reduce the average waiting time without going on diversion. Nevertheless, adopting AD policies could still reduce further the average waiting time. It is understandable that diverting patients from EDs might not be a safe decision for

the diverted patient, but the time to receive treatment might be smaller if the new ED destination is less saturated than the original destination.

Therefore, the call of different authors to discuss and analyze AD policies motivates this research. Even though the hospital being modeled is fictitious, its characteristics resemble the behavior and patterns seen in many EDs across the United States. Moreover, the methodology proposed in this chapter to analyze different AD policies by the mean performance and variability could be applied to other systems with similar characteristics.

The AD policies designed by providers usually include the observation of different state variables of the system, but the literature highlights that the decision of going on diversion usually is dominated by one factor. The three main factors found in literature as contributors for decision episodes are analyzed in this chapter. The results obtained from the analysis exhibit significant differences among the factors.

For instance, policies based on the lack of inpatient beds are more conservative and produce results whose range of time spent on diversion is much higher than the other policies. This can explain why some hospitals have a larger fraction of time on diversion status since the lack of inpatient beds is one of the most common causes to decide diverting ambulances.

Consequently, the design of AD policies should include the modeling of the system and the analysis of mean performance and variability of the results to allow the implementation of robust policies. This chapter proposes a methodology that enables decision makers to perform this task considering the tradeoff between

the time spent on diversion and the average waiting time of accepted patients in the ED.

The study of the mean performance is given by Pareto analysis using bi-criteria graphs that enable the comparison of different AD policies. In addition, the analysis includes using a quantitative metric that evaluates the policies in terms of the cardinality and coverage of their nondominated solutions. On the other hand, the analysis of variability includes the construction of simultaneous confidence ellipses and the creation of clusters that depend on the significant parameters of the policies in a regression equation. The clusters of every policy are evaluated through the R^2. Two types of centroids are compared for each cluster: the mean performance for both criteria of the results belonging to the cluster and the predicted performance obtained by a regression equation. Since the mean performance always minimizes the sum of squared error in a cluster, it can be used to compare the consistency of results across the policies and the effectiveness of using regression equations to predict the performance of new policy.

The findings in the experiments presented in this chapter suggest that policies based on number of patients waiting and number of patients boarding offer a good balance between the mean performance and the variability of their results. Furthermore, the use of periodic review produces more consistent performance. However, decision makers could use this methodology to evaluate their own broad possibilities of alternatives in order to assure the quality of their service in terms of the accessibility and timeliness of emergency care.

It is important to discuss the limitations of this study. It has been highlighted that the main object of study in this research is a fictitious hospital, but data used to build the model supports the representativeness of the model. On the other hand, the proposed methodology is applied to the analysis of a single-hospital; however, the nature of the problem implies that other hospitals might be affected. Nevertheless, the objective of this chapter is to present a methodology that can be applicable in the analysis of this important problem. Furthermore, AD diversion policies are designed and executed by authorized individuals of the hospital, complying with guidelines that government or private agencies may define. Besides, the effect of AD of one hospital over another hospital can be captured in the sample arrival rate that is used to build the simulation model. Therefore, the methodology proposed in this chapter intends to be presented as a set of tools that policy makers in each hospital can follow to define and compare their own policies.

In addition, there are other aspects that should be addressed about AD, such as the finance of ambulance patients and hospitals. Hospitals could see an opportunity cost for diverting patients; however, accepting patients in an overcrowding facility could make the hospital incurring in costs because of adverse events. On the other hand, hospitals may decide to go on diversion in order to save beds for elective admissions (i.e. scheduled surgeries). Nevertheless, the American College of Emergency Physicians (ACEP) discourages the use of financial reasons to divert patients. Thus, ACEP states that AD criteria must be based only on capacities or services of the hospital.

Another important aspect to discuss is the characteristic of the regional healthcare delivery network where AD is applied. As found in national reports, AD is a problem existing in metropolitan areas. Therefore, distances traveled by diverted ambulances are not as large as if the problem existed in rural areas. In fact, hospitals located in nonmetropolitan areas rarely go on diversion and their waiting times are much smaller than hospitals in metropolitan areas (Centers for Disease Control and Prevention 2006b).

2.7 Conclusions

AD has been adopted by several EDs across the United States as a way to reduce congestion. However, it has not been deeply discussed to what extent this objective is obtained by diverting patients. This research analyzes the impact on the average waiting time of the ED patients and on the time spent on diversion of policies that considers the main indicators in practice to go on diversion.

Through this research, it has been shown that the two performance measures are in conflict with each other; therefore it is responsibility of the decision makers to analyze the potential impacts of the policies that they design and choose the best option to balance diversion and waiting time according to the interests of each individual institution.

The procedure followed in this chapter to analyze AD policies can be adopted by real EDs to study the impact of diversion policies using experimentation based on simulation models. The results show significant differences in performance behavior of AD policies depending on the factor that they are based on. However, performance is not only policy dependent, but also

model or hospital dependent, because particular characteristics in arrival rate, admission probability, length of stay or acuity of patients can have a significant impact in the pace that AD reduces congestion.

In addition, results from the model analyzed show that policies based on inpatient occupancy level, which are very common in practice, have a higher percentage of time spent on diversion compared with other policies and also might not be very consistent. On the other hand, policies based on number of patients waiting in the ED or number of patients boarding using periodic review performs better than the others in terms of quality and consistency of results.

It is important to mention that results of this chapter show the potential improvement from AD using a local approach, specifically the average waiting time of accepted patients. However, the analysis of the overall improvement in the healthcare delivery system through AD must include nearby hospitals. Therefore, this project will extend to optimize the AD policy for a single hospital assuming that there is information available about a neighboring hospital. In addition, optimization of AD policies for multiple hospitals can be explored.

CHAPTER 3

OPTIMAL AMBULANCE DIVERSION CONTROL POLICIES

3.1 Introduction

Media and papers have been highlighting the overcrowding problem in emergency departments (EDs) in the United States (US) during recent years (Associated Press 2006). One of the major negative impacts of congestion in EDs is the long time that patients have to wait before starting to receive treatment, resulting in seriously adverse events, including death (KVAL 2010; CNN U.S. 2008). The risk of such adverse events increases when the condition of the patient is severe and when waiting times extend beyond a recommended safety time threshold (RSTT), which is set by the Center for Disease Control and Prevention (CDC) based on patient severity (which is assessed by various indicators of the health condition of the patient such as vital signs and stability), and the amount of resources required. United States General Accounting Office (2009) has drawn attention to the high fraction of patients that have to wait beyond RSTT. For example, in 2006, 73.9% of all patients that should have received "immediate" attention (no waiting at all) according to their severity index had to wait for some time in the EDs. In addition, 50.4% of patients with an RSTT of 14 minutes had to wait longer than that threshold before they started receiving treatment.

In order to reduce congestion and avoid potential implications of long wait times, EDs sometimes divert ambulances to other hospitals by requesting emergency medical services to bypass their facilities. This strategy is commonly implemented in US hospitals. According to United States General Accounting

Office, in 2003, 25% or more of the hospitals in several US metropolitan areas were on diversion more than 10% of the time. For 2006, 27.3% of hospitals reported going on diversion, and the average number of hours on diversion during that year were 473 hours (United States General Accounting Office 2009).

Although, EDs often divert ambulances to tackle overcrowding, this approach can have negative consequences when AD policies are not properly designed. For instance, Yankovic et al. (2010) indicate that AD might increase mortality among patients transported by an ambulance. Consequently, AD decisions should consider various factors such as the current congestion at the ED, severity of the patients, and the status of neighboring hospitals. For example, if a neighboring hospital is relatively near and currently less crowded, then it is more likely that an arriving patient in an ambulance can start receiving appropriate treatment earlier if he/she is diverted from an overcrowded facility. On the other hand, while ambulances can be diverted, EDs do not have control over walk-in arrivals, which, by law, have to be accepted and treated. Therefore, while on diversion, EDs still accept walk-in patients; these patients also contribute to congestion.

In this chapter, an optimal ambulance diversion control policy is developed. The optimal policy is defined to minimize the average time a patient waits longer than his/her RSTT. The following research questions are addressed in this chapter: (i) Can optimal AD policies significantly increase the safety of patient by minimizing the time that patients wait beyond their RSTT?; (ii) What is the structure of optimal AD policies?; (iii) What are the impacts of patient traffic

and severity mix on optimal AD decisions?; (iv) What is the value of information about the time to start treatment in the neighboring hospital(s) on optimal AD decisions and performance of the optimal policy?; and (v) How do policies applied in practice perform compared to optimal AD policies?

Empirical studies on the effectiveness of AD policies and the design of policies that minimize AD are available in the medical literature. In general, medical community is opposed to AD. They consider it an inefficient and risky decision. Therefore, they suggest avoiding or minimizing the use of AD. Instead, they propose to analyze the causes of overcrowding and take other actions to relieve congestion. Approaches to avoid AD include the redesign of AD guidelines to restrict the number of hours spent on diversion by hospitals serving a specific geographic region. The implementation of these guidelines has resulted in significant decreases in the number of hours on AD in the regions of study; this includes San Diego and Sacramento, California (Vilke et al. 2004b; Asamoah et al. 2008; Patel et al. 2006). Unfortunately, these studies do not discuss the effect of avoiding AD on other performance measures, such as the average waiting time.

Other empirical studies propose actions to reduce congestion from EDs and consequently reduce diversion. These actions include redesigning patient flow and improving capacity allocation in EDs (Cochran and Roche 2009; Allon et al. 2011). In addition, blocked admissions to inpatient units have been analyzed to reduce its effect on the patient flow in the ED (McConnell et al. 2005). Other studies predict crowding conditions in EDs to make appropriate changes in advance (Hoot et al. 2008; Chockalingam et al. 2010).

While reducing AD can increase access to emergency facilities, there is evidence that suggest that laws prohibiting AD (i.e., No AD) can put significant stress on the operations of EDs (Massachusetts Nurses Association 2009). The consequences of such laws include increases in the average patient waiting time and the number of patients boarding (i.e., patients waiting for an open bed in an inpatient unit).

On the other hand, analytical studies of AD suggest that appropriate policies could improve the performance of an emergency care system. For example, Deo and Gurvich (2011) modeled the decisions of two EDs using game theoretic approaches with the objective of minimizing the average patient waiting time for each hospital in a system with two EDs. The authors found that a centralized design of diversion policies is Pareto-improving compared to a decentralized strategy that leads to a defensive equilibrium. The authors also proposed a threshold-type AD policy, but they did not explore the optimality of this type of control policy. Using similar approaches, Hagtvedt et al. (2009) analyze AD and pointed out the need of a central agent that coordinates AD. Ramirez et al. (2011) presented a simulation model of an emergency care delivery system to analyze the effectiveness of diversion and destination policies. They evaluated the use of an effective combination of diversion-destination policies as an ambulance flow control mechanism in order to reduce the average time spent in activities with inappropriate level of care, which includes transportation to ED, waiting and boarding in the ED.

Even though admission control methods are commonly used in various manufacturing and service systems, the AD literature has not considered the use of such methods in the control of ambulance arrivals to date. Early studies on admission control typically focus on the control of a single customer class using $M/M/1$ queuing models (see Stidham (1985) for a survey). More recently, studies consider control of several demand classes requiring different levels of service. Ha (1997) discusses an inventory control problem of N demand classes that incur different lost sales costs when customers are not admitted into the system. Similar to the proposed setting, Carr and Duenyas (2000) discuss two demand classes, where one of the classes is always accepted into the system (similar to the walk-ins in the model presented in this chapter), and the company has an option to reject the arrivals from the other class (similar to the ambulance arrivals). Gupta and Wang (2007) consider one contracted demand class whose orders are always accepted and one transactional demand class whose orders can be rejected. Similarly, Feng and Pang (2010) consider a long-term contract market whose orders are always accepted, and the spot market whose orders may be subject to rejection. In the recent work of Chen et al. (2011), the authors discuss the admission control problem of the orders coming from an online retailer. All of above discussed studies control demand using accept/reject decisions, similar to accept/divert decisions. In addition, there is a rich literature on the control of admission using pricing and due date decisions. The readers are referred to the surveys of Elmaghraby and Keskinocak (2003) and Keskinocak and Tayur (2004)

for implementation of pricing and due date management for admission control, respectively.

In particular, this chapter contributes to existing literature in AD by proposing a mathematical model based on Markov Decision Processes (MDP) formulation to obtain the optimal AD control policies for a hospital. The objective is to minimize the long-run average expected tardiness per patient, where tardiness is defined as the length of time that a patient waits beyond his/her RSTT, before starting to receive treatment. Assuming Poisson arrivals, exponential treatment times and two severity levels, the structure of optimal policies is analyzed using both theoretical and computational analysis. Using theoretical analysis, this chapter shows that the optimal diversion policy can be characterized by a threshold curve, under the special condition that all ambulance patients are critical. Using computational analysis, the structure of the optimal AD policies is further studied by observing the impact of (i) patient arrival rates, (ii) the severity mix of patient population, and (iii) the "amount" of available information on the time to start treatment at the neighboring hospital(s). Next, a simulation study is presented, where various modeling assumptions are relaxed to represent more realistic scenarios, and compare the optimal policies with that of other simpler policies used in practice such as not diverting at all and diverting only when there are no available beds. Computational analysis verifies the superior performance of the optimal policies obtained using the proposed MDP model. In addition, a simple policy that diverts ambulances when there are no available beds for critical patients is shown to yield satisfactory results. Finally,

the possible drawbacks of the proposed approach in practice are discussed, and conclude that these drawbacks can be resolved by allocating sufficient capacity to EDs.

This chapter has to main contributions to the healthcare literature. First, to the best of our knowledge, this is the first study discussing optimal control of AD using an MDP formulation. Second, it considers a novel objective that minimizes the time that patients wait beyond a RSTT before starting to receive treatment. Although the AD literature includes various studies that discuss minimization of time spent in ED, this objective does not take into account the severity of more critical patients, whose treatment delays may result in death. Since RSTT depends on the severity level of patients, the objective considers the safety of the patient as a performance measure for AD policies, which is a significant measure to evaluate the effectiveness of AD policies according to (Asplin 2003). In addition, since the objective function is in time units, it does not require any cost parameterizations that have been commonly needed in previous literature.

The remaining sections of the chapter are organized as follows. Section 3.2 introduces the model. Section 3.4 analyzes the impact of the level of information on the time to start treatment in a neighboring hospital(s). Section 3.5 presents a simulation model to compare the policy prescribed by the MDP with policies used in real-life settings. Section 3.6 analyzes the issues related to the practical implementation of the AD policies prescribed by the MDP. Finally, Section 3.7 presents some conclusions and future extensions.

3.2 Model Formulation

The model considers an ED with two arrival streams, each following a Poisson process, which has been discussed as a reasonable approach to model arrivals to EDs (Green 2006): (i) ambulance arrivals with rate λ^A and (ii) walk-ins with rate λ^W. Arriving patients can have one of two types of severity levels: level 1 represents the critical patients, and level 2 are less emergent cases. The ED has two treatment areas dedicated to each severity level: A1 (critical care), which treats patients of level 1 severity, and A2 (fast-track), which treats patients of level 2 severity. Although most Emergency Severity Indices (ESI) consider three to five severity levels, the majority of patients can be grouped under two major categories in terms of the required treatment resources and priority. One group includes patients with an immediate and emergent need for emergency care, and another group includes patients with urgent and semi/non urgent needs. Furthermore, many hospitals in metropolitan areas have treatment spaces dedicated to patients with moderate or low severity level, similar to the area A2 considered in this model (Cochran and Roche 2009).

It is assumed that ambulance patients are level 1 with probability p^A_1, and level 2 with probability 1-p^A_1. In general, it is safe to assume that p^A_1 is relatively high (i.e., p^A_1>0.7). Similarly, walk-in patients are level 1 with probability p^W_1, and level 2 with probability 1- p^W_1. The tuple (p^A_1, p^W_1) is refer to as the severity mix. Upon admission to the ED, patients are first identified as level 1 or level 2, and they are served in the order of arrival at the corresponding area (A1 or A2). If all the beds in the appropriate area are occupied, then an arriving patient waits in a

queue corresponding to his/her treatment area. The number of beds is given by c_1 for A1 and c_2 for A2. Once a patient accesses a bed in the corresponding area, the patient remains the bed for some amount of time referred to as "treatment time". The treatment time considered in this chapter may include activities such as bedside assessment provided by nurses and doctors, delivery of medications, and the discharge process. It is assumed that the treatment time of a level i patient is a random variable distributed exponentially with rate μ_i for $i \in \{1, 2\}$. While the exponential distribution may not be a very good fit to represent the total treatment time, it is commonly used in the literature due to its analytical tractability (see for example Deo and Gurvich (2011)). In Section 3.5, this assumption is relaxed and a simulation model is developed using more realistic distributions. Other resources found in EDs such as doctor, nurses and medical equipment are not included in the model since they have low impact on the diversion decisions according to CDC (2006a).

The state of the system can be represented by the tuple $(n_1(t), n_2(t))$, where $n_1(t)$ and $n_2(t)$ represent the number of patients in the system with level 1 and level 2 severity at time t, respectively. The parameter t is dropped from the notation and the state space is denoted as $S = \{(n_1, n_2) : n_1 \geq 0; n_2 \geq 0\}$. The flow of patients is depicted in Figure 3.1.

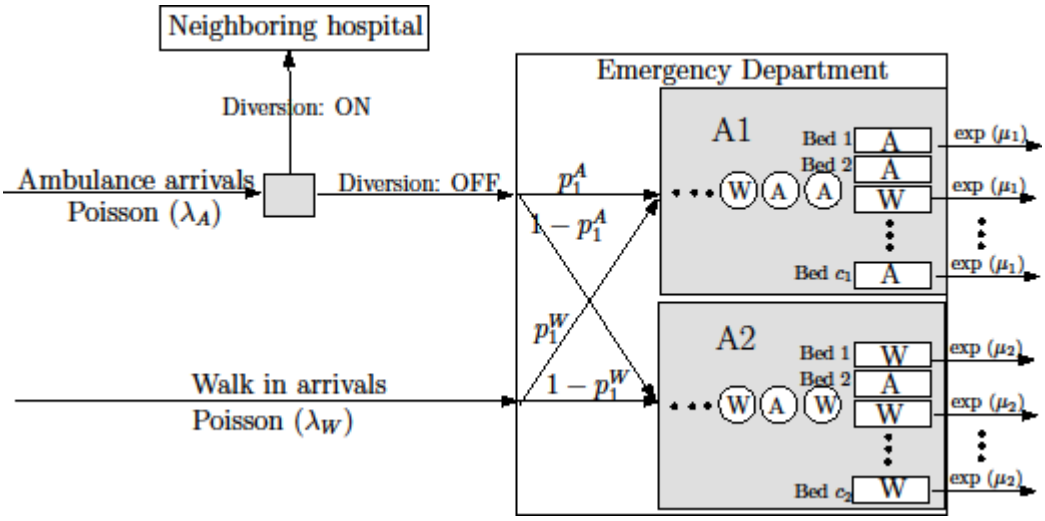

Figure 3.1. System representation.

To ensure the existence of a solution for the MDP model, it is assumed that total arrival rate is less than the total treatment rate in each treatment area (Bertsekas 2001), that is ,

$$p^A{}_1\,\lambda^A + p^W{}_1\,\lambda^W < c_1\mu_1, \text{ and } (1\text{-}p^A{}_1)\lambda^A + (1\text{-}p^W{}_1)\lambda^W < c_2\mu_2 \qquad (3.1)$$

In the computational analysis presented in Sections 3.3 and 3.4, the total number of patients in the system is limited such that it eliminates the need for a stability condition as given in Equation (3.1). Therefore, this condition is relaxed in the computational analysis.

The objective of this chapter is to find a state-dependent ambulance diversion policy that minimizes the long-run average expected tardiness per patient (denoted as ETP henceforth) over an infinite horizon. Different from the traditional settings, tardiness is referred as the non-negative difference between the total waiting time of the patient in the ED and the RSTT of the patient

(denoted as d_i for severity level $i \in \{1, 2\}$). It is assumed that $d_1 < d_2$; therefore, the difference in the RSTT provides a "natural" weight that severely penalizes the objective as the waiting time of critical patients increases. Let $T_i(n_i)$, $i \in \{1, 2\}$ denote the expected tardiness of an arriving patient with level i given that there are n_i level i patients in the system upon his/her arrival. Then,

$$T_i(n_i) = \int_{d_i}^{\infty} (x - d_i) f_{i,n_i}(x) dx \tag{3.2}$$

where $f_{i,ni}(.)$ denotes the probability density function (pdf) of the waiting time in the queue of a level i patient that observes n_i level i patients in the system upon his/her arrival. Since treatment times are exponential with rate $c_i\mu_i$, $f_{i,ni}(.)$ is the pdf of the Gamma distribution with parameters n_i-c_i+1 and $c_i\mu_i$, when $n_i \geq c_i$. If $n_i < c_i$, then $T_i(n_i) = 0$. $T_i(n_i)$ for $n_i \geq c_i$ is evaluated using Laplace transforms in Theorem 1 of Hafizoglu et al. (2011). The closed-form expression for $T_i(n_i)$ is provided in Equation (3.3).

$$T_i(n_i) = \begin{cases} \dfrac{e^{-d_i c_i \mu_i}}{c_i \mu_i} \displaystyle\sum_{k=0}^{n_i - c_i} (d_i c_i \mu_i)^k \frac{n_i - c_i + 1 - k}{k!} & \text{If } n_i \geq c_i \\ 0 & \text{If } n_i < c_i \end{cases} \tag{3.3}$$

When an ambulance is diverted, the patient is sent to a neighboring hospital for treatment. The time to start treatment by a diverted patient in the neighboring hospital is a random variable, X. In particular, X includes the additional transportation time to travel to a further away facility and the waiting

time inside the ED. The expected tardiness of a diverted patient with level i is denoted as T^D_i for $i \in \{1, 2\}$, and evaluated as in Equation (3.4).

$$T_i^D = \int_{d_i}^{\infty} (x - d_i) f(x) dx \tag{3.4}$$

where $f(x)$ is the pdf of X.

The context of the model assumes that the ambulance crew communicates with the ED to learn if the patient can be taken to the hospital or not. Then, the decision maker in the ED chooses to divert or accept the ambulance depending on the current state of the system. This assumption does not contradict the diversion guidelines formulated by the American College of Emergency Physicians stating that diversion criteria must be based only on hospital capacity and not on financial decisions (American College of Emergency Physicians 1999). In addition, it is assumed that the severity of the patient is not known at the time the diversion decision is made.

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, υ^* denote the optimal average expected tardiness per patient 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

$$\upsilon * \frac{\lambda^W + \lambda^A}{\nu} + h*(n_1 + n_2) = \tag{3.5}$$

$$\frac{\lambda^W p_1^W}{\nu}\left[T_1(n_1) + h*(n_1+1, n_2)\right] + \frac{\lambda^W (1-p_1^W)}{\nu}\left[T_2(n_2) + h*(n_1, n_2+1)\right]$$
$$+ \frac{\tilde{c}_1\mu_1}{\nu} h*(n_1-1, n_2) + \frac{\tilde{c}_2\mu_2}{\nu} h*(n_1, n_2-1)$$
$$+ \min\{\frac{\lambda^A p_1^A}{\nu}\left[T_1^D + h*(n_1, n_2)\right] + \frac{\lambda^A (1-p_1^A)}{\nu}\left[T_2^D + h*(n_1, n_2)\right],$$
$$\frac{\lambda^A p_1^A}{\nu}\left[T_1(n_1) + h*(n_1+1, n_2)\right] + \frac{\lambda^A (1-p_1^A)}{\nu}\left[T_2(n_2) + h*(n_1, n_2+1)\right]\}$$
$$+ \left(1 - \frac{\lambda^W + \lambda^A + \tilde{c}_1\mu_1 + \tilde{c}_2\mu_2}{\nu}\right) h*(n_1, n_2)$$

where,

$$\tilde{c}_i = \begin{cases} n_i & \text{If } n_i \le c_i \\ c_i & \text{If } n_i > c_i \end{cases}$$

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

The first two terms on the right hand side of Equation (3.5) refer to the walk-in patients with severity level 1 and level 2, respectively. The third and fourth terms represent the departure events, which decrease the number of patients in A1 or A2 by one, depending on the severity level of the departing patient. The first part inside the minimum statement represents the average tardiness if an arriving ambulance patient is diverted, whereas the second part represents the average tardiness if the ambulance patient is accepted to the hospital. The last term corresponds to a selfloop due to uniformization.

The objective function can be changed easily to minimizing the weighted average expected tardiness per patient in the long run by adding weights to the tardiness expressions. These weights may depend on the severity level with the weight given to the tardiness of level 1 patients being greater than the weight given to the tardiness of level 2 patients.

3.3 Properties of Optimal Diversion Policies

In this section, some properties of an optimal solution to the Bellman equation given in (3.5) are derived. Theorem 1 shows that the optimal diversion policy is characterized by a monotonic threshold curve under a special case where all ambulance patients are critical. This result also justifies the common use of threshold-type policies used previously in the AD literature (Deo and Gurvich 2011; Hagtvedt et al. 2009).

THEOREM 1. *If* $p^A_1 = 1$, *there exists a threshold curve* $\Delta(n_1)$, *where it is optimal to divert incoming ambulances when* $n_2 > \Delta(n_1)$, *and accept them when* $n_2 \leq \Delta(n_1)$. *Furthermore,* $\Delta(n_1)$ *is non-increasing in* n_1.

PROOF. The proof is in Appendix A. □

In Figure 3.2, Theorem 1 is illustrated, where $\Delta(n_1)$ is shown by the representative non-increasing curve. It is optimal to divert an ambulance if $n_2 > \Delta(n_1)$, that is, if the state is located above the curve. In words, the state space above the curve denotes the cases where ETP added from accepting ambulance patients (i.e., second term within minimization in Equation (3.5)) is greater than the ETP added from diverting them (i.e. first term within minimization in Equation (3.5)). In the remainder, $\Delta(n_1)$ denotes the threshold curve. The

threshold curve provides a simple and useful mechanism that optimally determines ambulance diversion decisions. Using a relative value iteration algorithm, one can solve 3.5, and determine the threshold curve.

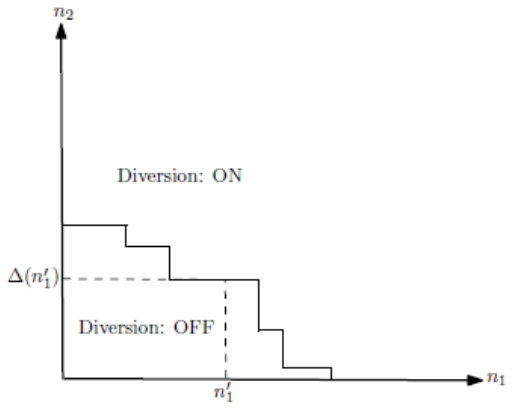

Figure 3.2. Illustration of Theorem 1.

In the proof of Theorem 1, it is required that $h^*(n_1,n_2)$ to be supermodular and the term $T_1(n_1)+h^*(n_1+1,n_2)-h^*(n_1,n_2)$ to be nondecreasing in n_1. While these properties can be shown to hold for the special case of $p^A{}_1=1$, the extensive complexity of Equation (3.5) does not allow to derive the desired properties when $p^A{}_1 < 1$. However, the computational study indicates that Theorem 1 holds in all the practical cases that are discussed in detail below.

Next, the behavior of optimal policy is explored using computational analysis. In all of the experiments, the values of c_1, c_2, μ_1, μ_2, d_1 and d_2 are fixed as indicated Table 3.1, and the values of the other parameters vary to test the system behavior under various scenarios.

Table 3.1. Factor levels used in the computational analysis.

Fixed Factors	Variable Factors
c_1 = 15 beds	Traffic $\in$ {Low, Medium, High}
c_2 = 5 beds	$p^A{}_1 \in \{0.8, 0.9\}$
μ_1 = 0.25 pat/hr	$p^W{}_1 \in \{0.3, 0.5\}$
μ_2 = 1 pat/hr	$E[X] \in$ {0.25,0.75,1.25,1.5,1.75 and 2 hrs}
d_1 = 0.25 hours	
d_2 = 1.5 hours	

The number of beds in the modeled ED is 15 for critical care and 5 for fast-track care. These numbers are close to the average number of beds in treatment spaces in real-life ED (an average of 14.6 beds in standard treatment spaces and 5 beds for other treatment spaces) (Centers for Disease Control and Prevention 2006b). The treatment rates were set to 0.25 and 1 patients per hour for critical and fast-track care, respectively. The first value is close to the average treatment time for immediate and emergent patients, given in several sources (Centers for Disease Control and Prevention 2006a; Cochran and Roche 2009; Hoot et al. 2008). The fast-track treatment rate is close to the value observed in Cochran and Roche (2009) for semiurgent and nonurgent patients. The RSTT set for level 1 is fixed to 0.25 hours, which corresponds to the second most emergent level in the ESI; this category is usually referred as "less than 15 minutes". The RSTT for level 2, on the other hand, is set to 1.5 hour, which corresponds to an average of the third and fourth ESI indices, usually referred as urgent ("1 hour") and semi-urgent ("2 hours") (Centers for Disease Control and Prevention 2006b).

The Utilization Due to Walk-in Arrivals (UDWA) is considered to quantify the low, medium and high levels of traffic, because walk-in arrivals

cannot be controlled using AD and they represent about 85% of the total arrivals. Hence, let

$$\text{UDWA}_1 = p^W_1\lambda^W/c_1\mu_1, \text{ and } \text{UDWA}_2 = (1-p^W_1)\lambda^W/c_2\mu_2 \quad (3.6)$$

where UDWA_1 and UDWA_2 denote the Utilization Due to Walk-in Arrivals in A1 and A2, respectively. Let $\max\{\text{UDWA}_1, \text{UDWA}_2\}$= 60%, 75% and 90% model the low, medium and high levels of traffic, respectively. The area with the highest utilization is referred to as the "congested area" in the remainder. For example, if traffic level is medium and p^W_1=0.3, then λ^W=5.36, which gives UDWA_1=43% and UDWA_2=75%, indicating that the congested area is A2. On the other hand, if traffic level is medium and p^W_1=0.5, then λ^W=5.63, UDWA_1=75% and UDWA_2=56.25%, which implies that the congested area is A1. Furthermore, for any combination of traffic intensity and severity mix, the value of p^W_1 determines the congested area in the ED. Hence, if p^W_1=0.3, then the congested area is A2; whereas if p^W_1=0.5, then the congested area is A1.

The proportion of the arrival rates of ambulances was fixed to be 15% of all the arrival rates. This value is very close to the national average of the percentage of ambulance arrivals to EDs in the United States, which is 15.5% (Centers for Disease Control and Prevention 2010). In addition, in all the analysis in this section, X is chosen to be deterministic. The impact of randomness of X is analyzed in Section 3.4.

An upper limit on the total number of patients in the system is considered in the implementation of the relative value iteration algorithm. Such an upper limit also allows relaxing the stability condition given in Equation (3.1). This upper limit is large enough to approximate the infinite capacity assumed in Section 3.2 while ensuring a reasonable execution time of the relative value iteration algorithm.

Figure 3.3 presents the threshold curves for four different values of (p^A_1, p^W_1) under medium traffic and a deterministic value of X = 0.75 hours.

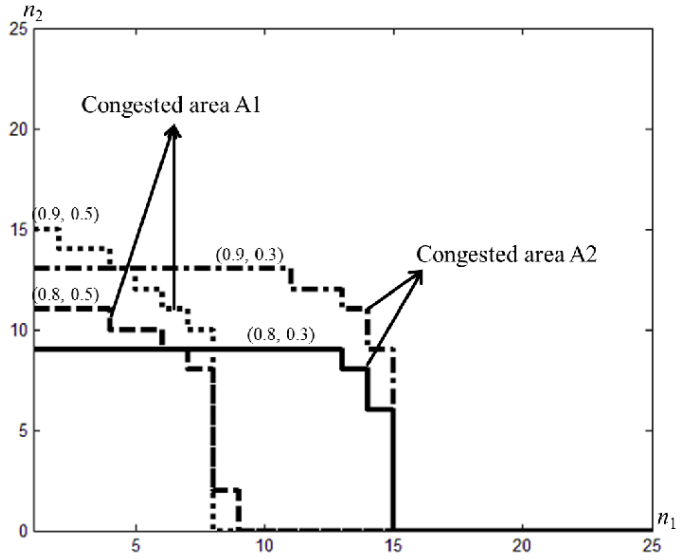

Figure 3.3. Illustration of thresholds for $p^A_1 \in \{0.8, 0.9\}$ and $p^W_1 \in \{0.3, 0.5\}$, under medium traffic and deterministic time to start treatment in other hospital (X) of 0.75 hours.

As shown in Figure 3.3, the congested area determines the shape of the threshold curve. It is expected that most ambulance patients will be critical patients; therefore if the congested area is A2, then the optimal policy initiates diversion only when all of the beds in A1 are occupied for low values of n_2.

However, if the congested area is A1, then the optimal policy initiates diversion much earlier in order to save beds in the congested area for future demand (possibly walk-ins). Hence, AD is an effective mechanism to alleviate congestion from A1, but it is not as effective when congested area is A2. The next sections present statistics that show significant reductions on average tardiness per patient when the congested area is the critical care. From a practical point of view, saving beds for future demand might not be acceptable by healthcare administrators, especially in the case of critical beds in the ED. Nevertheless, in the case of an emergency situation that affects a large proportion of the population (e.g. earthquake or terrorist attack), an ED located in the affected zone might divert ambulances to other hospitals and save resources for the walk-in arrivals of injured people.

In addition, the effect of the difference in threshold values and treatment rates can be seen in Figure 3.3. Any optimal policy plotted in this figure allows a queue in area A2 before diverting ambulances; that is, the thresholds observed in n_2 are greater than 5, which is the value set for c_2. The size of the queue allowed in A2 before diverting ambulances is smaller if that area is the congested one. On the other hand, queuing is not allowed in the critical care area A1; that is, the thresholds observed in n_1 are smaller than or equal to 15, which is the value set for c_1.

Next, the impact of traffic on the threshold curve is presented. Figures 3.4(*left*) and 3.4(*right*) demonstrate how threshold curves change with traffic intensity when X = 0.75 hours and congested areas are A1 and A2, respectively.

In general, the higher the utilization in the congested area is, the lower the threshold to initiate diversion is. However, changes in the threshold are more evident if the congested area is A1. That is, if the congested area is A1, AD policies might initiate diversion even when there are plenty of beds available in A1. Since patients arriving by ambulance are more likely to be critical patients, the optimal policy changes significantly in n_1 in order to manage the traffic. For example, when the congested area is A1 and there is high traffic intensity (90% UDWA), the optimal policy diverts all the time. For medium traffic, the optimal policy accepts some patients, but it saves almost half of the critical beds for future demand. For low traffic, the optimal policy practically waits to observe full occupancy in A1 before diverting ambulances. On the other hand, if the congested area is A2, the threshold in n_1 is also around the value of c_1 for low values of n_2, and the threshold in n_2 allows patients waiting in A2.

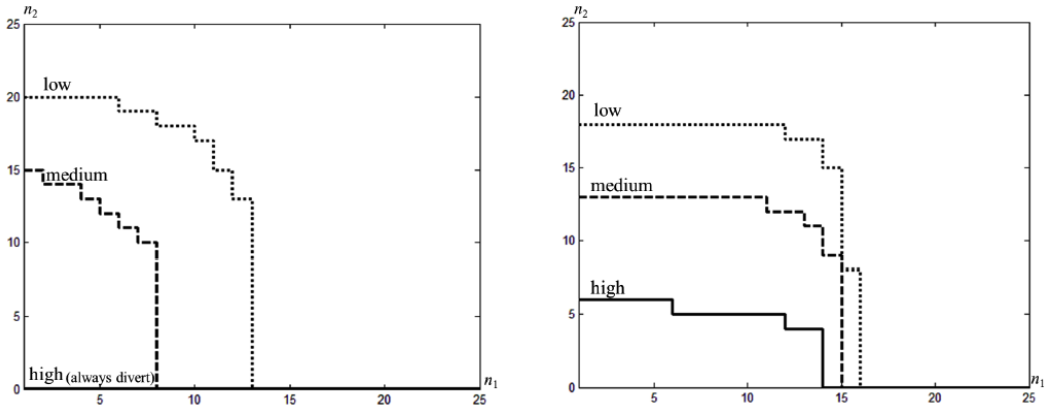

Figure 3.4. Illustration of thresholds for changing traffic levels when congested area is A1, p^A_1=0.9, and X=0.75 hours (*left*); and congested area is A2, p^A_1=0.9, and X=0.75 hours (*right*).

Finally, the impact of the magnitude of X on the thresholds is discussed. Noting that, T^D_1 and T^D_2 are functions of X, Theorem 2 shows that the increase in T^D_1 and T^D_2 pushes the threshold curve in the upward direction, under the special case of p^A_1=1.

THEOREM 2. *If* p^A_1 =1, *the threshold curve,* $\Delta(n_1)$ *is non-decreasing in* T^D_1 *and* T^D_2.

PROOF. The proof is in Appendix A. □.

Computational analysis results that are shown for p^A_1<1 are in line with Theorem 2. Figures 3.5(*left*) and 3.5(*right*) show the threshold curves for different values of X. The result is due to the fact that T^D_1 and T^D_2 increase in deterministic X, which implies that $\Delta(n_1)$ is non-decreasing in X as well.

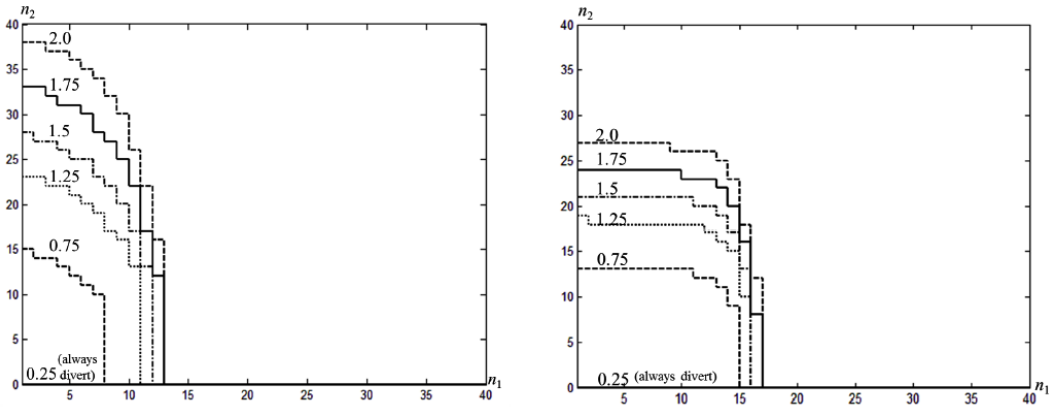

Figure 3.5. Illustration of thresholds for changing levels of the time to start treatment in the other hospital (X) for medium traffic intensity when congested area is A1, p^A_1=0.9 (*left*); and congested area is A2, p^A_1=0.9 (*right*).

For a deterministic value of X = 0.25 hrs, the optimal policy diverts all the time, regardless which area is the congested one. This is due to the fact that d_2<d_1=0.25, and hence, the policy diverts all the ambulance patients with a

guarantee of zero tardiness. However, as the time to be seen in another hospital increases, the optimal policy increases the threshold that initiates diversion. If the congested area is A1, the optimal policy might save beds in A1 for future demand when X has a moderate value. But as X increases beyond d_2, the optimal policy approaches initiating diversion under full occupancy in A1. If the congested area is A2 as shown in Figure 3.5(*right*), the threshold in n_1 increases in X, and might even allow a small queue, depending on the traffic intensity.

Diversion can be an effective tool for managing traffic in an ED and improving safety by minimizing tardiness. The optimal policy depends on several factors. However, optimal thresholds are more sensitive to these factors if the congested area is the critical care area because it is expected that most ambulance patients need to be treated in this area. Furthermore, diversion is initiated sooner if there is high traffic, to save beds for future walk-in demand, which cannot be diverted. Nevertheless, if the time to start treatment in a neighboring hospital increases significantly, or if the traffic intensity is low, or if the walk-in arrivals causes A2 to be the congested area, then the optimal threshold on n_1 practically waits to see full occupancy in A1 before diverting ambulances. In addition, if congested area is A2, the optimal threshold allows queuing in A2, because patients in that area have a relatively large RSTT and the treatment times are generally much shorter than those in A1.

3.4 Impact of Information of the Status of Neighboring Hospitals

The model presented in Section 3.2 assumes that the hospital under study has some information about the time to start treatment in a neighboring hospital if

patients are diverted. This time is related to multiple state variables of the neighboring hospital, such as crowding conditions, staffing, availability of lab and equipment, and even traffic conditions. The level of information of these variables determines the estimation of the random variable *X* proposed in the model.

In this section, the impact of information on the time to start treatment in a neighboring hospital is analyzed. In particular, various cases are considered where the decision maker has different levels of information on the random variable, *X*. Hence, in addition to the deterministic (D) *X* analyzed in Section 3.3, uniform (U) and triangular (T) distributed *X* are considered, as shown in Table 3.2. For each combination of distribution and expected value, there is one instance that has larger variability than the other. These cases are referred as small (S) and large (L) variability cases.

Table 3.2 Properties of *X* used in the computational analysis.

Type	Distr.	Parameters (mins)	Expected Value (mins)	Range (mins)	Standard Deviation (mins)	CV	Varia-bility
Deterministic		15	15				
		45	45				
		75	75				
Probabilistic	U	(10, 20)	15	10	2.8868	0.1925	S
	U	(5, 25)	15	20	5.7735	0.3849	L
	U	(30, 60)	45	30	8.6603	0.1925	S
	U	(15, 75)	45	60	17.3205	0.3849	L
	U	(50, 100)	75	50	14.4338	0.1925	S
	U	(25, 125)	75	100	28.8675	0.3849	L
	T	(10,15,20)	15	10	2.0412	0.1361	S
	T	(5,15,25)	15	20	4.0825	0.2722	L
	T	(30,45,60)	45	30	6.1237	0.1361	S
	T	(15,45,75)	45	60	12.2474	0.2722	L
	T	(50,75,100)	75	50	10.2062	0.1361	S
	T	(25,75,125)	75	100	20.4124	0.2722	L

The tardiness for uniform and triangular distributions is obtained as follows. Let X be a uniform random variable with parameters a and b, i.e., $f(x) = \frac{1}{b-a}$. Then,

$$T^D_i = \begin{cases} \frac{b+a}{2} - d_i & \text{if } d_i \le a \\ \frac{(b-d_i)^2}{2(b-a)} & \text{if } a < d_i \le b \\ 0 & \text{if } b < d_i. \end{cases} \tag{3.7}$$

Let X be a triangular random variable with parameters a and b and c, where $c = \frac{a+b}{2}$. Then,

$$T^D_i = \begin{cases} \frac{b+a}{2} - d_i & \text{if } d_i \le a \\ \frac{a}{6} + \frac{b}{3} + \frac{d_i}{2} - \frac{(a+b-2d_i)^2(2a-b-d_i)}{6(b-a)^2} & \text{if } a < d_i \le c \\ \frac{2(b-d_i)^3}{3(b-a)^2} & \text{if } c < d_i \le b \\ 0 & \text{if } b < d_i \end{cases} \tag{3.8}$$

Figure 3.6 depicts the threshold curves under medium traffic for the tuple (Distribution, $E[X]$, Variability), where Distribution $\in$ {D, U, T}, $E[X] \in$ {0.25, 0.75, 1.25 hrs} and Variability $\in$ {S, L}.

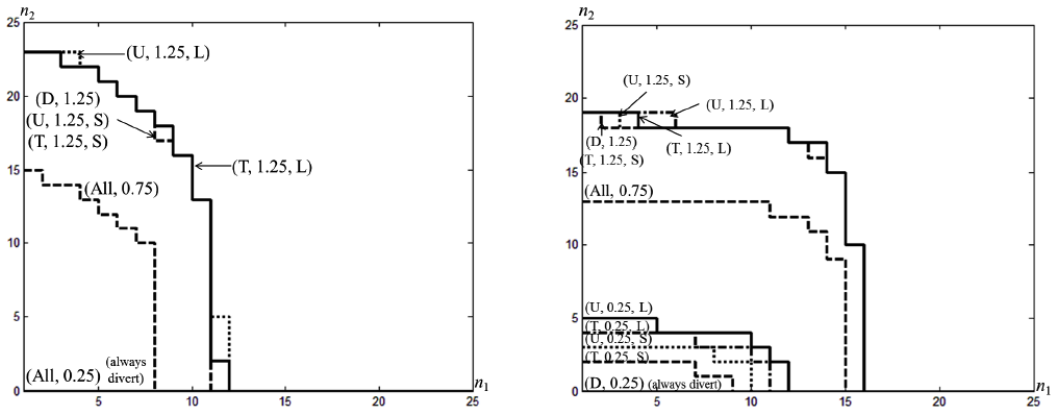

Figure 3.6 Illustration of impact of information on thresholds when congested area is A1, and traffic is medium (*left*); and when congested area is A2, and traffic is medium (*right*).

One clearly observes that the distribution and variability of X typically have a relatively small impact on threshold curves, since threshold curves with the same $E[X]$ generally overlap each other for most of the threshold curve. In particular, for $E[X]$ = 0.75 hours, the same threshold curves are obtained regardless of the distribution and the variability of X. In addition, under some cases, an upward shift in the threshold curve can be observed when the variability increases; this is particularly observed when $E[X]$ = 0.25 hours and congested area in A2. These results can be attributed to the relation of X with the expected tardiness values T^D_1 and T^D_2. Recalling Equation (3.4), a change in the distribution and variability of X may change the value of T^D_1 and T^D_2, resulting in a shift of threshold curves under the view of Theorem 2, and computational analysis results depicted in Figure 3.5. In other words, any change in X that causes T^D_1 and T^D_2 to increase (decrease) may shift the threshold curve in an upward

(downward) direction. Theorem 3 provides a crucial result discussing the impact of distribution of X on the threshold curves.

THEOREM 3. *Let* $\bar{X}$ *and* $\underline{X}$ *be two random variables where* $\bar{X} \geq_{st} \underline{X}$ (*i.e.,* $\bar{X}$ *is stochastically larger than* $\underline{X}$). *Furthermore, let* $\bar{\Delta}(n_1)$, $\underline{\Delta}(n_1)$ *be the threshold curves obtained by solving the problems with* $X = \bar{X}$ *and* $X = \underline{X}$, *respectively. If* $p^A{}_1 = 1$ *then* $\bar{\Delta}(n_1) \geq \underline{\Delta}(n_1)$.

PROOF. Let

$$\overline{T_i^D} = \int_{d_i}^{\infty} (x - d_i) f_{\bar{X}}(x) dx \quad \text{and} \quad \underline{T_i^D} = \int_{d_i}^{\infty} (x - d_i) f_{\underline{X}}(x) dx \tag{3.7}$$

Let the function $g_i(x)$ be

$$g_i(x) = \begin{cases} x - d_i & \text{for } x \geq d_i, \\ 0 & \text{for } x < d_i \end{cases} \tag{3.8}$$

Hence,

$$E\left[g_i\left(\bar{X}\right)\right] = \int_0^{\infty} g_i(x) f_{\bar{X}}(x) dx = \int_{d_i}^{\infty} (x - d_i) f_{\bar{X}}(x) dx = \overline{T_i^D}, \qquad i \in \{1, 2\} \tag{3.9}$$

Similarly, one can also obtain $E[g_i(\underline{X})] = \underline{T_i^D}$, $i \in \{1, 2\}$. From Proposition 9.1.2 of Ross (2004), it is known that $\bar{X} \geq_{st} \underline{X}$ and non-decreasing $g_i(.)$ functions, one has $E\left[g_i\left(\bar{X}\right)\right] \geq E\left[g_i\left(\underline{X}\right)\right]$, $i \in \{1, 2\}$, which gives $\overline{T_i^D} \geq \underline{T_i^D}$, $i \in \{1, 2\}$; therefore $\bar{\Delta}(n_1) \geq \underline{\Delta}(n_1)$. □

Since Theorem 3 follows the result of Theorem 2, it is proven for p_1^A=1. Although, the computational analysis allows analyzing the impact of several X distributions on threshold curves. Theorem 3 provides a more general finding. Given two random variables $\bar{X}$ and $\underline{X}$, with $F_{\bar{X}}(a) \leq F_{\underline{X}}(a)$ for all a, where $F(.)$ denotes the corresponding cumulative distribution functions, one can obtain higher threshold levels for $\bar{X}$. This result also explains the higher threshold levels obtained for instances with higher $E[X]$ values. Similar to Theorems 1 and 2, Theorem 3 is proven for special case p_1^A=1, however, its results are also observed in more general cases.

Table 3.3 presents the changes of T^D_1 and T^D_2 with respect to the distribution of X.

Table 3.3. Values of T^D_1 and T^D_2 (mins)

	$E[X]$ = 0.25 hrs				
	Det	Tria		Unif	
	-	S	L	S	L
T^D_1	0	0.83	1.67	1.25	2.5
T^D_2	0	0	0	0	0
	$E[X]$ = 0.75 hrs				
	Det	Tria		Unif	
	-	S	L	S	L
T^D_1	30	30	30	30	30
T^D_2	0	0	0	0	0
	$E[X]$ = 1.25 hrs				
	Det	Tria		Unif	
	-	S	L	S	L
T^D_1	60	60	60	60	60
T^D_2	0	4.17	8.33	6.25	12.5

As observed in Table 3.3, T^D_1 = 30 and T^D_2 = 0 for all cases under $E[X]$ = 0.75 hours, which explains the same threshold curves obtained when $E[X]$ = 0.75 hours. Furthermore, one can observe that the changes in the threshold curves for

$E[X]$ = 0.25 hours and $E[X]$ = 1.25 hours are due to the changes in T^D_1 and T^D_2, respectively.

Table 3.4 presents the ETP values. This table confirms that the impact of the variability on X is very small compared to the impact of $E[X]$, traffic intensity and RSTT of the congested area. Note that the ETP when $E[X]$ = 0.75 hours is always the same. This is due to the constant values obtained for T^D_1 and T^D_2 for each distribution. Therefore, the ETP increases if $E[X]$ and/or the traffic increase. In addition, if the congested area is A1, then the ETP is significantly larger than the case when the congested area is A2. This is due to the small RSTT of critical patients and the low treatment rate in A1.

In spite of the fact that the results presented in Figure 3.6 are only for medium traffic, similar results can be observed for other values of traffic intensity. Changing the traffic intensity shifts the threshold as in Figure 3.4, but the variability on the distribution still has a small impact. These findings suggest that in cases where the distribution of X is not known, hospital administrators may estimate $E[X]$ and confidently determine diversion decisions based on this expected value.

On the other hand, a bad estimation of X may result in policies that significantly increase the tardiness per patient. Therefore, the level of cooperation among hospitals to share information about their status is important to derive AD policies that work effectively. This implies having information and communication systems that monitor the state of the hospitals frequently and

translates this information to statistics required by other hospitals to decide diverting patients.

Table 3.4. Expected tardiness per patient in minutes for different levels of traffic and different distributions of X

		$E[X]$ = 0.25 hrs				
		Det	Tria		Unif	
Traffic	Cong. Area	-	S	L	S	L
Low	A1	0.58	0.69	0.80	0.75	0.91
	A2	0.21	0.26	0.27	0.27	0.28
Med	A1	4.77	4.89	5.00	4.94	5.11
	A2	2.03	2.14	2.23	2.19	2.30
High	A1	37.78	37.89	38.00	37.94	38.11
	A2	25.79	25.91	26.02	25.96	26.13
		$E[X]$ = 0.75 hrs				
		Det	Tria		Unif	
Traffic	Cong. Area	-	S	L	S	L
Low	A1	2.65	2.65	2.65	2.65	2.65
	A2	0.36	0.36	0.36	0.36	0.36
Med	A1	8.74	8.74	8.74	8.74	8.74
	A2	3.07	3.07	3.07	3.07	3.07
High	A1	41.83	41.83	41.83	41.83	41.83
	A2	29.34	29.34	29.34	29.34	29.34
		$E[X]$ = 1.25 hrs				
		Det	Tria		Unif	
Traffic	Cong. Area	-	S	L	S	L
Low	A1	3.62	3.62	3.62	3.62	3.63
	A2	0.38	0.38	0.38	0.38	0.38
Med	A1	11.98	11.98	12.01	11.99	12.04
	A2	3.36	3.36	3.36	3.36	3.37
High	A1	45.88	45.88	45.92	45.89	45.97
	A2	31.98	31.98	32.00	31.98	32.03

The model presented in Section 3.2 assumes stationary arrival rates and exponential treatment times. However, there is evidence that arrivals to EDs follow a non-stationary pattern. Furthermore, non-exponential distributions may provide a better model for the treatment times in EDs. In addition, it is very likely that congestion in neighboring hospitals is positively correlated; therefore, the value of $E[X]$ might also change throughout the day. In the next section, the

impact of these more realistic assumptions is considered using a simulation model. Simulation is used to evaluate the performance of the policy suggested by the MDP and compare it to other simple policies.

3.5 Simulation of Ambulance Diversion Policies

The MDP proposed in this chapter assumes stationary arrival rates and exponential treatment times. However, there is evidence that these assumptions do not represent the real-life settings. In this section, these assumptions are relaxed and patterns commonly observed in EDs across the United States are explored using a discrete-event simulation model. Furthermore, the AD policy prescribed by the MDP is compared with the following simple AD heuristics:

1. Full Beds in A1 (FB A1): Since most of the ambulance arrivals are critical patients, this policy diverts when all the beds in area A1 are occupied (i.e., when $n_1 \geq c_1$).
2. Full Beds in A1 or in A2 (FB A1/A2). This policy diverts an arriving ambulance when there is at least one area with all the beds occupied (i.e., when $n_1 \geq c_1$ or $n_2 \geq c_2$).
3. Full Beds (FB): This policy diverts an arriving ambulance only when all of the beds in the ED (both A1 and A2) are occupied (i.e., when $n_1 \geq c_1$ and $n_2 \geq c_2$).
4. Myopic policy (Myopic): This policy diverts an arriving ambulance only if the expected tardiness for the current ambulance patient at the neighboring hospital is smaller than the expected tardiness if he/she is accepted. Thus, under the myopic policy, the ambulance is diverted only when $p^A_1 T^D_1 + (1-p^A_1) T^D_2 \leq p^A_1 T(n_1) + (1-p^A_1)\ T(n_2)$. Note that this heuristic evaluates

$T(n_1)$ and $T(n_2)$ under the assumption that treatment times are exponentially distributed.

5. No AD policy (No AD): This policy does not divert patients at any time.

Several sources have identified a pattern in the ED arrivals across the US (Centers for Disease Control and Prevention 2008; Green 2006; Cochran and Roche 2009). This pattern observes low traffic between 1am and 8am approximately. Then, the arrivals increase between 8am and 10am, and remain at a high level between 10am and 11pm. Then, a decline of the arrivals is observed between 11pm and 1am. In order to consider this pattern in the simulation model, the arrival rate pattern used by Cochran and Roche (2009) is adopted. The authors present an hourly multiplicative index that indicates the traffic intensity compared to the average arrival rate. Figure 3.7 is taken from Cochran and Roche (2009), and it shows the change in the arrival multiplicative index throughout the day.

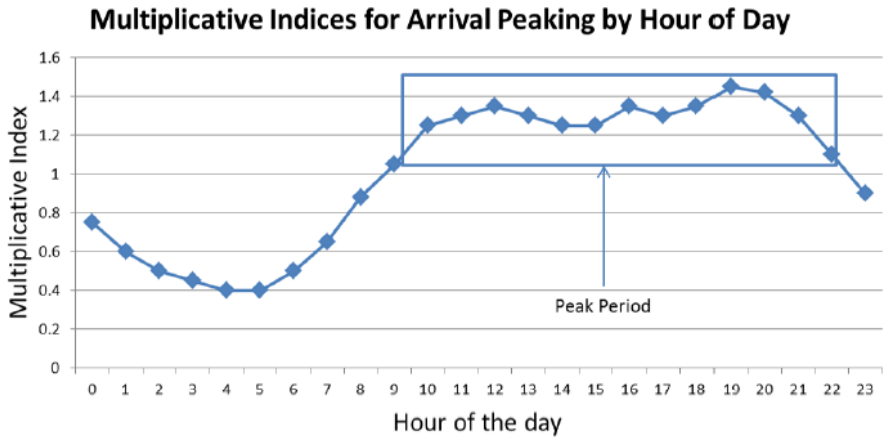

Figure 3.7. Arrival multiplicative indices, adopted from Cochran and Roche (2009).

In order to mimic the arrival pattern of Cochran and Roche (2009), the arrival rates are set as follows. First, λ^W is found such that max{UDWA$_1$, UDWA$_2$}=90%, and this value is set as the walk-in arrival rate for the highest peak hour, which is from 7pm to 8pm with the multiplicative index of 1.45. For example, for the setting (p^A_1, p^W_1) = (0.9, 0.5), the value obtained is λ^W = 6.75, which gives the arrival rate used for 7pm to 8pm in the simulation model. Next, the walk-in arrival rates are scaled using the multiplicative indices to obtain the arrival rates for every hour during the day. For example, for the setting (p^A_1,p^W_1)=(0.9, 0.5), the arrival rate between 1am and 2am, which has a multiplicative index of 0.6, is chosen as (0.6/1.45)6.75 = 2.793. Then, the hourly ambulance arrival rates are calculated such that they represent 15% of the total arrivals to the ED (Centers for Disease Control and Prevention 2010).

Consequently, the arrival rate pattern depicted in Figure 3.8 is obtained. Two different patient mixes (p^A_1, p^W_1) are used: (0.9, 0.5) and (0.9, 0.3), which make A1 and A2 the congested areas, respectively.

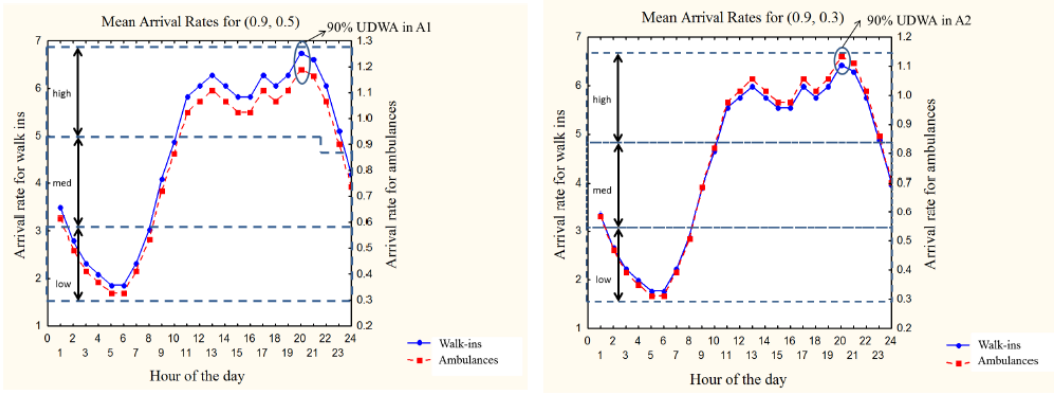

Figure 3.8. Arrival rates to the ED by walk-in patients and ambulances when congested area is A1, p^A_1 = 0.9, p^W_1 = 0.5 (*left*); and congested area is A2, p^A_1 = 0.9, p^W_1 = 0.3 (*right*).

As shown in Figure 3.8, three levels of traffic intensity are defined: low, medium and high. The arrival pattern observed in most EDs across the US and the large percentage of hospitals that go on diversion simultaneously (United States General Accounting Office 2003) suggest that traffic in neighboring hospitals is positively correlated. Therefore, it is very likely that if an ED experiences high traffic, a neighboring hospital also is experiencing high traffic, increasing the waiting time of the diverted patients. Hence, this section assumes that parameters for the distribution of X change depending on the traffic intensity of the ED under study. The random variable, X is assumed to have a triangular distribution with coefficient of variation of 0.2722 and three settings for the parameters of X are tested as shown in Table 3.5.

Table 3.5. Setting of X used in simulation.

	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)

The treatment times in areas A1 and A2 are assumed to be lognormally distributed, which is one of the distributions identified in Hoot et al. (2008) to represent treatment times in healthcare. The expected treatment times used in the simulation model remain 240 minutes and 60 minutes for patients treated in areas A1 and A2, respectively. The standard deviation was adjusted to match the coefficient of variation of treatment times found in Cochran and Roche (2009). Therefore, the standard deviation of treatment in A1 was set to 173.88 minutes, yielding a coefficient of variation of 0.72; and the standard deviation of treatment

in A2 was set to 6.12 minutes, yielding a coefficient of variation of 0.102. The probability density functions of the treatment times are shown in Figure 3.9.

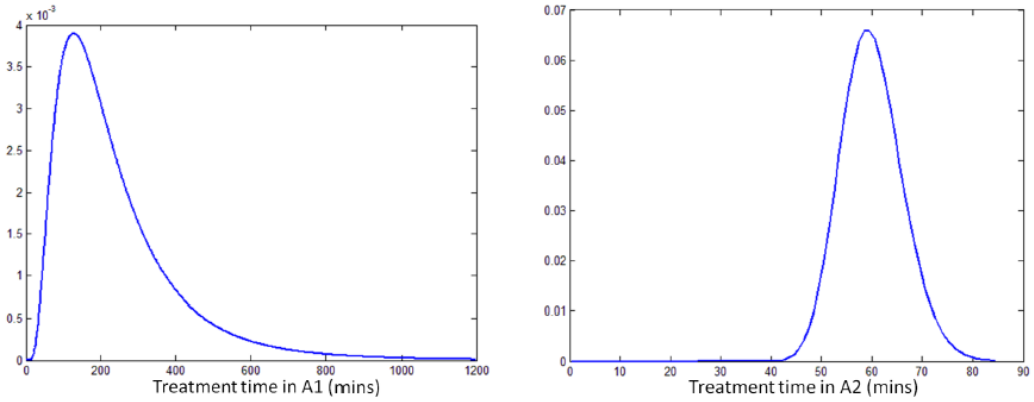

Figure 3.9. Probability density function of treatment time in A1 (*left*); and in A2 (*right*).

In addition to the heuristics listed previously in this section, the simulation model was used to test the policy prescribed by the MDP. In preliminary analysis, four different MDP strategies were tested considering stationary arrival rates of: (i) the average of high traffic hours; (ii) the average of medium traffic hours; (iii) the average of low traffic hours; and (iv) the overall daily average, while the parameters for X are the averages of the parameters throughout the day for each setting. The preliminary experiments show that the MDP solved with the overall daily average rate outperforms all others. Hence, the stationary arrival rate for the MDP model is chosen as the daily average in order to derive the optimal AD threshold. For each simulation setting, the MDP is solved first, and then the obtained AD control policy is implemented in the simulation model to obtain an estimate for its performance.

Each policy with all possible combinations of severity mix and setting for *X* was modeled in simulation models developed using Arena (Kelton et al. 2007). Pilot runs were used to determine a warm-up period of two months, replication length of one year and 30 replications in order to capture the performance of the system in steady state and estimate the average tardiness per patient using 95% confidence intervals with an average relative precision of 3.69%. In addition, common random numbers were used to reduce noise when comparing alternative AD policies (Banks et al. 2010). Figure 3.10 presents the confidence intervals for settings 1 and 3 of *X*, given in Table 3.5. The results for setting 2 are not shown in this figure because they fall somewhere between the results from settings 1 and 3.

Even though the simulation model includes several relaxations that invalidate the optimality of the policy suggested by the MDP, the policy prescribed performs consistently well in all scenarios compared with other heuristics. The FB A1 is a policy that also works consistently well in all the scenarios and, for some of them, there is not a significant difference in the performance compared with the AD control policy prescribed by the MDP. This heuristic works well because it takes advantage of the fact that most ambulance patients are critical; and hence, the policy tries to avoid queuing in the critical care area.

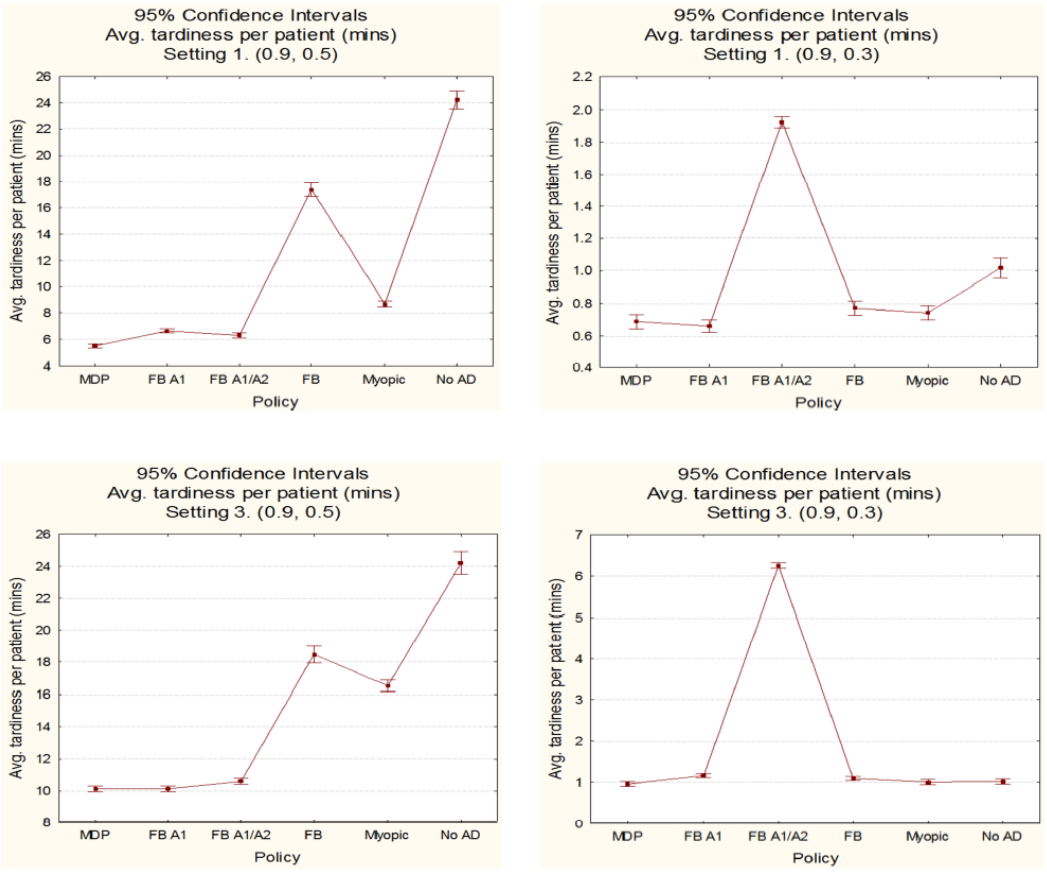

Figure 3.10. 95% Confidence intervals on the average tardiness per patient for Setting 1, Congested Area: A1 (*top left*); Setting 1, Congested Area: A2 (*top right*); Setting 3, Congested Area: A1 (*bottom left*); Setting 3, Congested Area: A2 (*bottom right*).

On the other hand, the policy that diverts when at least one area has all of the beds occupied (i.e., FB A1/A2) performs reasonably well if the congested area is A1, but it has the worst performance among heuristics if the congested area is A2. This is due to the fact that if the congested area is A1, then this policy avoids having several critical patients waiting for a bed; however, if congested area is A2, then the policy is very likely to start diverting when all the beds in this area are occupied, therefore it does not take advantage of the relatively high treatment rate of A2. The full occupancy policy (FB) does not perform as well as other

heuristics, especially if congested area is A1. In this case, the policy might delay diversion until beds in the non-congested area (A2) are fully occupied; therefore, some critical patients have to wait in the ED. The performance of the Myopic policy in comparison to other policies is better if congested area is A2 than if it is in A1. This may be due to the assumption of exponential treatment times used to compute the expected tardiness using Equation (3.3) before deciding if a patient is diverted or accepted.

Most of these heuristics work significantly better than No AD. Not diverting patients can produce a high average tardiness per patient, especially if congested area is A1, where the critical patients are treated. For this case, some of the AD policies, including the suggested by the MDP, can reduce the average tardiness by more than 10 minutes, which could make a significant difference in terms of mortality rate in critical patients. Therefore, these results suggest that intelligent design of AD policies can reduce the time to deliver appropriate treatment to patients, even if the time to start treatment in a neighboring hospital is relatively large.

Table 3.6 presents the relative performance of the heuristics, taking the policy prescribed by the MDP as a basis. The relative performance confirms that the MDP policy is significantly better than the other heuristic, except for FB A1 in some scenarios.

Table 3.6 Relative performance of heuristics compared with MDP (%).

	Setting/Congested Area					
	Setting 1		Setting 2		Setting 3	
Heuristic	A1	A2	A1	A2	A1	A2
FB A1	20.85	-4.17	5.59	3.30	-0.03	21.56
FB A1/A2	14.49	178.92	5.09	333.18	4.58	554.6
FB	216.25	11.84	135.85	8.46	82.31	14.43
Myopic	57.64	7.71	42.37	4.41	62.97	5.03
No AD	339.87	47.20	220.42	23.04	138.7	6.25

In order to determine which policy is the best for each scenario, MDP and FB A1 are compared using hypothesis testing on the difference of their means using 95% confidence level. There is significant evidence that the policy prescribed by the MDP is better than FB A1 for setting 1 and congested area A1, setting 2 and congested area A1, and setting 3 and congested area A2 scenarios. For setting 1 and congested area A2, setting 2 and congested area A2, and setting 3 and congested area A1 scenarios, there is not a significant difference on the performance of these two policies. Even though the FB A1 policy might be easier to implement, the threshold suggest by the MDP still works significantly better than FB A1 in several scenarios. Furthermore, the difference could be significantly high, like in the case of setting 1 and congested area A1 where suboptimality of FB A1 is more than 20%.

An important aspect to highlight from Table 3.6 is the large difference of the relative performance when comparing the columns that defines the congested area for the same setting. For example, the heuristic FB A1/A2 performs only 14.49% worse than the threshold prescribed by the MDP under setting 1 and congested area A1; but the same heuristic performs 178.92% worse than the MDP under same setting and congested area A2. Therefore, the knowledge of the

severity mix that defines the congested area is a key parameter that determines the effectiveness of an AD policy.

The results show that effective design of AD policies can decrease the average tardiness per patient significantly, even if the neighboring hospital is far away or crowded like in the case of setting 3. However, inappropriate heuristics can lead to a worse performance than not diverting at all, like in the case of policy FB A1/A2 and settings with congested area A2. In addition, the simulation model confirms that ambulance diversion is more likely to have a significant impact if the congested area is A1.

The MDP proposed in this chapter prescribes AD thresholds that perform consistently well despite the relaxation of important assumptions. However, the AD policy prescribed by the MDP may lead to a situation where the ED goes on and off diversion very often. Furthermore, the proposed Bellman's equation does not consider the percentage of time spent on diversion, which is an important performance measure for the EDs. The next section discusses insights about these aspects and presents a simple heuristic to avoid changing diversion status too frequently.

3.6 Insights on Implementation of AD Policies Prescribed by the MDP

Typically, real-life AD is implemented such that the ED maintains the diversion status for a predetermined period in which ambulances are diverted to other hospitals. In contrast, the optimal AD control policies prescribed by the MDP model comprise a single threshold that determine accepting or diverting individual ambulance arrivals. Hence, there may be some downsides of this

approach if implemented in practice: (i) EDs could go on and off diversion very often, increasing the cost of communicating with Emergency Medical Services; (ii) an ambulance could be rejected and another could be accepted within a short time frame, which may seem to be unethical to practitioners; (iii) the AD policy produced by the MDP formulation requires continuous monitoring of the state of the system.

In order to overcome these issues, a new heuristic is presented, MDPτ, that requires the diversion status to last for at least a predetermined duration of τ. This policy implements the threshold prescribed by the MDP to determine when to initiate the diversion status. Once the ED goes on diversion by exceeding the threshold, the ED maintains the diversion status for the next τ time units. After τ time units, the state of the system is evaluated. If the state of the system is above the threshold curve according to the prescribed MDP policy, then the diversion status is maintained for another τ time units. Otherwise, the ED removes the diversion status.

The remainder of the section analyzes these policies for the case where the congested area is A1 because AD is more effective in this scenario. In addition, setting 2 of Table 3.5 was chosen for analysis because it implies moderate values for the parameters of X; however, similar observations are made for other settings. Figure 3.11 shows the average tardiness per patient and the average number of diversion episodes per day for the MDP policy and the MDPτ policies, with $\tau \in \{30, 60, 90, 120\}$ (in minutes).

In general, the average tardiness per patient resulting from the MDPτ policies is greater than that of the policies prescribed by the MDP. However, the differences are quite small and often insignificant. MDPτ policies are not only more suitable to be implemented in practice; but also, they reduce the average number of diversion episodes per day significantly, as observed in Figure 3.11(*right*), and hence, they may avoid ethical problems related to admission control in emergency care.

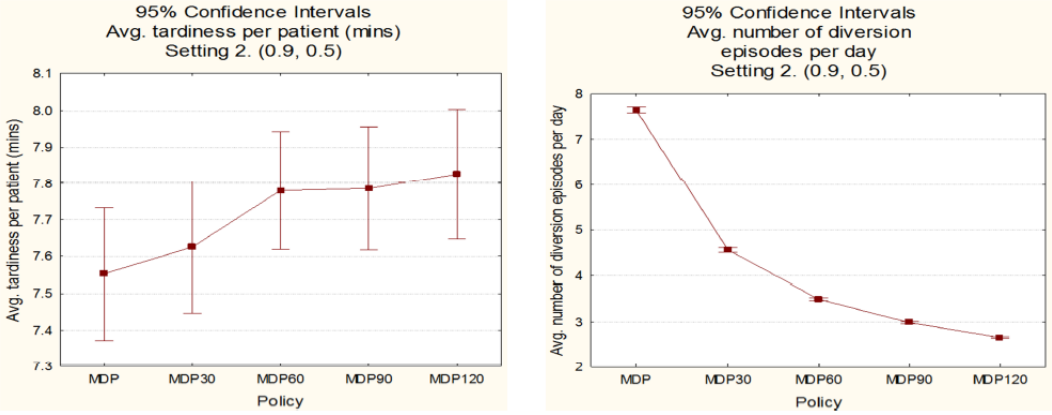

Figure 3.11. Performance of the MDP prescribed policy and MDPτ policies in terms of tardiness (*left*); and number of diversion episodes (*right*).

Decision makers in practice have the objective of providing timely care to patients requiring emergency care, as well as minimizing the duration of the diversion episodes and the fraction of time spent on diversion. The diversion episode length refers to the duration of the diversion status every time that the ED goes on diversion. Since the formulation presented in this chapter does not penalize being in the diversion status, the policies prescribed by the MDP may result in long diversion durations (particularly when treatment times at the other hospital are short).

In Figure 3.12, the average fraction of time spent on diversion and the average diversion episode length for the MDP and MDPτ policies are presented, using again setting 2 for the distribution of X and congested area in A1.

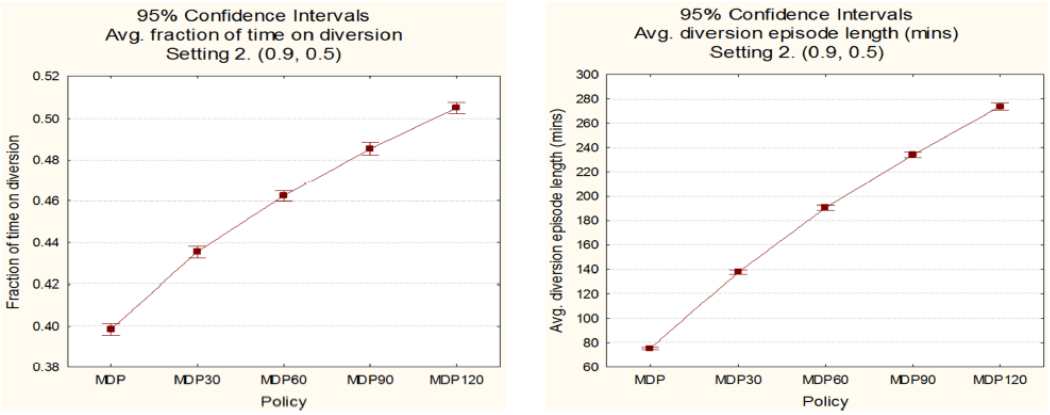

Figure 3.12. Performance of the MDP and MDPτ policies in terms of average fraction of time on diversion (*left*); and average diversion episode length (*right*) for c_1=15 beds.

An increase of τ results in more undesirable outcomes in terms of both metrics is observed. Furthermore, the fraction of time on diversion of the MDP policy is significantly higher than the values observed in real settings (less than 20% according to United States General Accounting Office (2003)). Therefore, the proposed MDP policy improves the performance of EDs in terms of ETP, however, it increases the fraction of time on diversion and the average diversion episode lengths, which may be undesirable for EDs.

In order to find long-term solutions that improve performance in both metrics, the decision makers must address the root cause of the problem, which may be the insufficient capacity to provide emergency care. Figure 3.13 presents a sensitivity analysis varying the number of beds in A1 and observing the impact on

the optimal average tardiness per patient and the fraction of time on diversion (under the optimal control policy) for setting 2.

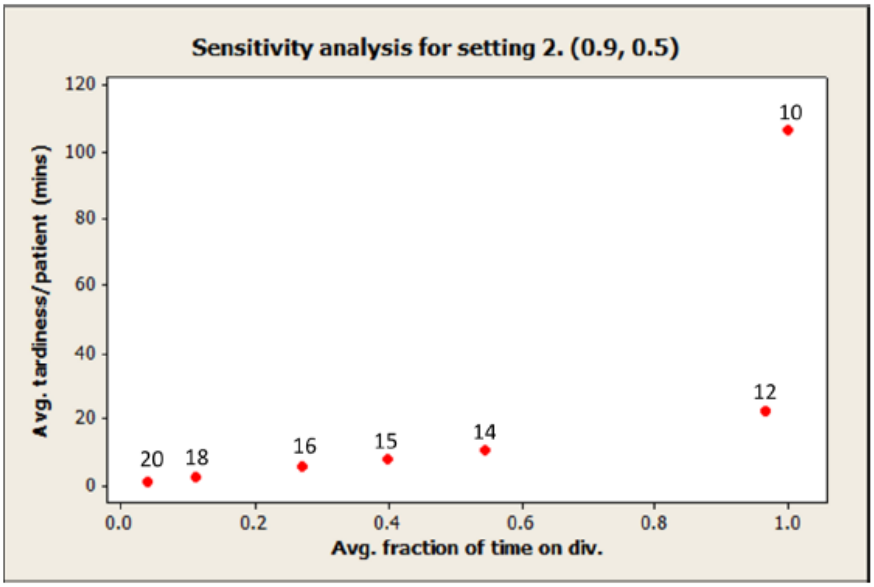

Figure 3.13. Fraction of time on diversion vs. average tardiness per patient for different number of beds in A1 and considering the MDP prescribed policy.

Capacity has a significant impact on the performance of the ED. Adding beds to A1 reduces the optimal average tardiness per patient, which approaches to zero as the number of beds becomes sufficient to serve the demand. In addition, adding beds reduces significantly the fraction of time on diversion under the optimal control policy. The MDP model may prescribe being always on diversion when the walk-in demand exceeds the capacity during the peak time. For example, during the peak period, the average arrival rate to A1 from walk-ins only varies between 2.56 and 3.375 patients per hour; while the average treatment rate, when there are only 10 beds in A1, is 2.5 patients per hour. On the other

hand, diversion is used less as the capacity becomes sufficient to serve all the arrivals.

Thus, under setting 2 and assuming c_1=10 beds, the best average tardiness per patient that the ED under study can obtain is 106 minutes, but this implies being almost always on diversion, and accepting only walk-ins. On the other hand, the ED can achieve an average tardiness per patient of 25 minutes, with two more beds in A1, but again, would be on diversion for a very a large fraction of time. In order to reduce both metrics simultaneously, the ED needs to invest further in critical care beds. For example, assuming 18 beds in A1, the ED could achieve an average tardiness per patient of only 2.45 minutes while being on diversion only 11% of the time. Hence, strategic planning of EDs must consider the capacity of the EDs to estimate the size and amount of resources required to reach a desired or acceptable level of tardiness and fraction of time on diversion.

3.7 Conclusions

This chapter presents an MDP model to determine the AD policy that minimizes the long-run average tardiness per patient for a single ED. Tardiness is defined as the amount of time that the patients wait beyond the recommended safety time threshold. The model considers two treatment areas, differentiated by patient severity, and assumes availability of (some) information on the time to start treatment in a neighboring hospital if an ambulance patient is diverted.

The structural properties of the model indicate the existence of a threshold curve “above" which one should divert ambulances. This threshold curve is sensitive to traffic intensity, severity mix and expected time to start treatment at

the neighboring hospital. Based on analytical results and computational experiments shown in this chapter, it is observed that the threshold curve is non-increasing in traffic intensity and non-decreasing in the expected tardiness experienced if patients are diverted to a neighboring hospital. Moreover, the optimal AD policy is more sensitive to changes in parameters, and can result in a larger reduction on the average expected tardiness if the critical care area has higher utilization than the fast-track area (A2), considering that most ambulance patients need immediate and emergent care. On the other hand, the variability regarding the time to start treatment in a neighboring hospital has a relatively small impact on the definition of the threshold and on the optimal value.

Depending on the traffic intensity and congested area, the optimal AD thresholds could allow queuing before diverting an ambulance patient, or it could save the beds for future demand. These variations on the optimal policy are found in particular for the case where A1 is the congested area. If the ED has high walk-in traffic and a significant proportion of them are critical patients, then the optimal AD policy initiates diversion earlier, saving beds in the critical area for future walk-in demand. Even though these types of actions might not be well received by the medical community, the policy could be adapted and implemented under specific conditions in order to increase the accessibility to emergency care for patients that go to an ED on their own.

Even though the MDP includes assumptions that might not be realistic in real-life settings, the optimal AD policy works consistently well under the incorporation of time-dependent arrival patterns and non-exponential treatment

time distributions. Furthermore, the policy prescribed by the MDP works significantly better than several policies applied in practice, including not diverting at all. Hence, effective AD policies can contribute to increasing patient safety by minimizing the time that they wait beyond their safety time threshold.

The AD policy prescribed by the MDP may have downsides for implementation. For example, the AD threshold curve may cause going on and off diversion relatively often. This chapter addresses this concern by proposing a heuristic that re-evaluates the diversion status at given time intervals, after going on diversion. The average tardiness per patient produced by this heuristic is slightly greater than the MDP prescribed policy, but it reduces significantly the number of diversion episodes per day.

On the other hand, the proposed formulation does not take into account some performance metrics of EDs such as the fraction of time on diversion and the average diversion episode length. Hence, the proposed MDP policy may prescribe policies that may result in relatively long diversion episodes, and relatively frequent diversions. A sensitivity analysis made on the number of beds in A1 shows that capacity should be addressed in a strategic manner in order to have significant improvements in both timeliness (measured by, e.g., average tardiness per patient) and in accessibility to emergency care (measured by, e.g., fraction of time on diversion).

The implementation of the policy prescribed by the MDP requires first the knowledge of the input parameters, which include the number of beds and treatment times. In addition, the computation requires knowledge about the

expected value of the time to start treatment in a neighboring hospital. Decision makers might be able to make an educated guess; however, collaboration among hospitals is encouraged to have better results systemwide. Information systems that can provide accurate information about the state of the hospital in real time would be a great support to assure that the AD policy is followed as recommended by the model.

In summary, this chapter demonstrates that an optimal design of AD policies can be an effective strategy to reduce the delays in receiving emergency care, which can potentially lead to significant reductions in mortality and morbidity. Decision makers, including hospital administrators and public health officers can design better policies by considering the proposed model. The next chapter extends this research for designing the optimal AD policies for multiple hospitals. Since MDP suffers from scalability issues, genetic algorithms combined with simulation is explored.

CHAPTER 4

CENTRALIZED DESIGN OF AMBULANCE DIVERSION POLICIES FOR MULTIPLE HOSPITALS

4.1 Introduction

Several reports, papers and articles highlight issues faced by emergency care delivery system in the United States. These issues include long periods waiting in emergency departments (EDs), high number of patients boarding, excess of transportation time by diverted ambulances, etc (American College of Emergency Physicians 2008; Asplin 2003). The most concerning consequences of these problems are adverse events that can increase the morbidity and mortality in patients (Green 2008; Pham et al. 2006).

Patient allocation in an emergency care delivery system can be an alternative to reduce congestion from EDs and avoid periods of inappropriate level of care. However, healthcare organizations do not have the mechanisms to control where patients go, except for those transported by ambulance. Thus, walk-in patients decide which ED to visit if they require emergency treatment, but ambulances can take patients to the most appropriate facility according to their health state and the state of the surrounding EDs. This chapter proposes the centralized design of ambulance diversion policies as part of an ambulance flow control mechanism that includes also ambulance destination policies. The combination of these two types of policies is referred as a patient allocation strategy in an emergency care delivery system (ECDS). The objective of the proposed ambulance flow control is to minimize the average time that patients

spend in activities that do not provide appropriate treatment through different stages of care. These activities, which are called non-value added activities in this chapter, include transportation, waiting in the ED and boarding in the ED waiting for a bed in an inpatient unit. Even though ambulance flow control acts only to a small proportion of all the patients visiting EDs (about 15% of all the arrivals to EDs are ambulances according to the Centers for Disease Control and Prevention (2010), an effective allocation can smooth the patient flow in the entire system because ambulance patients produce significant disruptions due to their characteristics, such as high priority level, long treatment times and high admission probability.

The remaining parts of this chapter are organized as follows. Section 4.2 presents findings in related literature. Section 4.3 describes the discrete-event simulation model built for an emergency care delivery system. Section 4.4 presents the methodology proposed for a centralized design of AD policies using GA. Section 4.5 describes the experimentation framework and shows the results of two case studies. The limitations of this research are discussed in Section 4.6 and finally conclusions and future extensions are presented in Section 4.7.

4.2 Literature Review

Ambulance Diversion is a way to relieve congestion from overcrowded EDs. However, diverting ambulance is also a problem because of the increase transportation. One of the first reports highlighting AD as an issue for healthcare delivery systems is the report submitted by the General Accounting Office to the US Senate in 2003 (United States General Accounting Office 2003). This report

observed a high incidence of diversion in statistical metropolitan areas. The main conditions identified as contributors to diversion include the inability to transfer patient from the ED to critical care beds, to telemetry beds or to other inpatient beds (Centers for Disease Control and Prevention 2006b). In addition, other studies identify the high number of patients waiting in the ED and high number of patients boarding as factors to trigger the diversion status (American College of Emergency Physicians 2008; Pham 2006).

The American College of Emergency Physicians recommends avoiding AD because of the potential consequences of longer transportation over the health condition of the patient (American College of Emergency Physicians 1999; American College of Emergency Physicians 2008). Hence, efforts to reduce diversion are available in literature as empirical studies. Furthermore, some of these studies suggest the hypothesis that periods on AD increase because of a reciprocating effect. Thus, if one facility goes on diversion, the surrounding EDs experience an overflow of incoming patients, which forces to these facilities going on diversion as well.

Vilke et al. (2004a) designed a plan to observe the reciprocating effect of AD in San Diego County. Two neighboring emergency departments were exposed to an experiment that restricted one of them of going on diversion. The team observed a significant reduction on the time spent on diversion on both hospitals. However, after experimentation and withdrawal of the constraint, the diversion episodes returned to their usual level. The authors of this study conclude that reciprocating effect is an important factor to observe as a contributor to AD.

Then, Vilke et al. (2004b) expanded the scope of the project to a two-year study that included 21 EDs and also the participation of the San Diego County Medical Society, paramedic agencies, the San Diego County Division of Emergency Medical Services and the local health care association. The group re-designed the AD guidelines of the hospitals restricting the diversion status. Under the new guidelines, the results showed a significant reduction in the mean hours spent on diversion per month in the whole region. The mean numbers of hours on diversion per month for the pre-trial, trial and post-trial periods were 4007, 1079 and 1774, respectively. The trial period refers to the period when the intervention started and post-trial refers to a control period. The authors conclude that a more restrictive AD guideline to go on diversion can reduce significantly the amount of time spent on diversion and increase the access to the facility requested.

A similar study is presented by Asamoah et al. (2008). This chapter presents a study about the implementation of a new AD protocol in a county of 600 000 people and 10 hospitals. The new protocol restricted the time spent on diversion to only one hour out of every eight. The mean number of hours on diversion in the system per month was 305, 275 and 54 for the pre-trial, interim and post-trial periods, respectively. Authors also found a small, but significant increase in the time that it takes for EMS personnel to become available for service after arrival with a patient to an ED (from 21.1 to 22.8 minutes). The authors conclude that a strict protocol that regulates the duration of AD can improve the accessibility to emergency care.

Patel et al. (2006) presents a diversion protocol for Sacramento, CA. This protocol establishes that facilities may go on AD only if ED cannot care for additional patients and it restricts the number of consecutive hours on diversion to three. The results of the analysis of this protocol carried out during 3-years with the participation of 17 hospitals include a reduction of the number of AD hours on diversion of the system by 1428 hours per month, which represents a decrease of 75%.

It is evident that these papers shows successful designs of AD policies to reduce diversion hours in multi-hospital systems, but they lack to in-depth quantitative study to assess the impact of this strategy on patients including the patient average waiting times within each facility or the number of patients boarding.

On the other hand, there are papers that analyze AD from an analytical perspective. One paper presented by Hagtvedt et al. (2009) proposes a game theoretical approach to analyze the behavior of hospitals regarding AD. The authors introduce a payoff function that includes the difference between the ideal and real loads in a hospital, plus a penalty for being on diversion. This payoff function was used to formulate the Prisoner's Dilemma to a system with two hospitals, where the decisions are going on diversion, or not going on diversion. The authors conclude that there is an incentive to go on AD if inflow is sufficiently higher than the ideal load, which could make cooperation difficult in a multi-hospital context; hence, in order to force cooperation among a multiple

hospitals, the system should include an external agent that regulates the AD strategies.

Deo and Gurvich (2011) present a queuing network formulation to analyze the effect of AD on the average waiting time within each ED. The authors found that AD could take advantage of the resource pooling effect in the system and a centralized definition of AD policies can be Pareto improving compared to not diverting at all. Since the optimal threshold is difficult to characterize, the authors introduce the number of beds as an AD threshold that can yield effective results.

The existing literature showing empirical studies does not show the potential benefits that could be gained with an effective design of diversion and destination policies. Moreover, these studies search for minimizing or avoiding diversion. However, there is evidence that not diverting patients from overcrowded hospitals can be risky for the health status of the patients. On the other hand, there is evidence from queuing formulations that suggest that diversion can reduce the average waiting time in the system. However, the assumptions made by queuing theory do not allow exploring the impact of other performance measures, such as transportation time, boarding time, etc.

This chapter proposes the effective design of ambulance diversion policies combined with destination policies to allocate ambulances to EDs in an ECDS. The methods proposed include using simulation and genetic algorithm to design the diversion policies for all the hospitals in the ECDS. Unlike models in the existing literature, the proposed model considers aspects that determine the effectiveness of the policies, such as: non-stationary arrival rates, severity levels

of patients and priorities, different treatment times and admission probability to hospitals and transportation time. In addition, the diversion policies explored in this chapter includes multiple state variables. Furthermore, the model includes destination policies that could also have an impact in the performance of the ECDS.

4.3 Emergency Care Delivery System Model

The model of the emergency care delivery system (ECDS) of this chapter is a discrete-event simulation model that comprises multiple hospitals which serve a geographical region. Each hospital includes an ED and an Inpatient Unit (IP). The simulation of the ECDS is executed through three main modules: the emergency patient generator, the ambulance destination decision and the hospital simulation module.

First, the emergency patient generator module creates patients with the need of ambulance transportation to one of the EDs. Thus, this module schedules the appearance of new patients and assigns a random location in the geographic zone. Next, an ambulance destination decision module determines the destination of the patient. This module observes the candidate destination hospitals based on their diversion status and the appropriate hospital is selected depending on a pre-defined destination policy. Then, the arrival of the ambulance patient to the selected ED is scheduled.

On the other hand, the hospital simulation module keeps each hospital operating according to their events. Besides the ambulance arrivals, each hospital receives walk-ins and direct admissions independently from the other hospitals.

The EDs might start a diversion period if the conditions prescribed in a diversion policy are satisfied. If a hospital goes on diversion, then it is removed from the candidate list of potential destination for other ambulance patients until the diversion status is back off. The general overview of the model is shown in Figure 4.1.

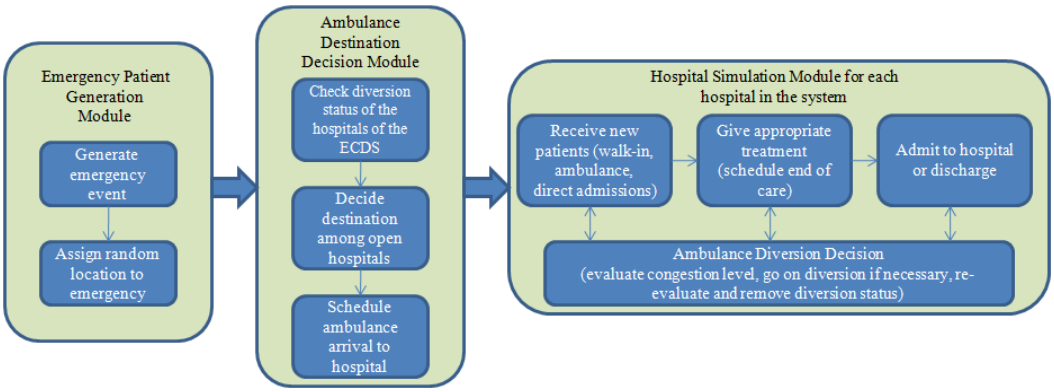

Figure 4.1. Overview of the simulation model.

There are two important assumptions regarding the acceptance of patients while a hospital is on diversion. First, if the ambulance destination decision module determines that a patient is taken to a specific hospital and that hospital goes on diversion before the patients arrives, then the patient is still received at that hospital. This assumption avoids redirecting an ambulance to another ED while it is on the road to the destination hospital. The second assumption avoids that all the hospitals in the ECDS go on diversion at the same time. Thus, if the last hospital off diversion in the ECDS observes that the condition to start a diversion period is reached, then all the hospitals in the ECDS go off diversion. This assumption avoids that a patients has to be taken to another region, which is an undesired aspect in real settings (Arizona Emergency Medical Systems 2000).

A generic model was built for each hospital, which comprises an ED and one inpatient unit whose main resources are the beds where patients receive treatment. Similar models can be found in Cochran and Bharti (2006), Kolker (2008) and Hoot et al. (2008). The main sources of information for the input data are Cochran and Bharti (2006), Cochran and Roche (2009) which present relevant data of hospitals located in Maricopa County, AZ. Additional information was obtained from the National Health Statistics Reports of the Centers for Disease Control and Prevention.

The patient flow inside each hospital is depicted in Figure 4.2. Patients arrive to the ED walking in or by ambulance; upon arrival to the ED, the patient is classified in one out of five severity levels, whose probability depends on the arrival mode. The five-level severity system has become the standard in many countries (Agency for Healthcare Research and Quality 2005) and some statistics are usually published in the Centers for Disease Control and Prevention reports. Patients with severity level 1 are the sickest patients and receive the highest priority, while patients with severity level 5 have the lowest priority. All the patients that go to the ED receive treatment in one bed. The mean treatment time depends on the severity level (Cochran and Roche 2009; Centers for Disease Control and Prevention 2006a). If all the beds are occupied, then the patients have to wait in a queue. As a bed becomes free, another patient starts receiving treatment. Patients are assigned to a bed considering first the priority and then first come – first served is the tie-breaker. If a patient waits too long for a bed, then the patient leaves the facility without receiving treatment (LWOT). After

ending treatment in the ED, the patients require admission to the inpatient unit with a probability that depends on the severity level. If all the beds in the IP are busy when admission from the ED is required, then the patient have to board in the ED bed until a bed in the IP unit opens. Beside admissions from the ED, the IP unit also considers direct admission. The patients are discharged after ending treatment in the IP unit or after ending treatment in the ED without admission.

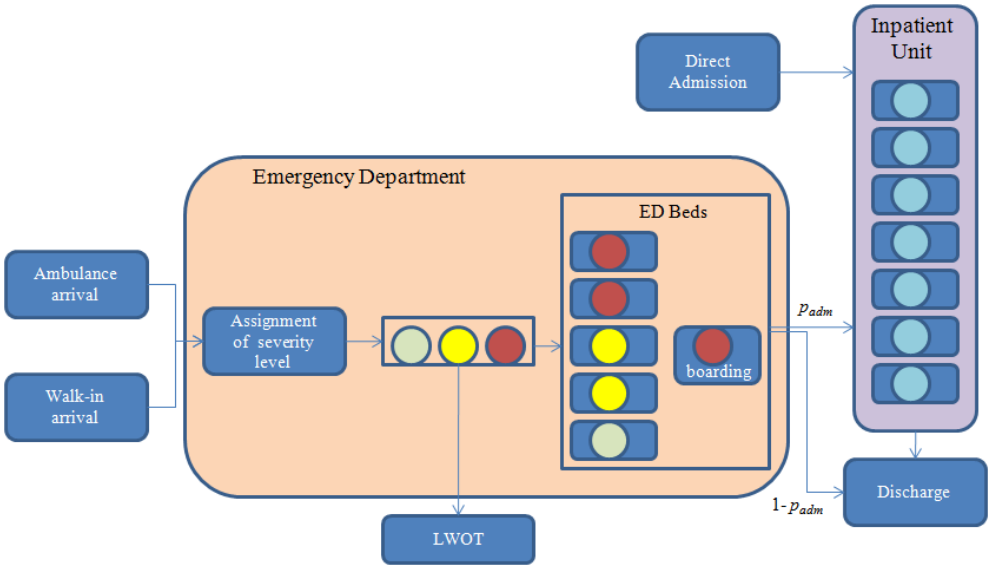

Figure 4.2. Patient flow inside each hospital.

The hospitals in the model include patterns observed in the real setting, such as hourly pattern of arrivals to ED, treatment times that depend on severity level and operational issues as congestion, boarding and patients leaving without treatment. The performance of the model of the generic ED is similar to the data presented in papers and reports. For example, Table 4.1 presents 95% confidence intervals on the main metrics of the proposed simulation model without diversion and compares them with available information of other EDs.

Table 4.1. Performance of generic ED compared with other EDs used as references.

Metric	Simulation Output 95% CI	Validation
Average waiting time	67.51 ± 8.46 minutes	55.8 minutes[a], 56 minutes[b]
Percentage of patients that left without treatment	0.92 ± 0.45	2% [a]
Average ambulance transportation time	6.84 ± 0.08 minutes	8 minutes[c]

[a]Centers for Disease Control and Prevention (2008)
[b]General Accounting Office (2009)
[c]Petzall et a. (2011)

There is very little information about national data regarding average boarding time because data is not collected or it is not available from the patient records (United States General Accounting Office 2009). The 95% confidence interval of the average boarding time for the generic ED simulated is 51.52 ± 9.3 minutes. According to the United States General Accounting Office (2003), the percentage of hospitals whose average boarding time is less than 2 hours is about 10%. Appendix B presents more details about input data.

4.4 Centralized Design of AD Policies

The centralized design of AD policies is a key factor for achieving the potential benefits of ambulance flow control. As highlighted in the literature review, an independent design of AD policies can have undesirable consequences, including long periods on simultaneous diversion in the system that can significantly increase the transportation time (Deo and Gurvich 2011; Vilke et al. 2004a).

This research proposes combining simulation and genetic algorithm (GA) to design the AD policies for all the hospitals in the ECDS and to find Pareto improving policies. In order to design effective AD policies from a centralized perspective, the chromosome structure of the GA comprises the diversion policies of all the hospitals in the system, as depicted in Figure 4.3. These AD policies, along with the destination policies, are evaluated using discrete-event simulation.

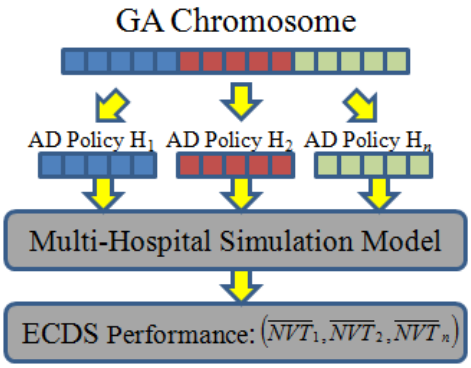

Figure 4.3. Centralized design of AD policies using GA.

The evaluation of the effectiveness of a strategy that implements a specific set of AD policies combined with a destination policy is through the vector that comprises the average patient non-value added time for each hospital: ($\overline{NVT}_1$, $\overline{NVT}_2$,…, $\overline{NVT}_n$), where n is the number of hospitals in the ECDS. For each hospital, the average patient non-value added time is:

$$\overline{NVT}_i = p_i^a \bar{T}_i + \sum_{k=1}^{5} w_{k,i} \bar{W}_{k,i} + p_i^{adm} \bar{B}_i \quad (4.1)$$

where,

$\overline{NVT}_i$: average patient non-value added time in hospital H_i.

p_i^a: fraction of ambulance arrivals to hospital H_i.

$\bar{T}_i$: average transportation time of ambulance patients received at hospital H_i. This includes patients whose final destination is H_i, and patients diverted from H_j to H_i, for all $i \neq j$.

$w_{k,i}$: weight given to the average waiting time of patients with severity level k in hospital H_i.

$\bar{W}_{k,i}$: average waiting time of patients with severity level k in hospital H_i.

p_i^{adm}: Fraction of ED patients admitted to hospital H_i.

$\bar{B}_i$: Average boarding time in hospital H_i.

The first term in the right side of Equation (4.1) takes into account only the ambulance patients. The second term is a weighted average of the waiting time of all the patients that went through the ED, except for those that left without treatment. The weights are defined by the decision maker and the purpose is to give more importance to the waiting of the most urgent patients than the least urgent. The third term of the computation considers information of admitted patients by including the boarding time.

Multi-objective genetic algorithm is presented to define the AD policies that produce Pareto improvements on the ECDS. This methodology allows the generation of new set of policies through recombination and mutation of chromosomes. Then, the set of AD policies of the ECDS is evaluated with the discrete-event simulation model described in Section 4.3. The fitness of the policy considers the tuple of average-patient non-value added times and selects the best fitted chromosomes to survive to the next generation. The NSGA-II algorithm,

proposed by Deb et al. (2002), is implemented in this research and its process is summarized in Appendix B.

The tuple of average-patient non-value added time and the calculation of each element are aligned with the objectives of this chapter. First, these calculations take into account the activities that patients spend in inappropriate treatment at different stages of emergency care, which are: transportation, waiting and boarding. Second, another objective of the chapter is to use strategies based on ambulance diversion and destination policies as ambulance flow control strategies to minimize the disruption of ambulance patients on the entire emergency system. This objective is achieved by including the data of all the patients in the ECDS in the calculations of the performance vector. Thus, Equation (4.1) includes information for ambulance and walk-in patients. Since ambulance patients are very likely to receive high priority, to have long treatment times and to be transferred to IP, the Pareto improvement of the performance vector implies finding the appropriate AD and destination policies that helps smoothing the patient flow in the ECDS. Hence, this allocation of ambulance patients reduces significant delays for all the patients.

4.4.1 Definition of ambulance diversion and destination policies

There are two methods of interest in this research than to control the flow of emergency patients: ambulance diversion and ambulance destination policies. The ambulance diversion policy defines the conditions in a hospital that start a period of full ambulance diversion. In addition, it considers the criteria to remove and revaluate the diversion status. Typically, ambulance diversion policies are

based on observing crowding indicators that alert about potential congestion in the facility and a risky situation for new patients (Arizona Emergency Medical Systems 2000).

On the other hand, the ambulance destination policy defines the hospital where a patient is taken. The selection is based on the open hospitals (hospitals off diversion) and current status, such as distance, crowding variables, etc. Usually, the emergency medical system (EMS) team makes this decision, but sometimes the patient is able to decide. The first part of the experimentation process takes into account three strategies of diversion policies and two destinations policies, which are described below.

Ambulance Diversion Policies:

- No Ambulance Diversion (No AD). This strategy does not allow hospitals going on ambulance diversion at any time.
- Optimized Single-Factor Ambulance Diversion Policy (SF AD). This strategy implements an ambulance diversion policy for each hospital that is based only on one factor.
- Optimized Multiple-Factor Ambulance Diversion Policy (MF AD). This strategy implements a diversion policy for each hospital that looks at several state variables to decide going on or off diversion.

Ambulance Destination Policies

- Take the patient to the nearest open hospital (NH).
- Take the patient to the least crowded hospital, which is the ED with the minimum number of patients waiting (LCH).

The first diversion policy analyzes the performance of the system if AD is prohibited. Some governments have banned ambulance diversion as a method to reduce congestion. However, this type of restriction might worsen the performance in EDs if other actions are not sufficient to relieve congestion (Massachusetts Nurses Association 2009).

The second type of diversion policy presents a policy based on a single factor or state variable to decide whether to go or not on diversion. Several reports and papers identify three main causes for going on diversion in practice: high number of patients waiting in the ED, high number of patients boarding in the ED and lack of beds in IP (American College of Emergency Physicians 2008; Centers for Disease Control and Prevention 2006b; McConnell 2005; Pham 2006). Therefore, the SF AD policy includes an upper threshold on one of these variables to decide if diversion status is set on, a lower threshold to remove the diversion status and a review frequency of the state of the system.

The third type of diversion policy includes several thresholds for each of the main state variables mentioned above. In addition, it includes thresholds for the number of patients waiting disaggregated per severity level. Thus, the MF AD policy triggers the diversion status when the state of the ED exceeds a specific number of thresholds.

Most analytical studies of ambulance diversion use single-factor in their analysis. For instance, Ramirez et al. (2010) propose six AD policies based on the three main causes for going on diversion. Deo and Gurvich (2011) analyze AD policies based only on the number of patients waiting in the ED; in addition, they

studied a diversion policy that sets the diversion status when all the beds in the ED are occupied. Allon et al. (2009) analyze AD policies based on a minimum and maximum of the number of patients boarding.

On the other hand, MF AD policies are common in real settings. For example, Hoot et al. (2008) presents a simulation study of an academic medical center that initiates AD if any of the following criteria is satisfied: 1) all critical care beds in the ED are occupied, patients are in hallway spaces and there are 10 or more patients waiting; 2) an acuity level exists that places additional patients at risk; or 3) all monitored beds within the ED are full.

Regarding the destination policies, guidelines of EMS suggest that patient should be taken to the nearest appropriate hospital (American College of Emergency Physicians 2006). Thus, ambulance crew and staff should make the decision based on distance and crowding levels. However, the decision might be suboptimal because of limited or unreliable information, bounded rationality or it can be based on a myopic perspective that does not weight the effect on the patients already in the system.

4.5 Chromosome Structure for AD Policies in a GA

This chapter presents multi-objective genetic algorithm combined with simulation a method to design and evaluate AD policies along with destination policies; therefore the structure of the chromosomes must be defined according to the types of policies described in Section 4.4.1. GA is applied only to the SF and MF policies to find the appropriate thresholds to go on and off diversion.

4.5.1 Chromosome for SF AD policies

The single-factor ambulance diversion policy for a specific hospital H_i is based on three main state variables:

NQ_i: Number of patients waiting in the ED of hospital H_i.

NB_i: Number of patients boarding in the ED of hospital H_i.

$NIPB_i$: Number of beds available in the IP unit of hospital H_i.

The total number of genes in a chromosome for this type of policy is $10n$, where n is the number of hospitals in the ECDS. Hence, the diversion policy for each hospital is represented by ten genes, which have the following structure:

Table 4.2. Chromosome partition that represents an SF AD policy in one hospital.

Gene	1	2	3	4	5	6	7	8	9	10
Variable	P_i	U_{NQi}	L_{NQi}	Δt_{NQi}	U_{NBi}	L_{NBi}	Δt_{NBi}	L_{NIPBi}	U_{NIPBi}	Δt_{NIPBi}

The first gene describes the type of factor to consider in the policy of hospital H_i. Thus, $P_i = 1$ implies that AD policy of hospital H_i is based only on number of patients waiting in the ED (NQ_i); $P_i = 2$ indicates that AD is based on the number of patients boarding (NB_i); and $P_i = 3$ means that AD is based on the number of beds available in the IP unit ($NIPB_i$). Therefore, the execution of an SF AD policy requires values for three parameters. If the policy is type 1, then the parameters are in the genes 2, 3 and 4. If it is type 2, then the genes of interest are 5, 6 and 7. If the type is 3, then the related genes are 8, 9 and 10.

The first of the three parameters that define a SF AD policy (gene 2, 5 or 8) is a threshold that triggers the diversion status. Thus, if policy is type 1, then the hospital H_i sets the diversion status when $NQ_i \geq U_{NQi}$. If it is type 2, then H_i

goes on diversion when $NB_i \geq U_{NBi}$. If it is type 3, then diversion is set when $NIPB_i \leq L_{NIPBi}$. After going on diversion, the state of the system is reviewed every Δt time units, represented by genes 4, 7 and 10 for policies type 1, 2 and 3, respectively.

The diversion status can be removed only at a review point and this decision depends on the current value of the state variable observed in the policy. Thus, if policy is 1, then the diversion status is removed if $NQ_i \leq L_{NQi}$. If policy is type 2, then diversion is removed if $NB_i \leq L_{NBi}$. If the policy is type 3, then diversion is removed if $NIPB_i \geq U_{NIPBi}$. Note that for all the policies the threshold U is greater than the threshold L. Policy type 3 has the U and L interchanged because of the meaning of the state variable (number of available beds in the IP). The three state variables analyzed in the SF policies are listed as the most common causes of going on diversion in practice. In addition, a study on single-factor AD policies presented by Ramirez et al. (2010) show that periodic review of the system after going on diversion produce smaller variability in the performance than continuous review. Therefore, the policies that include a periodic review are more consistent and allow a more precise prediction of the performance under a given policy.

An example of an SF AD policy is one that states: “go on diversion if there are at least 15 patients waiting in the ED, reevaluate the status every hour after going on diversion and remove the diversion status if there are 5 or less patients waiting”. This policy is encoded as shown in Table 4.3.

Table 4.3. Chromosome example for an SF AD policy.

Gene	1	2	3	4	5	6	7	8	9	10
Variable	1	15	5	60	Null	Null	Null	Null	Null	Null

In this example, genes 5 to 10 can take any value and the simulation code does not take them into account because the first gene specifies the type of policy.

4.5.2 Chromosome for MF AD policies

The MF AD policy considers more than one state variable to trigger the diversion status. Besides the variables used in SF policies, the MF includes the number of patients waiting in the ED of hospital H_i per severity level ($NQ1_i$, $NQ2_i$, $NQ3_i$, $NQ4_i$ and $NQ5_i$). The total number of genes for MF policies is also $10n$. Hence, for each hospital, the structure of the chromosome is:

Table 4.4. Chromosome partition that represents an MF AD policy in one hospital.

Gene	1	2	3	4	5	6	7	8	9	10
Variable	U_{NQi}	U_{NQ1i}	U_{NQ2i}	U_{NQ3i}	U_{NQ4i}	U_{NQ5i}	U_{NBi}	L_{NIPBi}	k	Δt

This type of policy comprises multiple thresholds presented in genes 1 to 8. Gene 9 is a variable k that represents the number of thresholds that must be reached in order to decide going on diversion. Thus, if $k = 1$, then the hospital will go on diversion when any threshold is reached, if $k = 8$, then hospital sets the diversion status only when all the thresholds are reached. On the other hand, it is possible that a threshold takes a null value, which means that that factor is not considered in the policy. After going on diversion, the status is reevaluated every Δt time units. At a review point, the diversion status is removed if the number of state variable above the thresholds (or below in case of $NIPB_i$) is less than k.

An example of an MF AD policy that states: “Hospital goes on diversion if at least two of the following conditions are satisfied:

1. The number of patients waiting in the ED is at least 20;
2. the number of patients with severity level 1 waiting in the ED is at least 2; and
3. the number of patients boarding is at least 3,

after going on diversion, the system will be reevaluated every 30 minutes”, can be expressed as shown in Table 4.5.

Table 4.5. Chromosome example for an MF AD policy.

Gene	1	2	3	4	5	6	7	8	9	10
Variable	20	2	Null	Null	Null	Null	3	Null	2	30

The details about the recombination and mutations strategies for SF and MF AD policies are presented in Appendix B.

4.6 Experimentation

The experimentation process for the centralized design of AD policies consists of two case studies; one of them comprises an ECDS with two hospitals and another with three hospitals.

4.6.1 Case study 1: ECDS with two hospitals

The first case to analyze is an ECDS with two hospitals located in a 100 squared-miles area, assuming the region is a square with corners in (0,0), (10,0), (0,10) and (10,10). Three location settings are defined for this part, one symmetric and two random locations. In the case of the symmetric location, the hospitals H_1 and H_2 are found at coordinates (2.5, 7.5) and (7.5, 2.5), respectively. The first random location places the hospitals at coordinates (5.96, 8.99) and (3.82, 1),

respectively. And the second random location places hospitals at (0.75, 6.23) and (5.81 and 6.89), respectively. These locations are depicted in Figure 4.4.

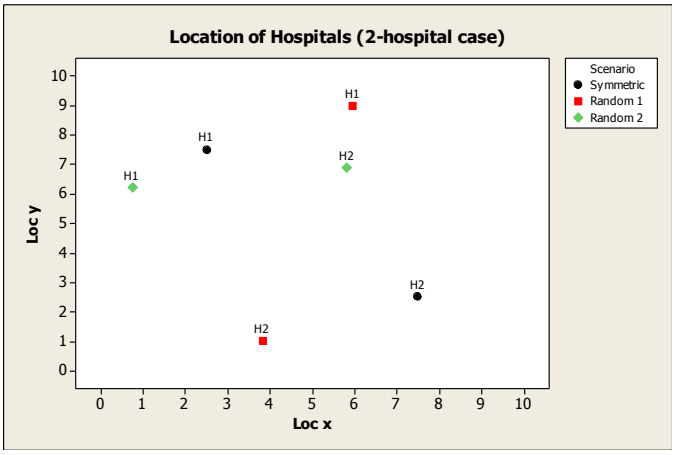

Figure 4.4. Location of hospitals for case study 1.

The simulation models of the hospitals follow the same structure and input parameters described in Section 4.3 and Appendix B. However, two relative sizes are tested: 1:1 and 1:1.2. Thus, 1:1 implies that the arrival rates for both hospitals are the same, while 1:1.2 implies that arrival rate to H_2 is 20% larger than the arrival rates to H_1. These two scenarios are used to expose the system to hospitals with similar characteristics and a system where one of the hospitals is more congested than the other. The scenarios and strategies used in the experimentation of this case study are summarized in Table 4.6.

Table 4.6. Scenarios and strategies used in the experimentation process.

Scenarios		Strategies	
Location (H_1, H_2)	Relative Size (H_1:H_2)	Diversion Policies	Destination Policies
Symmetric: (2.5, 7.5) & (7.5, 2.5) Random1: (5.96, 8.99) & (3.82, 1) Random2: (0.75, 6.23) & (5.81, 6.89)	1:1 1:1.2	No AD Optimized SF AD Optimized MF AD	NH LCH

The strategies combine diversion and destination policies to each potential scenario. Figure 4.5 shows the average-patient non-value added time (NVT) for each hospital in 2-dimension graphs. Each scenario is shown in a separate graph. For each scenario, all the strategies are presented.

The strategies that consider No AD are presented as a single point. In addition, a selection of non-dominated strategies obtained after the last generation of the GA for SF and MF diversion policies are presented in the results.

These results suggest that No AD combined with NH has an undesirable effect on the performance of the ECDS. One disadvantage of this strategy is that it unbalances the workload of emergency patients between the EDs if one hospital is more congested than the other, or if one hospital is located in a central area and another near the perimeter. If both hospitals are similar and location allows receiving a similar fraction of emergency patients, then a balanced performance is achieved, but there are other strategies that outperform this one.

The No AD - LCH strategy seems to have an acceptable performance compared with all the other strategies. In fact, this strategy balances the performance between the hospitals in the system, regardless their characteristics

or location. However, in all the scenarios there exists a strategy based on AD that dominates it.

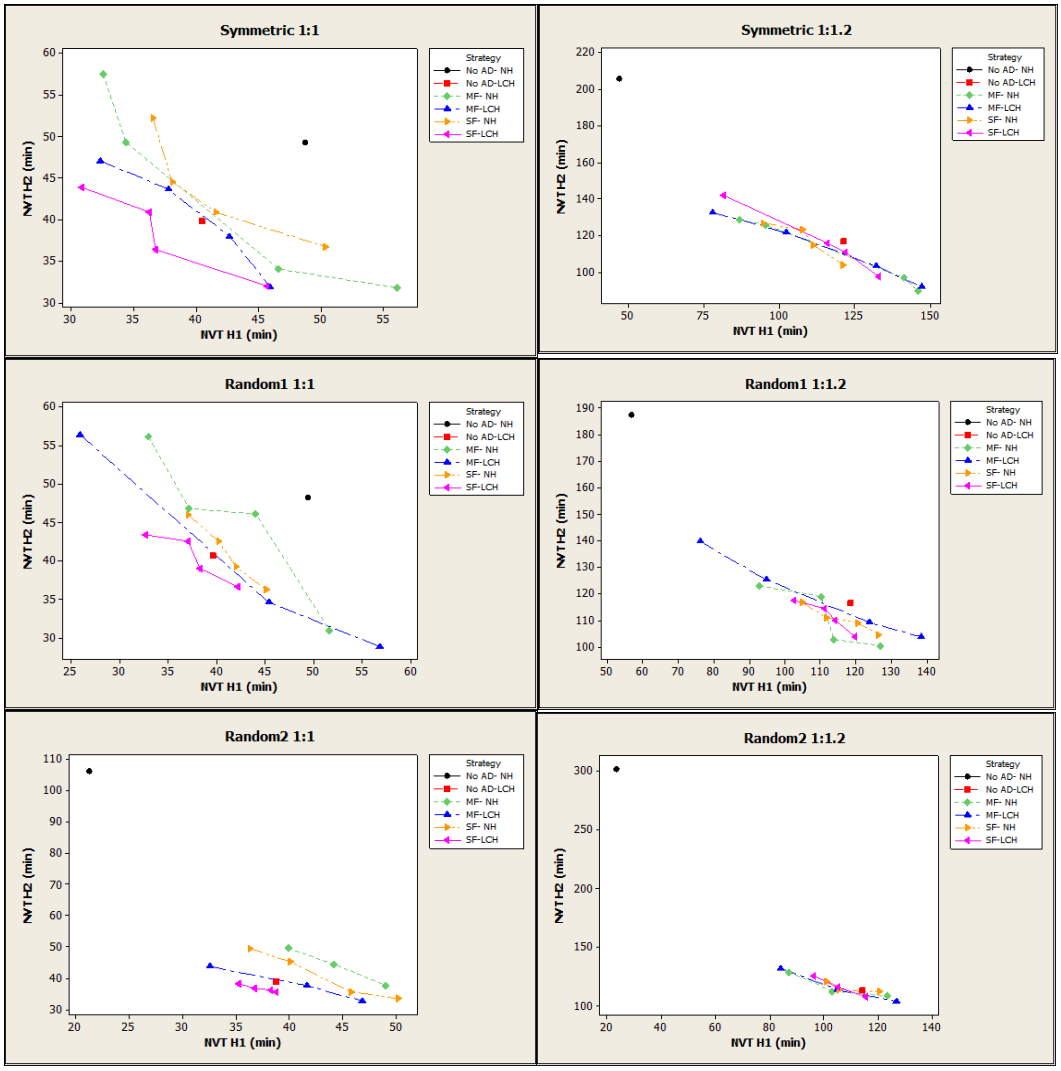

Figure 4.5. Results of GA and simulation for a 2-hospital ECDS.

There are important observations regarding the effect of AD policies. First, in the case of similar hospitals, the set of SF policies combined with LCH has a frontier better located than all the other strategies. Consequently, the overall set of nondominated strategies contains all the policies of this type. Furthermore, the Pareto fronts of the AD strategies combined with the LCH dominate most of

the strategies that combine AD with NH. In the case of a difference in the relative size, the lines that join the nondominated solutions for each strategy overlap each other. In general, strategies that allow AD have better performance than not allowing AD. Moreover, simple diversion policies like SF are at least as good as a more complex MF policy.

Table 4.7 shows the results for the strategies that do not use AD and strategies that allow diversion and have a balanced performance between hospitals. The No AD – NH strategy produces poor results if compared with other strategies. The No AD – LCH strategy produces results that balance the performance measure between both hospitals. However, in five out of the six scenarios, there is a diversion-based strategy that dominates this one.

Table 4.7. Comparison of No AD strategies vs. AD strategies.

	No AD Strategies				AD Strategies				
	Average patient NVT (mins) No AD - NH		Average patient NVT (mins) No AD - LCH			Average patient NVT (mins)		Percentage of time on diversion	
Scenario	H_1	H_2	H_1	H_2	Strategy	H_1	H_2	H_1	H_2
Symmetric 1:1	48.7	49.3	40.5	39.9	SF-LCH[a]	36.8	36.4	9.4	10.1
Symmetric 1:1.2	47.2	205.8	121.6	116.8	SF-NH	111.5	115.0	3.4	30.6
Random1 1:1	49.4	48.2	39.6	40.7	SF-LCH[a]	38.3	39.0	9.4	5.0
Random1 1:1.2	56.9	187.3	118.6	116.4	SF-NH	111.9	111.1	12.2	37.5
Random2 1:1	21.3	106.2	38.8	38.8	SF-LCH	36.7	36.8	9.0	9.9
Random2 1:1.2	23.5	301.3	114.4	112.9	SF-LCH	115.5	107.7	1.1	22.9

[a] Dominates No AD strategies

The diversion policies shown in the table are Pareto strategies and they balance performance measure between hospitals. In most of the scenarios, the diversion policy is accompanied by LCH destination. In order to balance the average-patient non-value added time across the hospitals, the percentage of time

spent on diversion can be significantly different between hospitals, especially if the hospitals have different relative size. For instance, in Scenario 1 (Symmetric 1:1), the percentages of time on diversion differ for only 0.7%; while in scenario 2 (Symmetric 1:1.2) the difference is 27.2%.

4.6.2 Case study 2: ECDS with three hospitals

The second case study consists of analyzing the performance of AD policies in a system with three hospitals. Two configurations of random locations are presented; one of them assumes that the ECDS is in a 10x10 squared-miles area (Random1), while another assumes that the area is 20x20 squared-miles (Random2). Figure 4.6 shows the location of each hospital for both configurations. Besides, two settings for the relative size of the hospitals are tested, one assumes the same relative size (1:1:1) and another assumes different sizes for all the hospitals, one of them has 10% more arrivals than the generic and another 20% more arrivals than the generic (1:1.1:1.2).

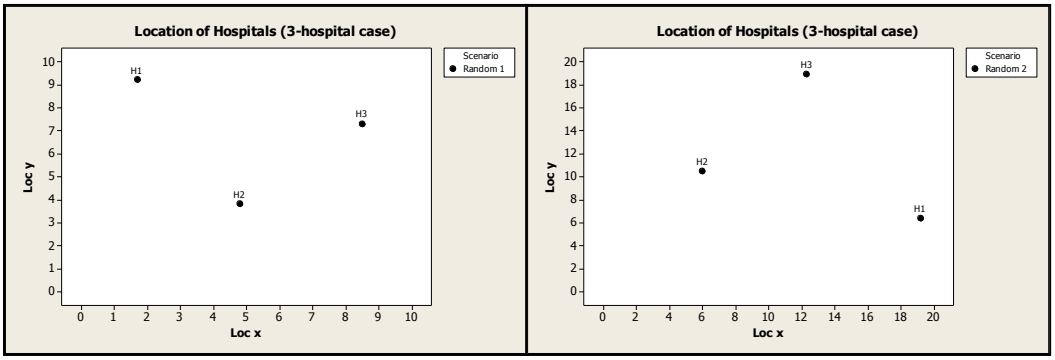

Figure 4.6. Location of the hospitals for case study 2.

Three strategies of diversion policies are used in the second case study. No AD policies, centralized SF policies and a simple policy (Simple AD). The first two of them are described in Section 4.4.1. The centralized MF policies are discarded in this part because they are more complex than SF policies and they do not guarantee to be significantly better than SF. The Simple AD policy consists of setting the diversion status when all the beds in the ED are occupied. This is a policy proposed by Deo and Gurvich (2011). In that paper, the authors propose this policy as a Pareto improving AD policy over the No AD policy, considering the waiting time in the ED as the performance metric of the system. Even though this chapter has significant differences in the structure of the model and in the evaluation of the performance, it is of the interest of this research to benchmark the proposed centralized SF policy to the Simple AD.

The destination policies used in this part include again NH, LCH and a new policy that adds up the expected transportation time to a hospital plus the current waiting time in the hospital (SUM). Thus, a new ambulance patient will be taken to hospital H_i such that:

$$H_i = \arg\min_i \left\{ E(T_i) + \sum_{k=1}^{5} w_{k,i} \overline{W}_{k,i} \right\} \tag{4.2}$$

where,

i: indices of open hospitals (hospitals off diversion).

$E(T_i)$: expected transportation time. Thus, $E(T_i) = \tau \times M(l, H_i)$, where τ is the average transportation time per mile and $M(l, H_i)$ is the Manhattan distance between emergency location l and hospital Hi.

$\sum_{k=1}^{5} w_{k,i}\overline{W}_{k,i}$: weighted average waiting time of the current number of patients waiting in the ED of hospital H_i.

Table 4.8. Scenarios and strategies used in the second experimentation process.

Scenarios		Strategies	
Location (H_1, H_2, H_3)	Relative Size (H_1: H_2: H_3)	Diversion Policies	Destination Policies
Random 1: (1.7, 9.2), (4.8, 3.8) & (8.5, 7.3) Random 2: (19.2, 6.4), (6, 10.5) & (12.3, 18.9)	1:1:1 1:1.1:1.2	No AD Simple AD Optimized SF AD	NH LCH SUM

The results for case study 2 are shown in Table 4.9, which includes the average-patient non-value added time per hospital for each strategy. Besides, it shows the sum of the non-value added time in the system, the standard deviation and the percentage of time spent on diversion in each hospital.

The results show that the proposed centralized design of SF AD policies significantly reduce the total average-patient non-value added time in the system compared with No AD. Furthermore, the Simple AD strategies outperform No AD; however; the centralized design of the SF AD achieves the best results. In every scenario, there is at least an AD strategy that dominates the No AD counterpart. In addition, SF AD strategies tend to balance the performance among hospitals, reducing the standard deviation of average-patient non-value added times across facilities.

Table 4.9. Results of diversion and destination strategies for an ECDS with three hospitals.

Scenario	Destination Policy	Diversion Policy	Average patient NVT (minutes) H_1	H_2	H_3	Sum NVT	Std. Dev. NVT
Random1 1:1:1	NH	No AD	21.28	155.63	45.29	222.21	71.65
		Simple AD	25.55	86.00	44.03	155.59	30.98
		Optimized SF AD	67.26	52.54	27.58	147.38	20.06
		Optimized SF AD[a]	20.47	101.09	36.96	158.52	42.59
	LCH	No AD	54.86	41.47	36.74	133.08	9.40
		Simple AD	37.12	43.32	40.73	121.17	3.12
		Optimized SF AD	36.48	42.78	37.67	116.93	3.35
		Optimized SF AD[a]	53.56	36.56	29.76	119.88	12.26
	SUM	No AD	37.80	49.34	42.95	130.09	5.78
		Simple AD[a]	36.48	46.91	41.49	124.89	5.21
		Optimized SF AD	25.48	51.08	43.66	120.22	13.17
		Optimized SF AD[a]	33.94	46.95	41.22	122.11	6.52
Random1 1:1.1:1.2	NH	No AD	25.24	273.94	160.93	460.11	124.52
		Simple AD	46.87	197.28	144.58	388.73	76.32
		Optimized SF AD	71.34	155.17	148.48	374.99	46.59
	LCH	No AD	132.93	123.55	119.53	376.01	6.88
		Simple AD	125.77	126.18	123.65	375.61	1.36
		Optimized SF AD[c]	117.39	106.46	111.13	334.98	5.48
	SUM	No AD	116.22	128.92	130.33	375.48	7.77
		Simple AD	116.68	126.45	131.27	374.40	7.43
		Optimized SF AD	140.70	54.49	147.88	343.08	51.97
		Optimized SF AD[c]	115.92	125.81	114.70	356.43	6.10
Random2 1:1:1	NH	No AD	34.12	144.28	26.26	204.66	65.99
		Simple AD	38.97	81.91	32.92	153.79	26.71
		Optimized SF AD	43.66	75.64	33.92	153.21	21.82
		Optimized SF AD[a]	26.37	104.73	25.56	156.66	45.48
	LCH	No AD	58.40	43.88	37.62	139.90	10.66
		Simple AD	42.06	46.15	41.27	129.48	2.62
		Optimized SF AD[b]	34.34	36.89	40.07	111.30	2.87
	SUM	No AD	41.81	54.77	40.90	137.48	7.76
		Simple AD[a]	38.71	49.65	38.45	126.81	6.40
		Optimized SF AD[c]	37.44	44.09	34.25	115.79	5.02
		Optimized SF AD[a]	35.46	46.94	39.80	122.20	5.80
Random2 1:1.1:1.2	NH	No AD	50.57	317.08	131.81	499.47	136.60
		Simple AD	83.50	238.65	137.97	460.12	78.71
		Optimized SF AD[b]	83.18	226.75	123.35	433.28	74.07
		Optimized SF AD	155.39	189.56	103.21	448.17	43.49
	LCH	No AD	168.27	156.33	150.16	474.76	9.21
		Simple AD[a]	159.98	153.01	149.22	462.20	5.46
		Optimized SF AD	76.71	61.11	290.24	428.05	128.02
		Optimized SF AD[a]	151.06	131.86	149.27	432.19	10.61
		Optimized SF AD[c]	153.12	144.20	141.33	438.66	6.15
	SUM	No AD	147.78	159.68	159.01	466.47	6.68
		Simple AD	152.43	163.26	162.50	478.19	6.04
		Optimized SF AD[c]	138.33	146.70	149.34	434.37	5.75

[a] Dominates No AD strategy

[b] Dominates Simple AD strategy

[c] Dominates No AD and Simple AD strategies

Regarding the destination policies, LCH and SUM policies clearly outperform the NH policy and they also produce balanced performance across hospitals. On the other hand, the difference in performance between LCH and SUM seem to be insignificant; there is not a clear domination of one policy over the other. The cause of the similar performance might be because in the SUM policy, the component related to current waiting time, which is highly correlated with LCH, is a determinant factor to decide where to take a patient. Table 4.10 shows the percentage of improvement on the total average-patient non-value added time in the system when SF AD and LCH are used.

Table 4.10. Percentage of improvement on the total average-patient non-value added time of optimized SF AD policies and LCH over other policies.

	Improvement of SF AD (%) compared with:			Improvement of LCH (%) compared with:		
Scenario	Destination Policy	No AD	Simple AD	Diversion Policy	NH	SUM
Random1 1:1:1	NH	33.68	5.28	No AD	40.11	2.24
	LCH	12.13	3.50	Simple AD	22.12	-3.07
	SUM	7.59	3.74	SF AD	20.66	-2.81
Random1 1:1.1:1.2	NH	18.50	3.53	No AD	18.28	0.14
	LCH	10.91	10.82	Simple AD	13.83	0.32
	SUM	8.63	8.37	SF AD	12.12	-2.42
Random2 1:1:1	NH	25.14	0.38	No AD	31.64	1.73
	LCH	20.44	14.0	Simple AD	15.81	2.06
	SUM	15.78	8.69	SF AD	27.36	-4.03
Random2 1:1.1:1.2	NH	13.25	5.83	No AD	4.95	1.75
	LCH	9.84	7.39	Simple AD	-0.45	-3.46
	SUM	6.88	9.16	SF AD	1.21	-1.48

The improvement percentage of using SF AD policies reduces when the destination policy considered is based on current crowding conditions; however, the reduced proportion of inappropriate level of treatment is still significant.

The results of both case studies strongly suggest that a centralized design of AD policies can smooth the patient flow by reducing the delays in activities with inappropriate level of care through different stages. Furthermore, the destination policy is a significant factor in the performance of the ECDS. Even though AD can be seen as a negative aspect in emergency care, this chapter showed that a centralized design of these policies is an effective patient allocation mechanism that can help avoiding congestion upstream (waiting) and downstream (boarding) in the ECDS. In addition, these results support the observations seen in systems where AD is prohibited, whose hospitals face straining in the operations that raises the waiting time and the number of patients boarding.

4.7 Limitations

The conclusions drawn in this chapter have certain limitations that must be highlighted. First, the periods on AD considered in this research are full period. However, some hospitals might divert only patients with specific severity level (e.g., trauma centers must receive all level 1patients).Nevertheless, the number of patients with the highest level of trauma is small (Centers for Disease Control and Prevention 2008); therefore, the advantages of AD observed in this chapter might still hold.

On the other hand, the results of this chapter show that a destination policy based on LCH has better results than NH; however, this might be true only in urban areas. The decision regarding the number of hospitals located in 100 squared-miles or 400 squared-miles was based on information obtain from Google Maps in an area of similar proportions in Maricopa County in AZ. But, these

results might not hold if the analysis is applied to rural areas. Nevertheless, the significant increase in transportation time in rural areas may jeopardize the health of the patients; therefore, AD is often not recommended in those regions.

The case studies prepared in this chapter include models of fictitious hospitals. The data available from real hospitals that might allow building a model for this research is insufficient. Moreover, different jurisdictions own the data needed to build a model like the proposed. Hence, the ECDS analyzed in this chapter was designed with hospitals whose characteristics are realistic. Appendix B describes the source of the input data. Furthermore, the models show statistics (i.e. average waiting time and LWOT percentage) that are validated through information published in different sources across the United States. Although the experimentation process is in an inexistent system, the methodology proposed for a centralized design of AD policies is independent of this scenario.

Nonetheless, the potential limitation on the applicability of this methodology is the level of cooperation among hospitals. Thus, the effectiveness of the centralized design of AD policies consists of having accurate and sufficient information to properly apply the diversion and destination policies. Therefore, cooperation mechanisms must be assured and empirical studies presented in the literature show that organizations are willing to cooperate in order to improve the healthcare system.

An important element of the hypothesis of this chapter is that smart diversion policies can provide benefits for the entire healthcare system. Articles and papers from medical sources tend to contradict this idea. However, this

chapter and other references show that not going on diversion might worsen the congestion in EDs (Massachusetts Nurses Association 2009). Furthermore, it is thought that diverting patients have negative implications on the economic aspect of the hospital that is on diversion because of the opportunity cost. Nevertheless, AD can reduce the adverse events caused by saturated systems, which may imply cost savings in the long term.

4.8 Conclusions

This chapter presented a centralized design of AD policies using GA and simulation to evaluate the performance. The AD policies are combined with destination policies in an ambulance flow control framework that allows the allocation of ambulance patient in an ECDS. The findings suggest that the centralized design of diversion policies and effective destination rules can reduce the time that patients spend in inappropriate level of care, including the patients that walk-in into an ED. This implies smoothing the patient flow using the appropriate diversion-destination strategy.

Two types of diversion policies were explored in this chapter: single-factor (SF) and multiple-factors (MF) policies. Even though SF is simpler than MF, the results show that they are at least as good as MF. In addition, the centralized design of any of these policies outperform other policies seen in real settings, such as No AD and setting the AD status when all the ED beds are occupied.

On the other hand, the least-crowded-hospital (LCH) destination policy outperforms the nearest-hospital (NH). Furthermore, a policy based on the sum of

expected transportation time and current average waiting time performs better than NH and similar to LCH. However, this might hold only on urban settings, like the one assumed in this research.

These results show the potential of reducing inappropriate level of care and avoiding adverse events by designing smart policies related to ambulance flow. Nevertheless, there are important challenges related to cooperation that must be overcome to obtain benefits from AD in real ECDS.

Future extensions of this research include the optimization of destination policies and combine them with the optimized diversion policies explored in this chapter. Approximate dynamic programming could be explored to optimize destination policies.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Overcrowding of EDs in several places around the work has required proposing solutions and making decisions to ensure that patients receive appropriate level of treatment timely. Diverting ambulances from overcrowded EDs started as a solution to potential periods of congestion, but the use of ineffective AD policies has caused concern on society and medical community. In fact, emergency physicians recommend avoiding diversion because of the potential harmful effect of longer transportation.

Nevertheless, overcrowding in EDs is still present in many regions and the safety of patients is at risk due to long waiting times and adverse events that occur in congested facilities. This dissertation presents modeling techniques to design and analyze AD policies, considering different measures related to safety, such as average waiting time, average tardiness and average patient non-value added time. The hypothesis underlying the dissertation is that an effective design of AD policies can improve the safety conditions of the patients in periods of high congestion. Chapters 2, 3 and 4 of this dissertation contain valuable information to assess decision makers in the design of their AD policies.

Chapter 2 presents a methodology to analyze the effectiveness of single-threshold AD policies in terms of mean performance and variability. The analysis was performed using bicriteria approach, which includes the average patient waiting time in the ED and the percentage of time spent on diversion. Given that patients arriving by ambulance have higher severity level, they tend to receive

priority to be treated and they spent longer time in treatment. Therefore, AD allows relieving congestion in the ED; however, the accessibility to emergency care may be compromised. The single-threshold AD policies evaluated in this chapter are based on one of the three main contributors to diversion in practice: number of patients waiting in the ED, number of patients boarding and inpatient occupancy level. In addition, the policies allow a periodic or continuous review of the system to go off diversion. The methodologies and analysis based on simulation include graphical and quantitative methods to evaluate mean performance and variability.

The results for Chapter 2 show that diverting ambulances from EDs can reduce significantly the average waiting time; however, the policies based on patients waiting and patients boarding reduce the average waiting time in a larger rate than the policies based on inpatient occupancy level. In addition, the policies that have a periodic review are more consistent; therefore, they allow a more accurate prediction of the performance of the ED for a given policy. Regression equations are proposed to derive the parameters of single-threshold AD policies that yields the desired results. These results contribute to the discussion of AD policies by proposing a simple methodology that allows the design and analysis of single-threshold policies and discusses the implications of using different state variables.

The structure of optimal AD control policies is explored in Chapter 3. A Markov Decision Process model is proposed with the objective of minimizing the average expected tardiness per patient. The measure of tardiness proposed in this

chapter differs from the manufacturing setting in that it represents the time that patients have to wait beyond a safety time threshold to start emergency treatment. Therefore, a measure correlated with safety is included. In addition, the model assumes that the distribution for the diverted patients to start treatment in a neighboring hospital is known. This aspect was not included in the previous chapter. The analysis of the MDP consists of theoretical and computational studies that suggest the following: the optimal AD policy has a threshold type, which is non-increasing in the number of patients in the ED and also in the time to start treatment in another hospital. The model indicates that diversion can help to manage the traffic with more effectiveness if the area where critical patients are treated is more congested than an area with a fast-track assessment.

Even though the MDP model includes several relaxations that are not realistic, a simulation study that includes time-dependent patterns observed in real settings confirm that a policy prescribed by the MDP performs significantly better than most heuristics used in practice. This chapter contributes to literature on being the first paper that explores the structural properties of optimal AD policies using MDP and it proposes a performance measure that is aligned with the objective of emergency care systems.

Since MDP model suffer from scalability issues, Chapter 4 proposes a genetic algorithm combined with simulation to design effective AD policies for multiple hospitals simultaneously. The simulation model includes different modules that allow the simulation of multiple hospitals and the generation of emergency events in a geographical region. The proposed model has the objective

of finding AD policies that yields Pareto improving solutions that minimizes the average patient non-value added time for each hospital. The average patient non-value added time comprises transportation, waiting and boarding time for patients requiring emergency care. Two types of chromosome structure are tested; one considers a single state variable to go on and off diversion; and the other uses information about several state variables. In addition, the AD policies are combined with simple ambulance destination policies to determine where an ambulance patient should be taken to.

The results shown in Chapter 4 suggest that policies that are based on a single-factor are simpler to implement and they could have similar or even better performance than multiple-factor AD policies. In addition, effective AD policies are Pareto improving and can reduce significantly the total average patient non-value added time in an emergency care delivery system. In addition, the destination policy has a significant effect on the performance of the system. Smart destination policies that balance the distance and crowding factors can boost the effectiveness of the AD policies. This chapter contributes to literature by proposing a genetic algorithm model that allows a centralized design of AD policies and it presents the combination of ambulance destination and diversion policies as an ambulance flow control mechanism that assesses the ambulance patient allocation in an emergency care delivery system.

Aware of the concerns of the medical community for the risks implied in diverting patients from EDs, this research demonstrates that an effective design of AD policies can improve performance measures related to patient safety.

Nevertheless, finding solutions to the root problems related to congestion and alternatives that have effect on the long term are encouraged.

Even though the models presented in this dissertation includes the most important aspects in complexity, relations and resources found in EDs; there are many more elements that decision makers would like to explore. Some of these models are flexible enough to include alternative elements to the analysis.

Future research in this topic includes exploring the optimal design of ambulance destination and diversion policies simultaneously. For this purpose, methods such as Approximate Dynamic Programming can be suitable to analyze the problem.

REFERENCES

Agency for Healthcare Research and Quality. 2005. Emergency severity index, version 4: Implementation handbook. Accessed on May 7, 2011. http://www.ahrq.gov/research/esi/esihandbk.pdf

Allon, G., S. Deo, W. Lin. 2011. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. Working Paper. Accessed on July 29, 2011. http://www.gsb.stanford.edu/facseminars/events/oit/documents/oit_10_09_deo.pdf.

American College of Emergency Physicians. 1999. Guidelines for ambulance diversion. Accessed on May 7, 2011. http://www.acep.org/content.aspx?id=30038.

American College of Emergency Physicians. 2006. Emergency ambulance destination. Accessed on May 7, 2011. http://www.acep.org/content.aspx?id=29196.

American College of Emergency Physicians. 2007. Model of the clinical practice of emergency medicine. Accessed on May 7, 2011. http://www.acep.org/content.aspx?id=29164.

American College of Emergency Physicians. 2008. Emergency department crowding: High-impact solutions. Accessed on May 7, 2011. www.acep.org/WorkArea/DownloadAsset.aspx?id=50026.

Arizona Emergency Medical Systems, Inc. 2000. Chapter 8: Transfer of care. Guideline statements on prehospital diversion. In *Regional EMS Directory (Red) Book*. Accessed on May 7, 2011. http://www.aems.org/aems/redbook/chap8.pdf.

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

Asplin, B.R. 2003. Editorial: Does ambulance diversion matter? *Annals of Emergency Medicine* 41(4): 477-480.

Associated Press. 2006. Report: ER care in U.S. at 'breaking point'. Updated June 14, 2006. http://www.msnbc.msn.com/id/13320317/ns/health-health_care/t/report-er-care-us-breaking-point/

Banks, J., J.S. Carson II, B.L. Nelson, D.M. Nicol. 2010. *Discrete-event system simulation*. Upper Saddle River, NJ: Pearson Education, Inc.

Bertsekas, D. P. 2001. *Dynamic programming and optimal control*, vol. two. Belmont, MA: Athena Scientific.

Bozkurt, B., J.W. Fowler, E.S. Gel, B. Kim, M. Koksalan, J. Wallenius. 2010. Quantitative comparison of approximate solution sets for multicriteria optimization problems with weighted Tchebycheff preference function. *Operations Research* 58(3): 650-659.

Broyles, J.R., J.K. Cochran. 2007. Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression. In *Proceedings of the 2007 Industrial Engineering Research Conference*: 613-618.

Burt, C.W., L.F. McCaig, R.H. Valverde. 2006. Analysis of ambulance transports and diversions among US emergency departments. *Annals of Emergency Medicine* 47(4): 317-326.

Carlyle, W.M., J.W. Fowler, E.S. Gel, B. Kim. 2003. Quantitative comparison of approximate solution sets for bi-criteria optimization problems. *Decision Sciences* 34(1): 63-82.

Carr, S., I. Duenyas. 2000. Optimal admission control and sequencing in a make-to-stock/make-to-order production system. *Operations Research* 48(5): 709-720.

Carter, M.W., S.D. Lapierre. 2001. Scheduling emergency room physicians. *Health Care Management Science* 4: 347-360.

Ceglowski, R., L. Churilov, J. Wasserthiel. 2007. Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society* 58: 246-254.

Centers for Disease Control and Prevention. 2006a. National hospital ambulatory medical care survey: 2004 emergency departments summary. Advance Data from Vital and Health Statistics No. 372.

Centers for Disease Control and Prevention. 2006b. Staffing, capacity, and ambulance diversion in emergency departments: United States, 2003-04. Advance Data from Vital and Health Statistics No. 376.

Centers for Disease Control and Prevention. 2008. National hospital ambulatory medical care survey: 2006 emergency departments summary. National Health Statistics Reports No. 7.

Centers for Disease Control and Prevention. 2010. National hospital ambulatory medical care survey: 2007 emergency departments summary. National Health Statistics Report No. 26.

Chen, J.A., Y.H. Chen, M. Parlar, Y.B. Xiao. 2011. Optimal inventory and admission policies for drop-shipping retailers serving in-store and online customers. *IEEE Transactions* 43(5): 332-347.

Chockalingam, A., K. Jayakumar, M.A. Lawley. 2010. A stochastic control approach to avoiding emergency department overcrowding. In *Proceedings of the 2010 Winter Simulation Conference*, ed. B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.: 2399-2411.

CNN U.S. 2008. Tape shows woman dying on waiting room floor. Updated July 01, 2008. http://articles.cnn.com/2008-07-01/us/waiting.room.death_1_hospital-staff-hospital-employee-kings-county-hospital-center?_s=PM:US

Cochran, J.K., A. Bharti. 2006. A multi-stage stochastic methodology for whole hospital bed planning under peak loading. *International Journal of Industrial and Systems Engineering* 1: 8-36.

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

Deb, K., A. Pratap, S. Agrawal, T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6(2): 182-197.

Deo, S., I. Gurvich. 2011. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science* 57(7): 1300-1319.

Elmaghraby, W., P. Keskinocak. 2003. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science* 49(10): 1287-1309.

Epstein, S.K., L. Tian. 2006. Development of an emergency department work score to predict ambulance diversion. *Academic Emergency Medicine Journal* 13 (4): 421-426.

Fatovich, D.M., R.L. Hirsch. 2003. Entry overload, emergency department overcrowding and ambulance bypass. *Emergency Medicine Journal* 20: 406-409.

Feng, Y.Y., Z. Pang. 2010. Dynamic coordination of production planning and sales admission control in the presence of a spot market. *Naval Research Logistics* 57(4): 309-329.

Green, L.V. 2006. Queuing analysis in healthcare. In *Patient flow: reducing delay in healthcare delivery*, ed. R.W. Hall, Springer's International Series: 281-307.

Green, L.V. 2008. Study confirms increased heart attacks deaths in NYC ambulance diversions. Updated November 10, 2008. http://www.eurekalert.org/pub_releases/2008-11/ifor-sci111008.php

Green, L.V., J. Soares, J.F. Giglio, R.A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1): 61-68.

Gupta, D., L. Wang. 2007. Capacity management for contract manufacturing. *Operations Research* 55(2): 367-377.

Ha, A.Y. 1997. Inventory rationing in make-to-stock production system with several demand classes and lost sales. *Management Science* 43(8): 1093-1103.

Hafizoglu, A.B., E.S. Gel, P. Keskinocak. 2011. Expected tardiness computations in multi class priority M/M/c queues. Working Paper.

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

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

Kelton, W.D., R.P. Sadowski, D.T. Sturrock. 2007. *Simulation with Arena*. New York, NY: The McGraw Hill Companies, Inc.

Keskinocak, P., S. Tayur. 2004. *Due date management policies*. Kluwer Academic Publisher, 485-554.

Kolker, A. 2008. Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion. *Journal of Medical Systems* 32: 389-401.

KVAL. 2010. Parents of woman who died after waiting in ER sue hospital. Updated July 13, 2010. http://www.kval.com/news/local/98377384.html.

Lagoe, R.J., M.S. Jastremski. 1990. Relieving overcrowded emergency departments through ambulance diversion. *Hospital Topics* 68(3): 23.

Law, A.M. 2007. *Simulation modeling & analysis*.New York, NY: The McGraw Hill Companies, Inc.

Massachusetts Nurses Association. 2009. State's "no diversion policy" is putting strain on Massachusetts hospitals. *The Massachusetts Nurse* 80(4): 8-9.

Mayhew, L., D. Smith. 2008. Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments. *Health Care Management Science* 11: 11-21.

McConnell, K.J., C.F. Richards, M. Daya, S.L. Bernell, C.C. Weathers, 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.

Medeiros, D.J., E. Swenson, C. DeFlitch. 2008. Improving patient flow in a hospital emergency department. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S.J. Mason, R.R. Hill, L. Monch, O. Rose, T. Jefferson, and J.W. Fowler, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.: 1526-1531.

Meng, L.Y., T. Spedding. 2008. Modeling patient arrivals when simulating an accident and emergency unit. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S.J. Mason, R.R. Hill, L. Monch, O. Rose, T. Jefferson, and J.W. Fowler, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.: 1509-1515.

Miller, M., D. Ferrin, N. Shahi. 2009. Estimating patient surge impact in several regional emergency departments. In *Proceedings of the 2009 Winter Simulation Conference*, ed. M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin, and R.G. Ingalls, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.: 1906-1915.

Montgomery, D.C. 2005. *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons, Inc.

National Academy of Engineering and Institute of Medicine. 2005. *Building a better delivery system. A new engineering/health care partnership*. Washington, DC: The National Academy Press.

Patel, P.B., R.W. Derlet, D.R. Vinson, M. Williams, J. Wills. 2006. Ambulance diversion reduction: The Sacramento solution. *American Journal of Emergency Medicine* 24: 206-213.

Petzall, K., J. Petzall, J. Jansson, G. Nordstrom. 2011. Time saved with high speed driving of ambulances. *Accident Analysis and Prevention* 43: 818-822.

Pham, J.C., R. Patel, M.G. Millin, T.D. Kirsch, A. Chanmugam. 2006. The effects of ambulance diversion: A comprehensive review. *Journal of the Academic Emergency Medicine* 13(11): 1220-1227.

Press Ganey Report. 2009. Emergency department pulse report 2009 – patient perspectives on American health care.

Ramirez, A., J.W. Fowler, T. Wu. 2009a. Modeling of regional healthcare delivery networks using distributed simulation. In *Proceedings of the 2009 Industrial Engineering Research Conference*: 681-686.

Ramirez, A., J.W. Fowler, T. Wu. 2009b. Analysis of ambulance diversion policies for a large-size hospital. In *Proceedings of the 2009 Winter Simulation Conference*, ed. M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin, and R.G. Ingalls, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.: 1875-1886.

Ramirez, A., J.W. Fowler, T. Wu. 2010. Bi-criteria analysis of ambulance diversion policies. In *Proceedings of the 2010 Winter Simulation Conference*, ed. B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.: 2315-2326.

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

Roche, K.T., J.K. Cochran. 2007. Improving patient safety by maximizing fast-track benefits in the emergency department – a queuing network approach. In *Proceedings of the 2007 Industrial Engineering Research* Conference: 619-624.

Ross, S.M. 2004. *Stochastic processes*. New York, NY: John Wiley & Sons, Inc.

Schull, M.J., K. Lazier, M. Vermeulen, S. Mawhinney, L.J. Morrison. 2003. Emergency department contributors to ambulance diversion: A quantitative analysis. *Annals of Emergency Medicine* 41(4): 467-476.

Stidham, S. 1985. Optimal-control of admission to a queuing system. *IEEE Transactions on Automatic Control* 30(8): 705-713.

Tan, P.N., M. Steinbach, V. Kumar. 2006. *Introduction to data mining*. Boston, MA: Pearson Education, Inc.

United States Department of Health and Human Services. 2010. National health expenditure data: NHE projections 2009-2019. Accessed on July 12, 2010. http://www.cms.hhs.gov/NationalHealthExpendData.

United States General Accounting Office. 2003. Hospital emergency departments: crowded conditions vary among hospitals and communities. GAO-03-460

United States General Accounting Office. 2009. Hospital emergency departments: crowding continues to occur, and some patients wait longer than recommended time frames. GAO-09-347.

Vilke, G.M., L. Brown, P. Skogland, C. Simmons, D.A. Guss. 2004a. Approach to decreasing emergency department ambulance diversion hours. *Journal of Emergency Medicine* 26(2): 189-192.

Vilke, G.M., E.M. Castillo, M.A. Metz, L.U. Ray, P.A. Murrin, R. Lev, T.C. Chan. 2004b. Community trial to decrease ambulance diversion hours: The San Diego county patient destination trial. *Annals of Emergency Medicine* 44(4): 295-303.

Wilhelm, E.M., J. Peck, S. Schoening, T. Lee. 2008. Reducing emergency department overcrowding-five patient buffer concepts in comparison. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S.J. Mason, R.R. Hill, L. Monch, O. Rose, T. Jefferson, and J.W. Fowler, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.: 1516-1525.

Williams, R.M. 2006. Ambulance diversion: Economic and policy considerations. *Annals of Emergency Medicine* 48(6): 711-712.

Yankovic, N., S. Glied, L.V. Green, M. Grams. 2010. The impact of ambulance diversion on heart attack deaths. *Inquiry* 47(1): 81-91.

APPENDIX A

PROOF OF THEOREMS

Proof of Theorem 1

A similar framework to Theorem 1 of Carr and Duenyas (2000) is followed. They characterize the production decisions in their paper using a monotonous threshold curve. Hence, most of the analysis and notations follow the proof of their Theorem 1. In the remainder of this proof the symbols ↑ and ↓ denote non-decreasing and non-increasing respectively.

First, two functions that allow notational simplicity are defined.

$$\Omega_{OFF}\, h^*(n_1,n_2) = T_1(n_1) + h^*(n_1 + 1, n_2), \tag{A.1}$$

$$\Omega_{ON}\, h^*(n_1,n_2) = T^D{}_1 + h^*(n_1,n_2). \tag{A.2}$$

Note that $\frac{\lambda^A}{\nu}\Omega_{OFF}\, h^*(n_1,n_2)$ and $\frac{\lambda^A}{\nu}\Omega_{ON}\, h^*(n_1,n_2)$ denote the additional ETP added when $p^A{}_1 = 1$ and the ambulance is accepted (i.e. diversion is off) or the ambulance is diverted (i.e. diversion is on), respectively. In other words, at state (n_1,n_2), it is optimal to divert an ambulance if $\Omega_{OFF}\, h^*(n_1,n_2) > \Omega_{ON}\, h^*(n_1,n_2)$, and accept the ambulance otherwise. For the existence of the threshold, $\Delta(n_1)$, the inequality $\Omega_{OFF}\, h^*(n_1,n_2) > \Omega_{ON}\, h^*(n_1,n_2)$ must hold for all $(n_1,n_2) \in S$ such that $n_2 > \Delta(n_1)$. Note that, if $\Omega_{OFF}\, h^*(n_1,n_2) - \Omega_{ON}\, h^*(n_1,n_2)$ is ↑ in n_2 for $(n_1,n_2) \in S$, as depicted in Figure A.1 (*left*), this condition is trivially satisfied. Hence, let **C1**(h^*) be a sufficient condition for the existence of the threshold.

C1(h^*) : $\Omega_{OFF}\, h^*(n_1,n_2) - \Omega_{ON}\, h^*(n_1,n_2)$ ↑ in n_2, for $(n_1,n_2) \in S$. (A.3)

Furthermore, if $\Omega_{OFF}\, h^*(n_1,n_2) - \Omega_{ON}\, h^*(n_1,n_2)$ ↑ in n_1 for $(n_1,n_2) \in S$, then $\Delta(n') \leq \Delta(n_2)$ for all $n' > n_2$, which gives the monotonocity of the threshold (see Figure A.1(*right*)). Let **C2**(h^*) be this sufficient condition and depict the change

of $\Omega_{OFF}\, h^*(n_1,n_2)$ - $\Omega_{ON}\, h^*(n_1,n_2)$ when n_1 is increased by one if **C1**(h^*) and **C2**(h^*) hold.

C2(h^*) : $\Omega_{OFF}\, h^*(n_1,n_2)$ - $\Omega_{ON}\, h^*(n_1,n_2) \uparrow$ in n_1, for $(n_1,n_2) \in S$. (A.4)

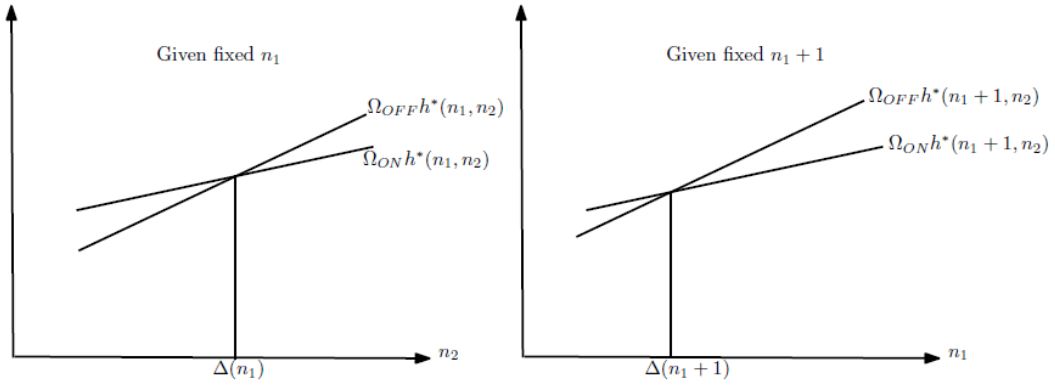

Figure A.1. Sample illustration of the changes in $\Omega_{OFF}\, h^*(n_1,n_2)$ and $\Omega_{ON}\, h^*(n_1,n_2)$ with respect to changes in n_2 under conditions C1(h^*) and C2(h^*). Since $\Omega_{OFF}\, h^*(n_1,n_2)$ - $\Omega_{ON}\, h^*(n_1,n_2)$ is $\uparrow$ in n_2, the threshold can be achieved (*left*). Furthermore, since $\Omega_{OFF}\, h^*(n_1,n_2)$ - $\Omega_{ON}\, h^*(n_1,n_2)$ is $\uparrow$ in n_1 the increase in n_1 decreases gap in between $\Omega_{OFF}\, h^*(n_1,n_2)$ and $\Omega_{ON}\, h^*(n_1,n_2)$, which results in a lower threshold value at n_1 +1 (*right*).

Simpler sufficient conditions are obtained for **C1**(h^*) and **C2**(h^*). From the definitions of $\Omega_{OFF}\, h(n_1,n_2)$ and $\Omega_{ON}\, h(n_1,n_2)$, one has:

$$\Omega_{OFF}\, h(n_1,n_2) - \Omega_{ON}\, h(n_1,n_2) \;=\; T_1(n_1) - T^D{}_1 + h^*(n_1+1,\, n_2) - h^*(n_1,n_2).$$

Next, **C3**(h^*), supermodularity of h^*, and **C4**(h^*) are defined. They ensure that **C1**(h^*) and **C2**(h^*) hold.

C3(h^*) : $h^*(n_1,n_2+1) - h^*(n_1,n_2) \uparrow$ in n_1, for $(n_1,n_2) \in S$. (A.4)

C4(h^*) : $T_1(n_1) + h^*(n_1+1,\, n_2) - h^*(n_1,n_2) \uparrow$ in n_1, for $(n_1,n_2) \in S$. (A.5)

Note that **C3**(h^*) implies $h^*(n_1+1,\, n_2) - h^*(n_1,n_2) \uparrow$ in n_2. Consequently, it is sufficient to show that **C3**(h^*), and **C4**(h^*) hold to complete the proof. The

desired conditions are proved using induction, and additional notation is defined similar to Carr and Duenyas (2000) that enables notational simplicity.

$$D_1t(n_1,n_2) = t(n_1,n_2 + 1) - t(n_1,n_2), \tag{A.6}$$

$$D_2t(n_1,n_2) = T_1(n_1) + t(n_1 + 1 ,n_2) - t(n_1,n_2). \tag{A.7}$$

DEFINITION 1. V is the set of functions on S such that if $t \in V$, then (i) $D_1t(n_1,n_2)$ is $\uparrow$ in n_1, and (ii) $D_2t(n_1,n_2)$ is $\uparrow$ in n_1 for $(n_1,n_2) \in S$.

Note that, if $h^* \in V$ then **C3**(h^*) and **C4**(h^*) hold, which completes this proof. In Lemma 1, it is shown that given a function $t \in V$, one can have $K_it(n_1,n_2) \in V$, $i \in \{1, 2, 3, 4, 5\}$, where $K_it(n_1,n_2) \in V$, $i \in \{1, 2, 3, 4, 5\}$ are defined in Equations (A.8)-(A10).

$$K_1t(n_1,n_2) = T_1(n_1) + t(n_1 + 1,n_2),\ K_2t(n_1,n_2) = T_2(n_2) + t(n_1,n_2 + 1) \tag{A.8}$$

$$K_3t(n_1,n_2) = t(n_1 ,n_2),\quad K_4t(n_1,n_2) = \min\{T^D{}_1 + t(n_1,n_2),\ T_1(n_1) + t(n_1 + 1,\ n_2)\} \tag{A.9}$$

$$\begin{aligned} K_5t(n_1,n_2) = {} & \frac{\lambda^W p^W{}_1}{\nu} K_1t(n_1,n_2) + \frac{\lambda^W \left(1 - p^W{}_1\right)}{\nu} K_2t(n_1,n_2) \\ & + \frac{\tilde{c}_1\mu_1}{\nu} K_3t(n_1 - 1,n_2) + \frac{\tilde{c}_2\mu_2}{\nu} K_3t(n_1,n_2 - 1) \\ & + \frac{\lambda^A}{\nu} K_4t(n_1,n_2) + \left(1 - \frac{\lambda^W + \lambda^A + \tilde{c}_1\mu_1 + \tilde{c}_2\mu_2}{\nu}\right) K_3t(n_1,n_2), \end{aligned} \tag{A.10}$$

LEMMA 1. *If $t \in V$ then $K_it \in V$ for $i = \{1, 2, 3, 4, 5\}$.*

Proof of Lemma 1. In this proof, $D_1K_it(n_1,n_2)$ is $\uparrow$ in n_1 for $i \in \{1, 2, 3, 4, 5\}$ is shown. The proofs for $D_2K_it(n_1,n_2)$ is $\uparrow$ in n_1 are omitted, but they can be obtained similarly.

Noting that $t \in V$, **C1**(t), **C2**(t), **C3**(t) and **C4**(t) hold, then

$$D_1K_1t(n_1,n_2) = t(n_1 + 1, n_2 + 1) - t(n_1 + 1, n_2), \tag{A.11}$$

$$D_1K_2t(n_1,n_2) = T_2(n_2 + 1) - T_2(n_2) + t(n_1, n_2 + 2) - t(n_1, n_2 + 1), \tag{A.12}$$

$$D_1K_3t(n_1,n_2) = t(n_1,n_2 + 1) - t(n_1,n_2). \tag{A.13}$$

$D_1K_1t(n_1,n_2)$, $D_1K_2t(n_1,n_2)$ and $D_1K_3t(n_1,n_2)$ are $\uparrow$ in n_1 from **C3**(t). Rewriting $D_1K_4t(n_1,n_2)$ in terms of $\Omega_{OFF}\,t$ and $\Omega_{ON}\,t$, Equation (A.14) is obtained.

$$D_1K_4t(n_1,n_2) = \min\{\Omega_{OFF}\,t(n_1,n_2 + 1), \Omega_{ON}\,t(n_1,n_2 + 1)\}$$
$$- \min\{\Omega_{OFF}\,t(n_1,n_2), \Omega_{ON}\,t(n_1,n_2)\}. \tag{A.14}$$

Next, it is shown that $D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) \geq 0$, which is sufficient to show that $D_1K_4t(n_1,n_2)$ is $\uparrow$ in n_1. For notational simplicity let $I(x)$ be the indicator function, where $I(x) = 1$ if $x \geq 0$ and $I(\mathrm{x}) = 0$ otherwise. Furthermore, let

$$\Theta = \begin{vmatrix} I(\Omega_{OFF}\,t(n_1,n_2 + 1) - \Omega_{ON}\,t(n_1,n_2 + 1)) & I(\Omega_{OFF}\,t(n_1{+}1, n_2{+}1) - \Omega_{ON}\,t(n_1{+}1, n_2{+}1)) \\ I(\Omega_{OFF}\,t(n_1,n_2) - \Omega_{ON}\,t(n_1,n_2)) & I(\Omega_{OFF}\,t(n_1{+}1, n_2) - \Omega_{ON}\,t(n_1{+}1, n_2)) \end{vmatrix} \tag{A.15}$$

For example, $\Theta = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}$ indicates that $\Omega_{OFF}\,t(n_1,n_2) < \Omega_{ON}\,t(n_1,n_2)$, $\Omega_{OFF}\,t(n_1{+}1, n_2) \geq \Omega_{ON}t(n_1{+}1, n_2)$, $\Omega_{OFF}\,t(n_1,n_2{+}1) \geq \Omega_{ON}\,t(n_1,n_2{+}1)$ and $\Omega_{OFF}\,t(n_1{+}1, n_2{+}1) \geq \Omega_{ON}\,t(n_1{+}1, n_2{+}1)$.

From **C1**(t) and **C2**(t), $\Omega_{OFF}\,t(n_1,n_2) - \Omega_{ON}\,t(n_1,n_2)$ is $\uparrow$ in n_1 and n_2. Consequently, there are six possible outcomes for Θ given as $\left\{\begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix}, \begin{vmatrix} 0 & 0 \\ 0 & 0 \end{vmatrix}, \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix}, \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}, \begin{vmatrix} 0 & 1 \\ 0 & 1 \end{vmatrix}, \begin{vmatrix} 1 & 1 \\ 0 & 0 \end{vmatrix}\right\}$, which are analyzed separately.

Case I. $\Theta = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix}$:

$$D_1K_4t(n_1 + 1, \text{n2}) - D_1K_4t(n_1,n_2) = \Omega_{ON}\, t(n_1 + 1, n_2 + 1) - \Omega_{ON}\, t(n_1 + 1, n_2)$$
$$- \Omega_{ON}\, t(n_1,n_2 + 1) + \Omega_{ON}\, t(n_1,n_2)$$
$$= [T^D{}_1 + t(n_1 + 1, n_2 + 1)] - [T^D{}_1 + t(n_1 + 1, n_2)]$$
$$- [T^D{}_1 + t(n_1, n_2 + 1)] + [T^D{}_1 + t(n_1,n_2)]$$
$$= t(n_1 + 1, n_2 + 1) - t(n_1 + 1, n_2)$$
$$- (t(n_1, n_2 + 1) - t(n_1,n_2)). \qquad \text{(A.16)}$$

Note that, from **C3**(t) right hand side of Equation (A.16) is nonnegative. Hence, $D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) \geq 0$ is satisfied.

Case II. $\Theta = \begin{vmatrix} 0 & 0 \\ 0 & 0 \end{vmatrix}$:

$$D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) = \Omega_{OFF}\, t(n_1 + 1, n_2 + 1) - \Omega_{OFF}\, t(n_1 + 1, n_2)$$
$$- \Omega_{OFF}\, t(n_1,n_2 + 1) + \Omega_{OFF}\, t(n_1,n_2)$$
$$= t(n_1 + 2, n_2 + 1) - t(n_1 + 2, n_2)$$
$$- (t(n_1 + 1, n_2 + 1) - t(n_1 + 1, n_2)). \qquad \text{(A.17)}$$

Similarly from **C3**(t) right hand side of Equation (A.17) is nonnegative, and the desired condition is satisfied.

Case III. $\Theta = \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix}$:

$$D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) = \Omega_{ON}t(n_1 + 1, n_2 + 1) - \Omega_{OFF}\, t(n_1 + 1, n_2)$$
$$- \Omega_{OFF}\, t(n_1, n_2 + 1) + \Omega_{OFF}\, t(n_1,n_2)$$
$$= [T^D{}_1 + t(n_1 + 1, n_2)] - [T_1(n_1 + 1) + t(n_1 + 2, n_2)]$$

$$= \Omega_{ON}\, t(n_1 + 1, n_2) - \Omega_{OFF}\, t(n_1 + 1, n_2). \qquad \text{(A.18)}$$

which is nonnegative under Case III.

Case IV. $\Theta = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}$:

$$\begin{aligned} D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) &= \Omega_{ON}\, t(n_1 + 1, n_2 + 1) - \Omega_{ON}\, t(n_1 + 1, n_2) \\ &\quad - \Omega_{ON}\, t(n_1, n_2 + 1) + \Omega_{OFF}\, t(n_1,n_2) \\ &= [T_1(n_1) + t(n_1 + 1, n_2 + 1)] - [T^D{}_1 + t(n_1, n_2 + 1)] \\ &= \Omega_{OFF}\, t(n_1, n_2 + 1) - \Omega_{ON}\, t(n_1, n_2 + 1). \qquad \text{(A.19)} \end{aligned}$$

which is nonnegative.

Case V. $\Theta = \begin{vmatrix} 0 & 1 \\ 0 & 1 \end{vmatrix}$:

$$\begin{aligned} D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) &= \Omega_{ON}\, t(n_1 + 1, n_2 + 1) - \Omega_{ON}\, t(n_1 + 1, n_2) \\ &\quad - \Omega_{OFF}\, t(n_1, n_2 + 1) + \Omega_{OFF}\, t(n_1,n_2) \\ &= 0. \qquad \text{(A.20)} \end{aligned}$$

Hence, $D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) \geq 0$.

Case VI. $\Theta = \begin{vmatrix} 1 & 1 \\ 0 & 0 \end{vmatrix}$:

$$\begin{aligned} D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) &= \Omega_{ON}\, t(n_1 + 1, n_2 + 1) - \Omega_{OFF}\, t(n_1 + 1, n_2) \\ &\quad - \Omega_{ON}\, t(n_1, n_2 + 1) + \Omega_{OFF}\, t(n_1,n_2) \\ &= T_1(n_1) + t(n_1 + 1, n_2 + 1) + t(n_1 + 1, n_2) \\ &\quad - (T_1(n_1 + 1) + t(n_1 + 2, n_2) + t(n_1, n_2 + 1)). \qquad \text{(A.21)} \end{aligned}$$

From $\Omega_{OFF}\ t(n_1 + 1,\ n_2) \leq \Omega_{ON}\ t(n_1 + 1,\ n_2)$, one can have $T_1(n_1 + 1) + t(n_1 + 2,\ n_2) \leq T^D{}_1 + t(n_1 + 1,\ n_2)$. Hence, using right hand side of Equation (A.21), the following inequality is obtained

$$\begin{aligned} D_1K_4t(n_1 + 1, n_2) - D_1K_4t(n_1,n_2) &\geq T_1(n_1) + t(n_1 + 1, n_2 + 1) + t(n_1 + 1, n_2) \\ &\quad - (T^D{}_1 + t(n_1 + 1, n_2) + t(n_1, n_2 + 1)) \\ &= T_1(n_1) + t(n_1 + 1, n_2 + 1) - (T^D{}_1 + t(n_1, n_2 + 1)) \\ &= \Omega_{OFF}\ t(n_1, n_2 + 1) - \Omega_{ON}\ t(n_1,\ n_2 + 1) \qquad \text{(A.22)} \end{aligned}$$

Note that right hand side of the Equation (A.22) is nonnegative. Hence the desired condition is satisfied.

In summary, all possible cases are considered, completing the proof for $D_1K_4t(n_1 + 1,\ n_2) - D_1K_4t(n_1,n_2) \geq 0$. Since, $K_5t(n_1,n_2)$ is the linear combination of $K_it(n_1,n_2)$ for $i \in \{1, 2, 3, 4\}$, $K_5t(n_1,n_2)$ is $\uparrow$ in n_1 as well.

□

Now consider a value iteration algorithm that solves Equation (3.5) recursively by

$$h_{k+1}(n_1,n_2) = K_5h_k(n_1,n_2), \qquad \text{(A.23)}$$

where $h_0(n_1,n_2) = 0$ for $(n_1,n_2) \in S$. Since $h_0(n_1,n_2) = 0$ for $(n_1,n_2) \in S$, trivially $h_0 \in V$ is obtained, which indicates that $h_1 \in V$ from Lemma 1 and Equation (A.23). Now assume for some k, $h_k \in V$. From Lemma 1 and Equation (A.23) again, one can obtain that $h_{k+1} \in V$. From Inequality (3.1) the model is unichain, therefore $\lim_{k\to\infty} h_k(n_1,n_2) = h^*(n_1,n_2)$ (Bertsekas 2001), and hence, $h^* \in V$, as well. As a result, **C3**(h^*) and **C4**(h^*) hold, which completes the proof. □

Proof of Theorem 2

In this proof, it is only shown that $\Delta(n_1)$ is non-decreasing in $T^D{}_1$ and omit the proof for $T^D{}_2$, which follows similar lines.

Two different problems are considered that are solved by $T^D{}_1 = \underline{T^D{}_1}$ and $T^D{}_1 = \overline{T^D{}_1}$, where $\underline{T^D{}_1} \leq \overline{T^D{}_1}$. The optimal relative effect of starting in state (n_1,n_2) is denoted as $\underline{h^*}(n_1,n_2)$ and $\overline{h^*}(n_1,n_2)$ in first and second problems, respectively. Following the proof of Theorem 1, for monotonocity of $T^D{}_1$ in $\Delta(n_1)$, $\Omega_{OFF}\ h^*(n_1,n_2)$ - $\Omega_{ON}\ h^*(n_1,n_2)$ is needed to be non-increasing in $T^D{}_1$. Hence, **C5**($\underline{h^*}$, $\overline{h^*}$) provides a sufficient condition for this proof.

$$\textbf{C5}(\underline{h^*}, \overline{h^*}) : \Omega_{OFF}\ \overline{h^*}(n_1,n_2) - \Omega_{ON}\ \overline{h^*}(n_1,n_2) \leq \Omega_{OFF}\ \underline{h^*}(n_1,n_2) - \Omega_{ON}\ \underline{h^*}(n_1,n_2),$$

$$\text{for } (n_1,n_2) \in S, \tag{A.24}$$

where $\Omega_{OFF}\ \underline{h^*}(n_1,n_2)$ and $\Omega_{OFF}\ \overline{h^*}(n_1,n_2)$ can be obtained from Equation (A.1), and $\Omega_{ON}\ \underline{h^*}\ (n_1,n_2)$ and $\Omega_{ON}\ \overline{h^*}(n_1,n_2)$ are given as in Equation (A.25).

$$\Omega_{ON}\ \overline{h^*}(n_1,n_2) = \overline{T^D{}_1} + \overline{h^*}(n_1,n_2), \qquad \Omega_{ON}\ \underline{h^*}(n_1,n_2) = \underline{T^D{}_1} + \underline{h^*}(n_1,n_2). \tag{A.25}$$

Next, **C6**($\underline{h^*}$, $\overline{h^*}$) is defined and it is a sufficient condition for **C5**($\underline{h^*}$, $\overline{h^*}$) from Equations (A.1) and (A.2).

$$\textbf{C6}(\underline{h^*}, \overline{h^*}) : \overline{h^*}(n_1+1, n_2) - \overline{h^*}(n_1,n_2) - \overline{T^D{}_1} \leq \underline{h^*}(n_1+1, n_2) - \underline{h^*}(n_1,n_2) - \underline{T^D{}_1},$$

$$\text{for } (n_1,n_2) \in S. \tag{A.26}$$

Similar to the proof of Theorem 1, induction is used to show that **C6**($\underline{h^*}$, $\overline{h^*}$) holds. For notational simplicity, $D_3t(n_1,n_2)$ and the set V' are defined.

$$D_3t(n_1,n_2) = t(n_1 + 1, n_2) - t(n_1,n_2) - T^D{}_1 \quad \text{(A.27)}$$

DEFINITION 2. V' is the set of functions on S such that if $(\underline{t}, \bar{t}) \in V'$, then $D_3\bar{t}(n_1,n_2) \leq D_3\underline{t}(n_1,n_2)$ for $(n_1,n_2) \in S$.

Similar to the definitions of $\Omega_{ON}\ \underline{h^*}(n_1,n_2)$ and $\Omega_{ON}\ \overline{h^*}(n_1,n_2)$, in the remainder of this proof, $\underline{T^D{}_1}$ and $\overline{T^D{}_1}$ are used within the definitions of $D_3\underline{t}$, $D_3\bar{t}$, $D_3K_i\underline{t}$, $D_3K_i\bar{t}$, for $i \in \{1, 2, 3, 4, 5\}$. Lemma 2 shows that the desired conditions are preserved under the operation K_i, $i \in \{1, 2, 3, 4, 5\}$.

LEMMA 2. *If* $(\underline{t},\bar{t}) \in V'$ *then* $(K_i\underline{t}, K_i\bar{t}) \in V'$ *for* $i \in \{1, 2, 3, 4, 5\}$.

Proof of Lemma 2. In this proof, the inequality $D_3K_i\bar{t}(n_1,n_2) \leq D_3K_i\underline{t}(n_1,n_2)$ is shown for $i \in \{1, 2, 3, 4, 5\}$. One observes that, from $(\underline{t},\bar{t}) \in V'$, **C5**$(\underline{t},\bar{t})$ and **C6**$(\underline{t},\bar{t})$ hold. Furthermore, from Theorem 1, **C1**$(\underline{t})$, **C2**$(\underline{t})$, **C1**$(\bar{t})$ and **C2**$(\bar{t})$ hold. Plugging in K_i, i = 1, 2, 3 into Equation (A.27), the following equations are obtained.

$$D_3K_1t(n_1,n_2) = T_1(n_1 +1) + t(n_1 + 2, n_2) - T_1(n_1) - t(n_1 + 1,n_2) - T^D{}_1, \quad \text{(A.28)}$$

$$D_3K_2t(n_1,n_2) = t(n_1 + 1, n_2 + 1) - t(n_1, n_2 + 1) - T^D{}_1, \quad \text{(A.29)}$$

$$D_3K_3t(n_1,n_2) = t(n_1 + 1, n_2) - t(n_1, n_2) - T^D{}_1, \quad \text{(A.30)}$$

Using $D_3K_it(n_1,n_2)$ for i = 1, 2, 3 and **C6**$(\underline{t},\bar{t})$, it is trivial to show that $D_3K_i\bar{t}(n_1,n_2) \leq D_3K_i\underline{t}(n_1,n_2)$ for i = 1, 2, 3. To show that $D_3K_4\bar{t}(n_1,n_2) \leq D_3K_4\underline{t}(n_1,n_2)$, a case by case analysis is implemented and Θ' is defined as follows.

$$\Theta' = \begin{vmatrix} I(\Omega_{OFF}\ \underline{t}(n_1,n_2) - \Omega_{ON}\ \underline{t}(n_1,n_2)) & I(\Omega_{OFF}\ \underline{t}(n_1+1,n_2) - \Omega_{ON}\ \underline{t}(n_1+1,n_2)) \\ I(\Omega_{OFF}\ \bar{t}(n_1,n_2) - \Omega_{ON}\ \bar{t}(n_1,n_2)) & I(\Omega_{OFF}\ \bar{t}(n_1+1,n_2) - \Omega_{ON}\ \bar{t}(n_1+1,n_2)) \end{vmatrix} \quad \text{(A.31)}$$

Given **C5**($\underline{t}, \bar{t}$), **C1**($\underline{t}$) and **C1**($\bar{t}$), there are six possible values for Θ' given as $\left\{\begin{vmatrix}1&1\\1&1\end{vmatrix}, \begin{vmatrix}1&1\\0&0\end{vmatrix}, \begin{vmatrix}1&1\\0&1\end{vmatrix}, \begin{vmatrix}0&0\\0&0\end{vmatrix}, \begin{vmatrix}0&1\\0&1\end{vmatrix}, \begin{vmatrix}0&1\\0&0\end{vmatrix}\right\}$.

Case I. $\Theta' = \begin{vmatrix}1&1\\1&1\end{vmatrix}$. Using Equations (A.9) and (A.27):

$$D_3K_4t(n_1,n_2) = [T^D{}_1 + t(n_1+1, n_2)] - [T^D{}_1 + t(n_1, n_2)] - T^D{}_1$$

$$= t(n_1+1, n_2)] - t(n_1, n_2) - T^D{}_1. \quad \text{(A.31)}$$

Using **C6**($\underline{t}, \bar{t}$), one clearly obtains

$$D_3K_4\bar{t}(n_1,n_2) = \bar{t}(n_1+1, n_2) - \bar{t}(n_1,n_2) - \overline{T^D{}_1}$$

$$\leq \underline{t}(n_1+1, n_2) - \underline{t}(n_1,n_2) - \underline{T^D{}_1}$$

$$= D_3K_4\underline{t}(n_1,n_2). \quad \text{(A.32)}$$

Case II. $\Theta' = \begin{vmatrix}1&1\\0&0\end{vmatrix}$:

$$D_3K_4\bar{t}(n_1,n_2) = [T_1(n_1+1) + \bar{t}(n_1+2, n_2)] - [T_1(n_1) + \bar{t}(n_1+1, n_2)] - \overline{T^D{}_1} \quad \text{(A.33)}$$

$$D_3K_4\underline{t}(n_1,n_2) = \underline{t}(n_1+1, n_2) - \underline{t}(n_1,n_2) - \underline{T^D{}_1}\ . \quad \text{(A.34)}$$

Since, $\Omega_{OFF}\ \bar{t}(n_1+1, n_2) \leq \Omega_{ON}\ \bar{t}(n_1+1, n_2)$, one obtains the following inequality.

$$D_3K_4\bar{t}(n_1,n_2) \leq [\overline{T^D{}_1} + \bar{t}(n_1+1, n_2)] - [T_1(n_1) + \bar{t}(n_1+1, n_2)] - \overline{T^D{}_1}$$

$$= -T_1(n_1). \quad \text{(A.35)}$$

Similarly, using the inequality, $-\Omega_{OFF}\underline{t}(n_1,n_2) \leq -\Omega_{ON}\underline{t}(n_1,n_2)$, the Inequality (A.36) can be written

$$D_3K_4\underline{t}(n_1,n_2) \geq \underline{t}(n_1+1, n_2) - T_1(n_1) - \underline{t}(n_1+1,n_2) = -T_1(n_1). \quad \text{(A.36)}$$

Combining Inequalities (A.35) and (A.36), the following inequality is obtained:

$$D_3K_4\bar{t}(n_1,n_2) \leq D_3K_4\underline{t}(n_1,n_2).$$

Case III. $\Theta' = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}$. In this case, $D_3K_4\underline{t}(n_1,n_2)$ is given as defined in Equation (A.34). $D_3K_4\bar{t}(n_1,n_2)$ is given in Equation (A.37)

$$D_3K_4\bar{t}(n_1,n_2) = [\overline{T^{D}_1} + \bar{t}(n_1+1, n_2)] - [T_1(n_1) + \bar{t}(n_1+1, n_2)] - \overline{T^{D}_1}$$

$$= -T_1(n_1). \quad \text{(A.37)}$$

Using the inequality $\Omega_{OFF}\underline{t}(n_1,n_2) \geq \Omega_{ON}\underline{t}(n_1,n_2)$, one clearly obtains

$$D_3K_4\bar{t}(n_1,n_2) \leq D_3K_4\underline{t}(n_1,n_2).$$

Case IV. $\Theta' = \begin{vmatrix} 0 & 0 \\ 0 & 0 \end{vmatrix}$:

$$D_3K_4\bar{t}(n_1,n_2) = [T_1(n_1+1) + \bar{t}(n_1+2, n_2)] - [T_1(n_1) + \bar{t}(n_1+1, n_2)] - \overline{T^{D}_1}. \quad \text{(A.38)}$$

$$D_3K_4\underline{t}(n_1,n_2) = [T_1(n_1+1) + \underline{t}(n_1+2, n_2)] - [T_1(n_1) + \underline{t}(n_1+1, n_2)] - \underline{T^{D}_1}. \quad \text{(A.39)}$$

and the desired condition can be simply achieved using **C6**($\underline{t}$,$\bar{t}$).

Case V. $\Theta' = \begin{vmatrix} 0 & 1 \\ 0 & 1 \end{vmatrix}$:

$$D_3K_4\bar{t}(n_1,n_2) = [\overline{T^{D}_1} + \bar{t}(n_1+1, n_2)] - [T_1(n_1) + \bar{t}(n_1+1, n_2)] - \overline{T^{D}_1}. \quad \text{(A.40)}$$

$$D_3K_4\underline{t}(n_1,n_2) = [\underline{T^{D}_1} + \underline{t}(n_1+1, n_2)] - [T_1(n_1) + \underline{t}(n_1+1, n_2)] - \underline{T^{D}_1}. \quad \text{(A.41)}$$

Rearranging the terms, the following equality is obtained: $D_3K_4\bar{t}(n_1,n_2) = D_3K_4\underline{t}(n_1,n_2)$, which is sufficient for $D_3K_4\bar{t}(n_1,n_2) \leq D_3K_4\underline{t}(n_1,n_2)$.

Case VI. Θ' = $\begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix}$. In this case, $D_3K_4\bar{t}(n_1,n_2)$ and $D_3K_4\underline{t}(n_1,n_2)$ are given in Equations (A.38) and (A.41), respectively. The desired condition can be shown using the inequality $\Omega_{OFF}\,\bar{t}(n_1 + 1, n_2) \leq \Omega_{ON}\,\bar{t}(n_1 + 1, n_2)$.

Since, $D_3K_5\bar{t}(n_1,n_2)$ and $D_3K_5\underline{t}(n_1,n_2)$ are linear combinations of $D_3K_4\bar{t}(n_1,n_2)$ and $D_3K_4\underline{t}\ (n_1,n_2)$, $i \in \{1, 2, 3, 4\}$, it is trivial to show that $D_3K_5\bar{t}(n_1,n_2) \leq D_3K_5\underline{t}(n_1,n_2)$.

The remainder of the proof follows from the induction argument discussed at the end of proof of Theorem 1. □

APPENDIX B

INPUT DATA FOR THE EMERGENCY CARE DELIVERY SYSTEM MODEL AND PARAMETERS FOR GA

Input Data for Simulation model

The generic hospital described in Chapter 4 was built using C++ with information published in different sources. Arrival rates are derived from Cochran and Roche (2009), which shows a non-stationary arrival process to a single ED. Arrival pattern shown in Figure 2.3 was also used to model the arrivals in the model of Chapter 4. As stated in all the chapters, the arrival rate shown in that Figure has also been observed in other papers and official reports across the United States.

The walk-ins are scheduled independently for each hospital in the simulation model according to the time of the day and scenario. Thus, a scenario considering a different relative size among hospitals multiplies the arrival rate shown in Figure 2.3 by the indicated factor: 10% or 20%.

On the other hand, the rate of ambulance arrivals described in Figure 2.3 is used to model the frequency of new emergency patients in the simulation. Thus, the emergency patient generator module creates patients according to the ambulance rate and assigns a random location in the ECDS. Uniform distribution is assumed to define the location of the new emergency patient in the two-dimensional space. Upon the location assignment, the simulation schedules the arrival of the patient to the appropriate hospital based on the destination and diversion policies. Green (2006) proposes a set of arguments to assume Poisson process for the arrivals to healthcare systems. Hence, this chapter assumes Poisson process for all its arrivals.

In order to schedule ambulance arrivals to an appropriate hospital, the transportation time is estimated by τ x $M(l, H_i)$, where $M(l, H_i)$ is the Manhattan distance between the emergency location and the selected hospital, and τ is the transportation time per mile. In order to take into account the uncertainty of the transportation time, this chapter assumes a distribution for τ, such that $\tau \sim$ Normal(1.25, 0.5). This implies that the average transportation time is 1.25 minutes per mile, which is similar to the data presented by Google Maps as transportation time per mile in Maricopa County, AZ.

The severity level assigned to each patient depends on the arrival mode. The percentages of each severity level are the same than those shown in Table 2.1. The overall percentages are similar to information presented in Cochran and Roche (2009) and close to the national average (Centers for Disease Control and Prevention 2010).

The treatment times in the ED depend on the severity level of the patients and they were derived from Cochran and Roche (2009). The mean treatment time per severity level is shown in Table B1. This chapter assumes that the treatment time follows an Erlang distribution with shape parameter of 3. This assumption produces a distribution with coefficient of variation of to $1/\sqrt{3}$, which is similar to the value observed by Cochran and Roche (2009) for the coefficient of variation of treatment in an ED. In addition, the probability density function of the Erlang distribution is similar to other distributions used to characterize the treatment times in EDs (Hoot et al. 2008).

Table B.1. Mean Treatment Times in the ED.

Severity Level	Mean Treatment Time (min)
1	273
2	273
3	140
4	106
5	30

After ending treatment in the ED, the patients can be admitted to the IP unit with a probability that depends on the severity level. These probabilities are presented in Table B.2. The overall admission percentage is 15% which is in the range the average seen in metropolitan areas in the United States (Centers for Disease Control and Prevention 2010).

Table B.2. Admission probabilities to IP unit.

Severity Level	Admission Percentage
1	70
2	34
3	10
4	5
5	3
Overall	15

Direct admissions to IP occur according to a Poisson process with a mean of one admission per hour, which is similar to the total external arrival rates of the hospital analyzed in Cochran and Bharti (2006). The treatment time in the IP is also assumed to be an Erlang distribution with shape parameter equal to 3 and a mean of four days, which is similar to the data found by Cochran and Bharti (2006) and close to mean length of stay in IP units according to the Centers for Disease Control and Prevention (2010).

In order to model the LWOT patients, this chapter incorporates an approach presented by (Miller et al. 2009). The LWOT routine consists of removing patients from the queue if they have not been placed in a bed within 24 hours. This chapter assumes that LWOT patients go home or visit a non-emergency physician; therefore, they are not scheduled to arrive to another hospital in the model.

The hospitals in the model have 20 beds in the ED and 200 IP beds. The number of beds considered for the ED is similar to the median in the United States (Centers for Disease Control and Prevention 2006b) and the size of the IP unit is suitable of a medium-size hospital.

The simulation length for the research is fixed to six months after a warm-up period of one month and ten replications per strategy are considered. These parameters were defined after a set of pilot runs to obtain precise estimation of the performance measure of interest. In addition, Common Random Numbers (Banks et al. 2010) are used to expose the different strategies to the similar conditions and reduce the noise among them.

Parameters for the Genetic Algorithm

The GA used in this dissertation is a nondominated sorting genetic algorithm (NSGA-II) proposed by Deb et al. (2002), which uses front ranking and a crowding distance operator for the survivor selection of chromosome from generation to generation. The main elements and processes of the GA are described below.

Population: The initial population comprises policies randomly generated and a selection of good policies derived from a pilot run. The number of policies (chromosomes) kept from generation to generation is 20.

Parent Selection: Binary selection tournament is used to select two parent chromosomes from the population pool to generate a new chromosome (offspring). The selection of the parent chromosome is based on the front number and crowding distance is used as a tiebreaker.

Front: The front number of a specific policy *P* is related to the number of policies which dominate *P* (domination count). The nondominated policies of a set of policies have front number equal to one. Then, policies in front one are removed from the total set and the process repeats. The new set of nondominated policies is assigned to front two. The process repeats until a front number is assigned to all the policies.

Crowding distance: The crowding distance is related to the diversity of the policies. The crowding distance of a specific policy *P* is an estimation of the perimeter of the cuboid formed by the nearest policy neighbors of *P*. The policies with larger crowding distance are more likely to be included in the parent selection since diversity encourages exploring areas with low density of policies.

Recombination: The recombination strategy depends if the chromosome belongs to an SF policy or to an MF.

- Crossover for SF AD policy in an ECDS with two hospitals: Once that parents are selected (chromosomes X & Y), the offspring (chromosome Z) is obtained

by forming policy of hospital H_1 from chromosome X and policy of hospital H_2 from chromosome Z. This process is depicted in Figure B.1.

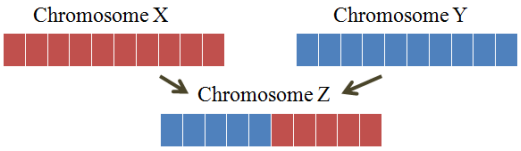

Figure B.1. Crossover for SF AD policy in an ECDS with 2 hospitals.

- Crossover for MF AD policy in an ECDS with two hospitals: Once that parents are selected (chromosomes X & Y), the offspring (chromosome Z) is obtained by uniform crossover. Thus, each gene in chromosome Z is defined by the value of the same gene in Chromosome X with probability *p* or the value is taken from Chromosome Y with probability 1-*p*. The value of *p* selected in this research is 0.5. This type of crossover is depicted in Figure B.2.

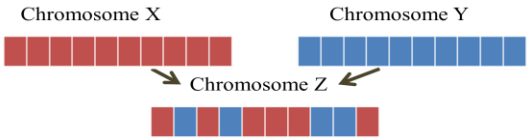

Figure B.2. Crossover for MF AD policy in an ECDS with 2 hospitals.

- Crossover for SF AD policy in an ECDS with three hospitals: Once that parents are selected (chromosomes X & Y), the offspring (chromosome Z) is obtained by forming policy of hospital H_1 from chromosome X with probability *p* or from chromosome Y with probability 1-*p*. The value of *p* selected in this research is 0.5. Policies for hospitals H_2 and H_3 in chromosome Z are defined in the same way. Figure B.3 depicts the process of

this crossover and shows some potential configurations of policies in chromosome Z.

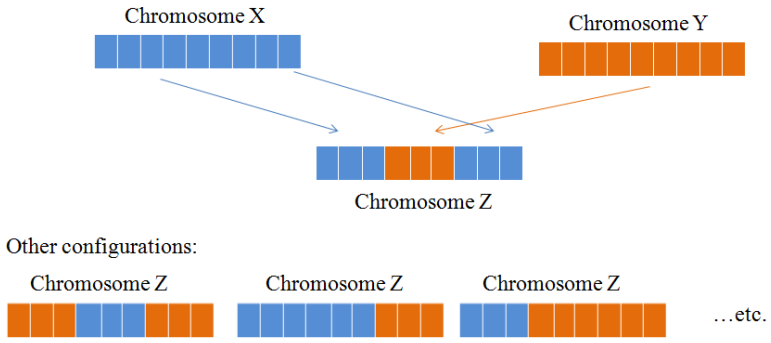

Figure B.3. Crossover for SF AD policy in an ECDS with 3 hospitals and potential outcomes.

These recombination strategies allow obtaining feasible chromosomes that can be used in the evaluation of the policies via simulation.

Mutation: The mutation procedure is a search mechanism that modifies the value of a gene in the policy of a hospital. The change in the gene is done with a probability p (p = 0.7 in this research). The new value of a gene g is given by:

$$g_{\text{new}} = g_{\text{current}} + z\sigma_g \tag{B.1}$$

where $z \sim$ Normal(0, 1). Thus, the new value of the gene is a neighbor value. The range of search is defined by the value of σ_g, which depends on the nature of the variable represented by the gene. Hence, for an MF AD policy, it is expected to have $\sigma_{NQ1} < \sigma_{NQ5}$.

Offspring generator: The pool of chromosomes, including the initial population in the generation and the offspring, comprises 40 policies.

Simulation: The fitness of a chromosome is given by the tuple of average-patient non-value added time in each hospital. The vector is obtained via simulation using the framework described in Section 4.2.

Survivor selection: The new population for the next generation in the genetic algorithm is determined by the front and crowding distance. The chromosomes with small number value and large crowding distance have higher priority to survive to the next generation.

Finally, the simulation-optimization algorithm stops after 30 generations and the best chromosomes that survived in the last generation are kept for the analysis.