DESIGNING EMERGENCY MEDICAL SERVICES PROCESSES TO MINIMIZE THE IMPACT OF AMBULANCE OFFLOAD DELAY

by

Mengyu Li

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
July 2019

To my loved ones.

Contents

List of Tables

List of Figures

Abstract

Ambulance offload delay (AOD) occurs when the care of incoming ambulance patients cannot be transferred immediately from paramedics to staff in a hospital emergency department (ED). This is typically due to ED congestions. In such situations, paramedics are responsible to provide patient care until an ED bed becomes available. AOD can negatively affect ambulance availability to future calls and reduce the efficiency of the emergency medical services (EMS) system. Hence, this problem has become a significant concern for many healthcare providers and is the focus of this dissertation.

In this dissertation, we develop several models to analyze AOD. With 12-months of emergency call data provided by the provincial EMS provider and local hospitals in Nova Scotia, Canada, we conduct an empirical analysis to measure the effects of AOD on the EMS system. The analyzed performance metrics include the number of ambulances at EDs, ambulance turnaround time, total call time, response time, and ambulance availability. The results indicate significant negative effects on all these metrics within the region experiencing AOD. AOD also has a negative impact on ambulance availability in adjacent regions for an EMS system with shared resources.

We then develop a decision-support tool using a novel hybrid decision tree model to predict the severity of AOD within 1 to 5 hours based on the current system status. The objective of this study is to provide a prediction model for EMS decision makers so that proactive interventions at different system states can be initiated to mitigate AOD. The hybrid algorithm shows improvements in the classification of this real-world problem when tested against a basic decision tree algorithm.

Finally, we develop an optimal ambulance destination policy using a discrete time, infinite-horizon, discounted Markov Decision Process. This model helps determine when it is advantageous to send appropriate patients to out-of-region EDs, which have longer transport times but shorter offload times. The optimal policy can significantly reduce AOD, time-to-ED bed for patients, and out-of-service time for paramedics at the expense of increased ambulances travel distances.

List of Abbreviations Used

AD	ambulance diversion
AOD	ambulance offload delay
CAD	computer aided dispatch
CART	classification and regression tree
CTAS	Canadian triage and acuity scale
DT	decision tree
ED	emergency department
EDIS	emergency department information system
EHS	emergency health services
EMS	emergency medical services
ePCR	electronic patient care reporting system
GIS	geographic information system
HRM	Halifax regional municipality
LOS	length of stay
MDP	Markov decision process
MIN	master incident number
ML	machine learning
NEDOCS	national emergency department overcrowding scale
NSHA	Nova Scotia health authority

OR	operations research
OZ	offload zone
PTU	patient transfer unit

Acknowledgements

I am most grateful to my supervisor Dr. Peter Vanberkel for his generous guidance, support and inspiration throughout my Ph.D. studies. Thank you for believing in me and taking me in as a graduate student with no formal education experience in the industrial engineering field.

I would like to thank the members of my supervisory committee, Dr. John Blake, Dr. Alix Cater, Dr. Jing Chen, and Dr. Corinne MacDonald. I am very grateful for their insightful comments and helpful suggestions that improved this research. A special thank goes to Dr. Xiang Zhong at University of Florida for sponsoring me as a visiting scholar at the university, as well as her kind support and inspiring advise on my research.

I wish to express my gratitude to all my friends and colleagues at the department of industrial engineering who were always there when I needed them, and for making this an enjoyable experience. Special thanks to Pin, Zhuojun, Lauren, and Tara, for their support during my graduate studies.

A very special gratitude goes out to all personnel from NSHA, EHS and EMCi who have helped and supported me with this research. With a special mention to Jan Jensen, Judah Goldstein, Terence Hawco, Del Kenley, EHS Operations Supervisors, and EMCi dispatch call centre. It was fantastic to have the opportunity to work with these talented and passionate people.

I am fortunate and grateful to have my partner, who later became my husband, Andrew MacIntosh, accompanying me through this long journey, with his unconditional love and support. Without him, I wouldn't have achieved this success.

Finally, I am forever indebted to my mother for her constant care, support, and encouragements.

Chapter 1

INTRODUCTION

Healthcare is an area of growing importance and cost around the world [1]. It is also a challenging area for operations research (OR) due to its unique characteristics. As our society ages, the demand and pressure on the health care system rises rapidly; therefore, the system faces increasing challenges related to process efficiency.

One of the key components in healthcare is pre-hospital care provided by Emergency Medical Services (EMS). EMS are public safety systems responsible for providing emergency assistance and for protecting public health and safety [2]. The goal of such systems is to respond quickly to population calls, to provide first aid services, and to transfer patients to the emergency department (ED) of an appropriate hospital when needed [2–4]. In life-threatening emergency situations, the ability of EMS providers to quickly respond will mean less adverse effects for the patients involved as timely care is crucial. Any delay is highly undesirable from a patient safety perspective [2]. Therefore, EMS providers continuously seek best practices, especially in a world where an aging population adds pressure to the health care system [5]. Meanwhile, EMS systems also need to sustain themselves financially (i.e., capital and operation costs). EMS providers are challenged to perform their services more effectively and efficiently to meet their own budgetary and performance targets. To achieve both timeliness and economic objectives, limited EMS resources (e.g., emergency vehicles, paramedics) must be managed efficiently in an environment with a high level of uncertainty related to demand characteristics and resources availability [3].

My Ph.D. research focuses on a relatively new EMS operation challenge, the ambulance offload delay (AOD) problem, which is a direct consequence of health care system congestion. The definition of this problem is presented in detail in Section 1.3. This research measures and quantifies this problem with a real-world case study of the EMS system in Nova Scotia, Canada. Furthermore, studies have been carried out, using several operations research (OR) methodologies, to predict the problem, develop and evaluate interventions to mitigate AOD, and to improve the performance of EMS as a public interest.

1.1 EMS Operations

To help readers understand how EMS operates, the typical events associated with an ambulance response in a Canadian EMS system are summarized in this section. The information is based on the general EMS procedures reviewed in literature [1, 6–8] and the author's observations over 50 hours of on-site training at Nova Scotia's EMS service provider. It describes typical EMS operations involved in responding to emergency/urgent calls in Nova Scotia, Canada.

When a new call is received by the EMS dispatch call centre, the call taker assesses the call (known as the call screening process). The main function is to determine the severity of the incident and its degree of urgency. Each call is then classified into a call priority and the dispatch logic applies to decide on the type and number of ambulances to dispatch to the accident scene [6]. For all but low priority calls, typical dispatch logic specifies that, the closest vehicle is sent to ensure that vehicles arrival on scene as quickly as possible. For high-priority calls, a second vehicle with advanced skilled paramedics may also be dispatched to ensure the correct level of care can be provided at the scene [1]. After an ambulance crew is given the details of the call, the vehicle starts driving to the scene. In some cases, there may be a short mobilization delay before the crew departs which is normally negligible if the ambulance is posted on the road. However, the mobilization delay can be a few minutes if, for example, the ambulance crew is resting at a station. Based on the priority of the call, ambulances may travel either with lights and sirens on or without. Vehicle travelling speeds can also be different due to the call priority. High-priority calls typically require higher speeds, while lower priority calls are responded to with standard traffic speeds. Upon arrival at the scene, paramedics assess the patient, perform first-aid care, and decide if the patient needs to be transported to a hospital. If no patient transport is required, paramedics clear the scene and become free for future service or reposition. Otherwise, the ambulance crew departs the scene and transports the patient to a hospital ED [1]. Depending on the at-scene assessment of the patient, the transport can be either at higher or normal travel speeds. Once the ambulance arrives at the hospital with the patient, the paramedics transfer the patient care to the ED staff (ambulance offload), then clean and restock the ambulance, complete patient care reports, nourishment, etc. [7]. However, when the ED is congested, this transfer of patient care (ambulance offload) is often delayed, until an ED bed becomes available (see Section 1.3 for more details). After this process is completed, the ambulance and crew become free and available for the next call.

Spaite et al. [9] categorized these events associated with an ambulance response, or "time-on-task", into standard time intervals. Later, Cone et al. [10] presented a figure adapted from that study as a summary of the time intervals of ambulance response events (Figure 1.1). It can also be viewed as a process chart of the ambulance events when responding to an emergency/urgent call. Readers can refer to this figure to further understand the EMS operating procedures.

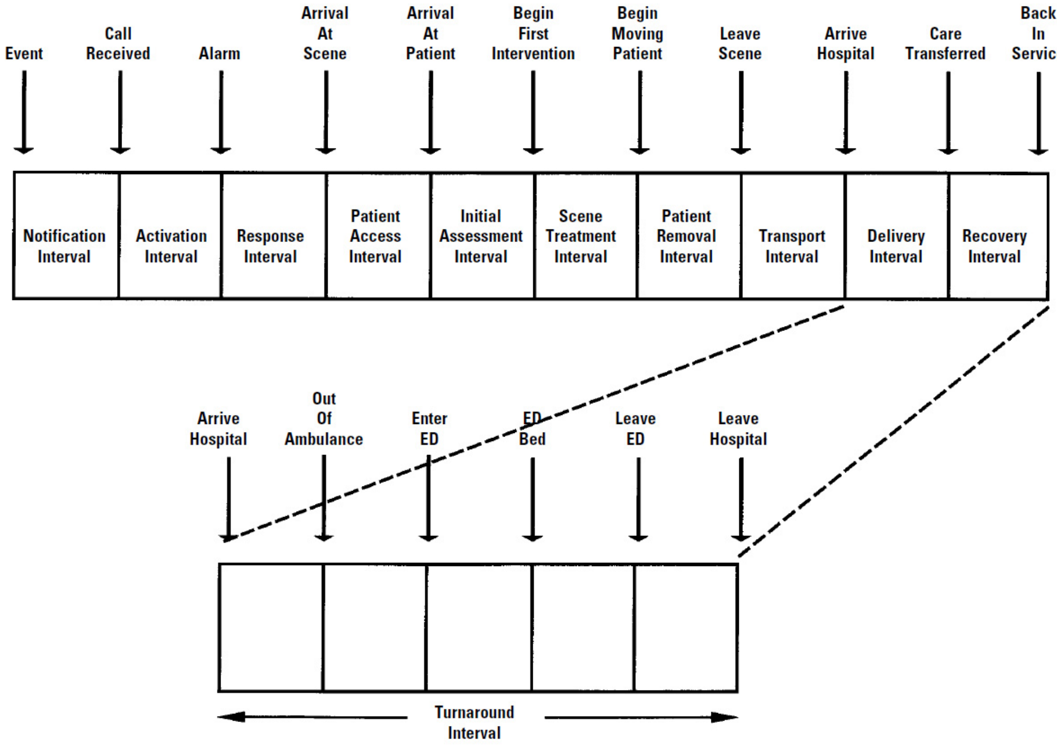


Figure 1.1: A summary of the time intervals of ambulance response events. Adapted from Cone et al. [10].

1.2 OR in EMS management

OR in healthcare operations management has been an active and popular research field. There are many problems in EMS systems that can be addressed from an OR perspective. For instance, the ambulance locations for providing maximum coverage to a given population, the ambulance/paramedic schedule for maintaining an adequate service level, etc.

Much research has been conducted into EMS systems around the world. Researchers have shown great interest in analyzing a variety of EMS processes to make suggestions for improvements in: response time, dispatch time, deployment and redeployment, etc. [11–14]. Various OR methods (such as mathematical programming, queueing theory, simulation and statistical modelling) have been applied to analyze EMS systems and contribute to the development of EMS solutions to improve resource efficiency [1, 15]. Literature reviews have been written regarding the work conducted for EMS systems using different OR methods [3, 6, 16]. Specifically, Brotcorne et al. [6] conducted a review on mathematical programming applied to ambulance location and relocation models. Fomundam and Herrmann [16] surveyed the applications and contributions of queuing theory in the field of healthcare. Aboueljinane et al. [3] focused on reviewing computer simulation models that have been used for the analysis and improvement of EMS. Readers can refer to these reviews to find models and applications for different approaches to EMS system performance improvements.

1.3 The ambulance offload problem

When an ambulance arrives at the hospital with patient(s), the paramedics transfer patient care to the ED staff, then complete patient care reports, clean and restock the ambulance before becoming available for the next call. This total time that an ambulance spends at the hospital while on call is known as the ambulance "turnaround interval" [7, 9]. It can be further separated into two sub-intervals: the “delivery interval” and the “recovery interval” [10] (Figure 1.2). The “delivery interval”, which is also known as the ambulance offload time [7], starts when the paramedics arrive at the hospital with the patient(s), and ends when the patient care is transferred from the ambulance service to the ED. The “recovery interval” starts when the patient transfer of care is finished, and ends when the ambulance crew are ready to return to service [7, 9]. Due to the increasing demand of the health care system, hospital EDs often operate at their capacities [17–19]. When a hospital ED cannot accept the incoming ambulance patient immediately (often due to congestion), paramedics wait with their patient(s), and continue to provide patient care until an ED bed becomes available and the ED personnel assume responsibility for the patient(s). This delay period in transfer of care is referred to as AOD. The ambulance crews who get delayed at the hospital are unable to return to service. The time to transfer a patient from EMS to the ED can be significant when there is no ED bed available for an extended period of time for the incoming ambulance patient [20]. The AOD problem has become a growing concern in many health care systems, thus, has attracted attentions of many health care providers and researchers [7, 20, 21].

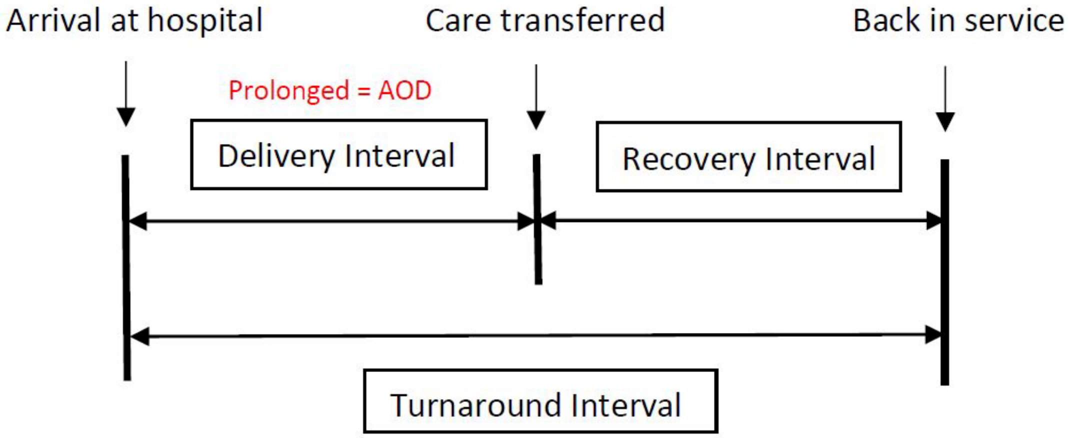


Figure 1.2: The time-interval diagram of ambulance patient transportation process.

Many researchers have suggested that AOD is caused by ED crowding [5, 22–25] and may cause substantial consequences to patients and to EMS systems [26–28]. Consequences to patients include delay to definitive care, poor pain control, delayed time to antibiotics, etc., which may compromise patient safety [20, 27, 28]. Consequences to EMS systems include negative impacts on the system status and resource availability. It can negatively affect the availability of the ambulance service to respond to the next call, prolong the response time and time spent on task, resulting in decreased efficiency of the EMS system, and the need for additional staffing [26, 28]. In addition, financial burdens to EMS systems and legal concerns regarding the AOD problem have also been reported [5, 29, 30]. A systematic review of literature that addresses the AOD problem is also conducted and published [31] as a contribution of this dissertation. Part of the review is presented as Chapter 2 in this dissertation. Readers can refer to this review to find more research on the following topics: improved understanding and assessment of the AOD problem, analysis of the root causes and impacts of the problem, and development and evaluation of interventions from both hospital and EMS system levels.

Despite EMS decision-making being a well investigated subject area for OR, the EMS interface with hospital EDs, more specifically, the AOD problem, has seen less attention in this field [7, 10, 28]. Optimization models of ambulance services generally do not address the amount of time that ambulances spend at hospitals waiting to transfer patients. However, in recent years, the AOD problem has been raised by health care providers and researchers [7, 10, 28]. There are retrospective studies with the goal of understanding and analyzing this growing issue [7, 22, 28]. However, only a few OR models have been presented on the AOD problem indicating the need for long term plans to prevent or mitigate it [5, 32, 33]. Therefore, a formal process based on scientific evidence is needed for EMS systems to determine the impacts of AOD and to design mitigation interventions to reduce it.

1.4 Research objectives

The objectives of my Ph.D. research include the following:

- to understand and measure the effects of AOD on the EMS system of Nova Scotia, Canada;

- to design and evaluate proactive EMS interventions to minimize the effects of AOD on the performance of EMS system;

- to establish understanding of the impacts of AOD in a Canadian EMS setting with combination of urban and rural regions.

To achieve these goals, this research was carried out in three phases. The first phase is to understand the AOD problem in general with a review of literature (Chapter 2); quantify the particular AOD problem presented in Nova Scotia, Canada, and measure its effects on the local EMS system (Chapter 3). The second phase is to predict AOD status of the system in real time, to allow the problem to be addressed proactively (Chapter 4). The third phase is to define EMS intervention ideas in collaboration with key personnel from the local ambulance service provider, and to develop models to test interventions that have great potentials (Chapter 5). A primary contribution of this research is to measure the efficacy of the selected intervention, and to help the local ambulance service provider determine if it can help achieve a desired system performance. This is desired by both ambulance service provider and the Nova Scotia Health Authority (NSHA), to move from the current reactive practice to a proactive, predictable response by all involved parties.

1.5 Thesis structure

The rest of the dissertation is organized as follows:

- Chapter 2 provides a systematic review on literature that addresses the AOD problem. This chapter has been published as a review paper in the journal of healthcare system management science [31].
- Chapter 3 reports an empirical analysis of the effect of AOD on the efficiency of the EMS system in Nova Scotia, Canada.
- Chapter 4 proposes a hybrid decision tree model for the prediction of the EMS system status in relation to the AOD problem.
- Chapter 5 presents an optimal ambulance destination policy developed when facing AOD by using a Markov Decision Process (MDP) model.
- Chapter 6 includes the conclusion and discussion of this research, as well as some suggestions for future work.

1.6 Summary of content

Chapter 2 reviews literature which addresses the ambulance offload delay problem. The review is organized by the following topics: improved understanding and assessment of the problem, analysis of the root causes and impacts of the problem,

Chapter Title	Approach
Chapter 2. A review on ambulance offload delay literature	
Chapter 3. An empirical analysis of the effect of ambulance offload delay on the efficiency of the ambulance system	Statistical analysis & regression model
Chapter 4. Predicting ambulance offload delay using a hybrid decision tree model	Machine learning algorithms
Chapter 5. Determining ambulance destinations when facing offload delays using a Markov decision process model	Markov decision process model

Table 1.1: Chapter scopes and approaches.

and development and evaluation of interventions. The review found that many researchers have investigated areas of emergency department crowding and ambulance diversion; however, research focused solely on the ambulance offload delay problem is limited. Of the 137 articles reviewed, 28 articles were identified which studied the causes of AOD, 14 articles studied its effects, and 89 articles studied proposed solutions (of which, 58 articles studied ambulance diversion and 31 articles studied other interventions). A common theme found throughout the reviewed articles was that this problem includes clinical, operational, and administrative perspectives, and therefore must be addressed in a system-wide manner. The most common intervention type was ambulance diversion. Yet, it yields controversial results. A number of recommendations are made with respect to future research in this area. These include conducting system-wide mitigation interventions, addressing root causes of ED crowding and access block, and providing more OR models to evaluate AOD mitigation interventions prior to implementation. In addition, measurements of AOD should be improved to assess the size and magnitude of this problem more accurately.

Chapter 2 is based on the following article:

- **M. Li**, P. Vanberkel, & A. Carter. (2018). A Review on Ambulance Offload Delay Literature. *Health Care Management Science.* https://doi.org/10.1007/s10729-018-9450-x.

In *Chapter 3*, we conduct an empirical analysis of the effects of AOD in Nova Scotia, Canada. The efficiency of the EMS system was measured using 12-months of emergency call data from the partnering ambulance service provider and local hospitals. Performance measures associated with AOD include the number of ambulances at EDs, ambulance turnaround time, total call time, response time, and ambulance availability. The results suggest that AOD occurring in the Central Region of Nova Scotia leads to negative effects on all perspectives of these performance measures in that region. It reduces the efficiency of the EMS system by prolonging the ambulance turnaround time and total call time, and increasing the response time to future calls. Furthermore, AOD has a negative impact on ambulance availability of the region experiencing it. It also shows some impacts on ambulance availability of the other adjacent regions in the same EMS system with shared resources. The results of this study offer insight into a more comprehensive understanding of the impacts of AOD on the EMS network. This approach can also be generalized to be used by other EMS providers to assess the impact of AOD on their operations.

Chapter 3 is based on the following article:

- **Mengyu Li**, Xiang Zhong, Judah Goldstein, Terence Hawco, Jan Jensen, Alix Carter, & Peter Vanberkel. An empirical analysis of the effect of ambulance offload delay on the efficiency of the ambulance system (working paper).

In *Chapter 4*, we develope a decision-support tool using a hybrid decision tree model to predict the severity of AOD occurring within 1 to 5 hours in an EMS system. The primary objective of this study is to provide a prediction model for the AOD states based on the current system status as well as hours of the day and day of the week, so that the decision makers can activate proactive interventions to mitigate AOD. Various prediction models are developed based on different prediction focuses and periods tailored to the client's needs. Furthermore, we demonstrate the value of predictive analysis to improve operational efficiency. This research demonstrates a novel hybrid decision tree method applied with administrative data. A naïve Bayes classifier was employed first to remove the noisy training observations before the decision tree induction. This hybrid decision tree algorithm was tested against the basic classification and regression tree (CART) algorithm, using classification accuracy, precision, sensitivity and specificity analysis. The results indicate that the hybrid algorithm shows improvements of performance in the classification of the real world problem. It is anticipated that the prediction model for AOD produced from this study will be directly transferable. It can be generalized to other EMS systems with a similar operational setting where ambulance offload is impacted by ED congestion.

Chapter 4 is based on the following article:

- **Mengyu Li**, Peter Vanberkel, & Xiang Zhong. Predicting ambulance offload delay using a hybrid decision tree model (working paper).

In *Chapter 5*, one of the AOD interventions is chosen from two focus group discussions with key personnel from the local ambulance service provider. We then formulate a discrete time, infinite-horizon, discounted MDP model to determine when it is advantageous to send appropriate patients to out-of-region EDs, which have longer transport times but shorter offload times. Based on the MDP model, an optimal ambulance destination policy is constructed using the policy iteration algorithm. A computational study is applied using 12-months of data from an EMS provider which experiences AOD regularly. We find that the optimal policies can significantly reduce AOD, time to bed for patients, and out-of-service time for paramedics at the expense of increased ambulances travel distances. The model can be generalized and used as a decision support tool for EMS systems to mitigate the impact of AOD on their operations.

Chapter 5 is based on the following article:

- **Mengyu Li**, Alix Carter, Judah Goldstein, Terence Hawco, Jan Jensen, & Peter Vanberkel. Determining ambulance destinations when facing offload delays using a Markov decision process model (working paper).

Since *Chapter 2* through *Chapter 5* are either published paper or working papers, some repetitions of introductory information and terminology can be expected due to the nature of the work. Some of these similar sections have been removed from later chapters of this dissertation to avoid repetition, while some of them are kept for structural purpose and the flow of chapters.

In *Chapter 6*, we conclude this research and discuss some future research directions.

Chapter 2

A REVIEW ON AMBULANCE OFFLOAD DELAY LITERATURE

2.1 Introduction

Emergency medical services (EMS) are public safety systems responsible for providing emergency assistance and for protecting public health and safety [2]. The goal of such systems is to respond quickly to population calls, to provide first aid services, and to transfer patients to the appropriate hospital when needed [3]. In life-threatening emergency situations, the ability of EMS providers to quickly respond will mean less adverse effects for the patients involved. Therefore, EMS providers continuously seek best practices, especially in a world where an aging population adds pressure to the health care system [5]. Furthermore, EMS providers are challenged to perform their services more effectively and efficiently to meet their own budgetary and performance targets.

Much research has been conducted into EMS systems around the world. Researchers have shown great interest in analyzing a variety of EMS processes to make suggestions for improvements in: response time, dispatch time, deployment and redeployment, etc. [11–14] . Various operations research (OR) methods (such as mathematical programming, queueing theory, simulation and statistical modelling) have been applied to analyze EMS systems and contribute to the development of EMS solutions, commonly through improving resource efficiency [1, 15]. Literature reviews have been written regarding the work conducted for EMS systems using different OR methods [5, 6, 30]. Specifically, Brotcorne et al. [6] conducted a review on mathematical programming applied to ambulance location and relocation models. Fomundam and Herrmann [16] surveyed the applications and contributions of queuing theory in the field of healthcare. Aboueljinane et al. [3] focused on reviewing computer simulation models that have been used for the analysis and improvement of EMS. Readers can refer to these reviews to find models and applications for different approaches to EMS system performance improvements.

Despite EMS decision-making being a well investigated subject area, the EMS interface with hospital emergency departments (EDs) has seen less attention. Optimization models of ambulance services generally do not address the amount of time that ambulances spend at hospitals waiting to transfer patients. However, in recent years, the ambulance offload delay (AOD) problem has been raised by health care providers and researchers [7, 10, 28].

When an ambulance arrives at the hospital with patient(s), the paramedics transfer patient care to the ED staff, then complete patient care reports, clean and restock the ambulance before becoming available for the next call. This total time that an ambulance spends at the hospital while on call is known as the ambulance "turnaround interval" [7, 9]. It can be further separated into two sub-intervals: the “delivery interval” and the “recovery interval” [10] (Figure 2.1). The “delivery interval”, which is also known as the ambulance offload time [7], starts when the paramedics arrive at the hospital with the patient(s), and ends when patient care is transferred to the ED staff. The “recovery interval” starts from when the patient transfer of care is finished, and ends when the ambulance and crew are ready to return to service [7, 9]. When the ED cannot accept the incoming ambulance patient immediately (often due to congestion), paramedics wait with their patient(s), and continue to provide patient care until an ED bed becomes available and the ED personnel assume responsibility for the patient(s). This delay period in transfer of care is referred to as AOD. This AOD problem is a growing concern for health care providers, as the delayed ambulance and crew are unable to return to service, and this delay can be significant [20]. Keeping EMS crews at hospital EDs can have a significant adverse impact on ambulance availability and response times for future population calls [22, 34].

The AOD problem has only recently become an active research area. There are retrospective studies with the goal of understanding and analyzing this growing issue [7, 22, 28]. There are also various analytical models on AOD indicating the need for long term plans to prevent or mitigate the problem [5, 32, 33]. However, we have not found a literature review focused on the AOD problem. The goal of this review is to analyze the literature examining the AOD problem found in journal articles,

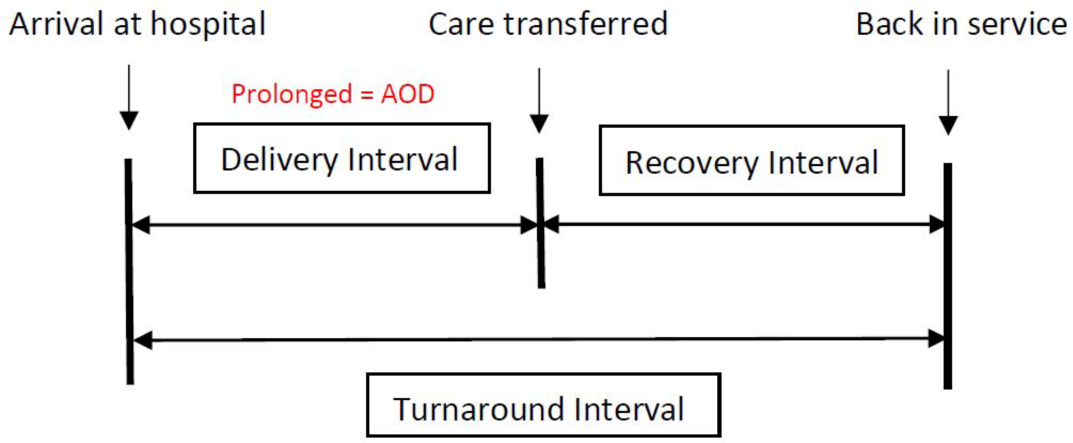


Figure 2.1: The time-interval diagram of ambulance patient transportation process.

conference proceedings, grey literature, and books that represents 30 years of work in this field. Our discussion summarizes this growing issue, the development and contributions of OR to this field, and provides a description of the novel literature for coping with AOD.

This review is organized as follows: Section 2.2 describes the search strategy for the literature and review criteria. Section 2.3 presents the current understanding of this problem, and the measures to assess and/or evaluate the impacts of AOD. Section 2.4 discusses some of the potential root causes of the AOD problem that have been reported in literature. Section 2.5 summarizes the impacts of AOD, including the consequences on patient outcomes and EMS system performance, its financial impacts, and some legal concerns. Section 2.6 reviews the current interventions that have been studied and trialed to minimize the impact of AOD and potential future implementations to improve EMS performance.

2.2 Search Strategy

We conducted a comprehensive search of the existing literature applied to the AOD problem found in journal articles, conference proceedings, grey literature, and books. We defined the scope of this review to include articles that met one or both of the following review criteria: (1) they studied the AOD problem or the interface between

EMS and hospital EDs as a primary objective, in relation to EMS operations, including measures, causes, effects, and solutions; (2) they studied interventions related to AOD or the interface of EMS and hospital EDs in the context of general EMS practices, rather than a specialty service.

The databases consulted include: PubMed MEDLINE, CINAHL Full Text, Web of Science Core Collection, and ProQuest Dissertations & Theses. A broad set of search terms in the title and abstract fields was identified by a preliminary search on related topics to encompass each facet of the review criteria. Search keywords included: ambulance offload; ambulance diversion; ambulance ramping; ambulance handover; ambulance availability; offload delay; offload time; offload zone; turnaround interval; hospital interval. All searches were conducted on May 29, 2017, with restriction to English-language publications. The searches returned 470 studies with 137 duplicates, which resulted in 333 unique articles. Articles that clearly did not meet one or more of the review criteria were not considered further. The reviewer identified 100 articles meeting the review criteria, and 37 more articles were found through reference searching. The method, focal areas, and main contributions of each paper are outlined in the electronic accompaniment in Appendix A.

2.2.1 Search Results

The searches returned 470 studies with 137 duplicates, which resulted in 333 unique articles. The author examined the results to identify potential articles of interest. Articles that did not meet any of the review criteria according to the title and abstract were not considered further. Full-text of the potential relevant articles were then reviewed, and the reviewer identified 100 articles meeting one or both of the review criteria. 37 more articles were found through reference searching of the reference lists (Figure 2.2). The method, focal areas, and main contributions of each paper (n=137) are provided in the electronic supplementary material. Readers can refer to this for a summary and a quick reference guide.

2.3 Understanding and Assessing AOD Problem

This section reviews studies that have worked to describe and quantify the size of the AOD problem in different parts of the world. We identified nine such studies

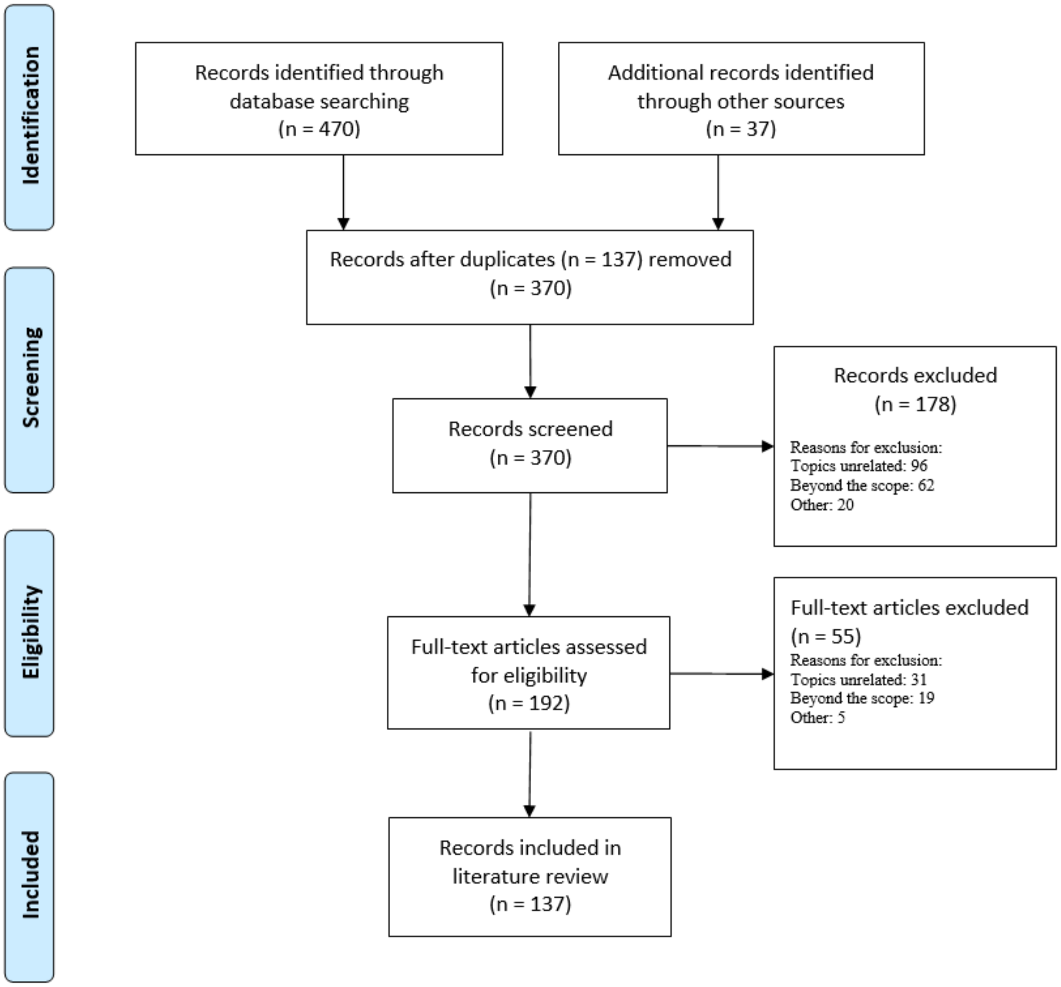

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 470)
Additional records identified through other sources (n = 37)
Records after duplicates (n = 137) removed (n = 370)
Records screened (n = 370)
Records excluded (n = 178)
Reasons for exclusion:
Topics unrelated: 96
Beyond the scope: 62
Other: 20
Full-text articles assessed for eligibility (n = 192)
Full-text articles excluded (n = 55)
Reasons for exclusion:
Topics unrelated: 31
Beyond the scope: 19
Other: 5
Records included in literature review (n = 137)

Figure 2.2: Literature search and screening flowchart.

measuring AOD in regions of North America, Europe, and Australia.

2.3.1 Empirical Assessment of AOD

Eckstein and Chan [26] analyzed a total of 21,240 incidents when the AOD occurs in Los Angeles, CA, USA between April 2001 and March 2002. Incidents were included when the ambulance turnaround time was greater than the local standard of 15 minutes. These accounted for 1 out of every 8 ambulance transports in the studied area. Among these incidents, 8.4% were in excess of 1 hour. The median waiting time per incident was reported to be 27 minutes, with an interquartile range of 20 to 40 minutes. They concluded that the decreased ambulance availability may have a significant negative impact on the EMS systems' ability to provide timely response. Their study also suggested a direct link between ED crowding and the ability of EMS to provide a timely response to future emergency calls.

In a study conducted by Segal et al. [35], the authors examined the ambulance turnaround time for 152 ambulance arrivals to a local hospital ED in Montreal, QC, Canada during a six-week period from June to August 2003. The results show that the total time ambulances spent in hospitals represents 45% of the total call time (45.24 minutes and 101.06 minutes, respectively). The majority of the turnaround time occurred after the completion of triage with a mean time of 31.33 minutes. The authors suspected that the prolonged post-triage time may be a reflection of the difficulty ambulances are having in transferring patient care to the ED.

Silvestri et al. [23] conducted an observational study to evaluate offload delay intervals and the association between out-of-hospital patient triage categorization and admission. The overall mean offload time was reported to be 32.7 minutes (among the 167 patients in the study group), including 122 green-level (least severe), 36 yellow-level (moderately severe), and 9 red-level (most severe) patients. The mean offload times for green, yellow, and red criteria were 34, 39, and 1.6 minutes, respectively. Over 52% of all patients were offloaded within 15 minutes of arrival, with an additional 16% within 30 minutes, 17% within 60 minutes, and 15% in excess of 60 minutes. The author concluded that the patient triage categorization cannot determine need for admission therefore should not be used to evaluate offload time intervals.

Cone et al. [36] reported that AOD is a relatively common problem at the interface of the EMS systems and hospital EDs in New South Wales, Australia. They conducted a retrospective study in 2009 to quantify the AOD experienced by the Ambulance Service of New South Wales, and to investigate patient and system factors associated with AOD. Of 141,381 transports, 12.5% of patients experienced an AOD of 30 – 60 minutes, and 5% a delay of $\geq$ 60 minutes. AOD was most pronounced at large hospitals, in urban areas and during winter.

2.3.2 Measurements of AOD

Many hospital EDs and EMS systems have started to treat AOD as a new performance benchmark to ensure quality patient care [5, 28, 37]. Researchers, therefore, have begun to explore different ways to help assess the AOD problem properly and accurately. Hammond et al. [38] introduced a standard definition of this process developed through in-depth interviews, focus groups and chart audits within the Queensland Ambulance Service and 10 EDs across Southeast Queensland, Australia. The study identifies significant inconsistencies in the practice and reporting of AOD across all EDs. Taylor et al. [39] conducted an observational study in Bath, UK, to determine the difference between the recorded arrival of an ambulance outside an ED and the actual delivery of the patient to the clinical area of the ED. This study demonstrates a small but significant delay between these two time records. The author recognized that this delay is inevitable, and it is difficult to see how it can be significantly reduced.

A concern was expressed by Segal et al. [35] that little data is available that directly relate AOD to specific factors (i.e., ED crowding). Cooney et al. [37, 40] assessed the AOD problem at a hospital ED in Syracuse, NY, USA, to explore if the National Emergency Department Overcrowding Scale (NEDOCS) score could be used to predict increasing AOD. NEDOCS is a performance measure (ranges between 0 and 200) implemented in most of the North American's EDs in to assess the degree of crowding (the higher, the busier). The authors studied a sample of 483 patients arriving via ambulance to the SUNY Upstate Medical University Hospital ED during a 12-month period, by recording the NEDOCS score and offload time for each patient at the time of arrival, as well as demographical information. Among these visits, AODs were ranged from 0 (no delay) to 157 minutes with a mean of 17.07 minutes. 15.5% of them were reported ≥ 30 minutes. When examining the delay time alongside the NEDOCS score groups, significant AOD time differences were reported between these groups. The authors thus concluded that the NEDOCS score had a positive correlation with AOD and could potentially be utilized by EMS personnel for determining the appropriate destination for ambulance patients to avoid crowded EDs. Later, Cooney et al. [41] conducted another study with similar data format to assess AOD at an academic level 1 trauma center with separate adult and pediatric EDs. A 12-month sample of 1,892 patients was evaluated with 21.8% pediatric ($<$ 19 years old) and 78.2% adult ($>$ 18 years old). AOD ranged from 0 to 122 minutes, with a mean of 14.01 minutes. Significant differences were found in delay time between the

NEDOCS score range groups (defined in their previous study [40]): group 1 = 9.18 minutes, group 2 = 12.72 minutes, group 3 = 18.14 minutes, group 4 = 20.62 minutes. This indicates that NEDOCS score has a positive correlation with AOD. 769 of these cases were also evaluated by using the Emergency Severity Index triage level (1 – 5). The authors reported that the mid-level severity (level 3) was associated with the longest average AOD, 11.62 minutes. There were significant differences between all five triage levels when measuring the average AOD. The authors suspected that nursing perception of patient severity may affect AOD.

According to Carter et al. [7], most EMS systems find it challenging to accurately measure the offload time (delivery interval). Instead, they measure the ambulance's total time at hospital (turnaround interval) and most AOD research and policy is based on this proxy. Therefore, this research group tested the validity of using the turnaround interval as a surrogate for the delivery interval. Their analysis showed a good correlation (0.753) between turnaround time and actual offload time. Steer et al. [41] introduced a novel method to monitor the offload time by using radio frequency identification (RFID) tags to the ambulance cots and a reader in the ED ambulance entrance. This way the ambulance traffic in ED can be passively recorded. 1,920 complete visits were recorded in this 16 weeks observational study starting December 2009. The offload time averaged at 13.2 minutes, with a median of 10.7 minutes. A total of 43% of the patients were offloaded in less than 10 minutes, while 27% took greater than 15 minutes.

The summary of these measurements of the AOD problem is shown in Table 2.1.

2.4 Causes of AOD

Emergency department crowding refers to the situation where an ED is functionally impeded due to the physical or staffing capacity shortage of the ED [42]. It has been reported by many authors as an important contributor to AOD [5, 21–25], thereby a major concern to EMS providers, as the negative effects are substantial [43–45]. Due to the increasing volume of patients, ED staff can no longer prioritize the quick turnaround of ambulances. This creates risks for delayed EMS responses to future population calls [22].

Paper	AOD Measure	Data stratifications	Study Region
Eckstein & Chan (2004)	Median: 27 minutes	N/A	Los Angeles, CA, USA
Segal et al. (2006)	Mean: 31.3 minutes	N/A	Montreal,QC, Canada
Silvestri et al. (2006)	Mean: 32.7 minutes	By severity	Orlando, FL, USA
Cooney et al. (2011; 2013a)	Mean: 17.07 minutes	By NEDOCS score	Syracuse, NY, USA
Cooney et al. (2013b)	Mean: 14.01 minutes	By severity and NEDOCS score	Syracuse, NY, USA
Steer et al. (2016)	Mean: 13.2 minutes, Median: 10.7 minutes	N/A	Akron, OH, USA
Cone et al. (2012)	12.5% of patients:30 − 60 minutes;5% of patients:≥ 60 minutes	By lengths of AOD	New South Wales, Australia
Taylor et al. (2006)	Measured the difference between the recorded arrival of an ambulance and the actual delivery of the patient to the clinical area of the ED	N/A	Bath, UK
Hammond et al. (2009)	Held interviews to define AOD	N/A	Southeast Queensland, Australia
Carter et al. (2014)	Calculated the correlation (0.753) between ambulance total time at hospital and AOD time	N/A	Richmond, VA, USA

Table 2.1: The summary of articles that measure the AOD problem.

Other observational and analytical studies have supported this conclusion. An investigation conducted for the Ministry of Health and Long-Term Care in ON, Canada [46] reported that the principal cause of AOD is the congestion in downstream stages of patient care (i.e., hospital bed shortage). Eckstein and Chan [26] suggested that ED crowding results in delays for paramedics waiting to transfer patients (AOD). Majedi [32] expressed concern that the delayed transfer of an admitted patient from the ED to an inpatient bed contributes to ED crowding, and subsequently the AOD problem. Eckstein et al. [22] and Almehdawe et al. [5] both suggested that this escalating problem of extremely high inpatient occupancies (capacity shortage) has resulted in ED crowding, the AOD problem, and eventually a reduction in the quality of EMS service to the community.

ED crowding is an increasingly common issue faced by many health care systems

[17–19]. In Andrulis et al.'s survey [47] on crowding in 239 American teaching hospitals, three quarters of responding hospitals reported holding times increased for admitted patients over the preceding three years from 1991, and the use of methods to decrease crowding was also growing. The increase in ED crowding has also be reported in published articles since this survey [27, 48, 49].

The causes of ED crowding are complex and multifaceted [28]. Many researchers have investigated this area to identify the contributing factors and strategies to reduce ED crowding. Derlet et al. [48] distributed a survey to EDs in 50 American states to determine the factors associated with ED crowding as perceived by ED directors. Among the 575 responded EDs, 91% reported ED crowding as a problem, and 33% reported that some patients had poor outcomes as a result of it. Their study summarized some common causes of ED crowding reported by the ED directors, including high patient acuity, hospital bed shortage, high ED patient volume, radiology and lab delays, and insufficient ED space. Some other factors contributing to ED crowding were outlined by Derlet and Richards [43, 50], and Olshaker & Rathlev [51]. Such factors included shortage of support staff, consultation delays, shortage of on-call specialists, ED space limitations, language and cultural barriers, increased medical record documentation requirements, and difficulty in arranging follow-up care. Figure 2.3 shows a summary diagram of the common causes of ED crowding, which leads to AOD. Readers can also refer to Hoot & Aronsky's review [52] to find more research regarding causes, effects, and solutions of ED crowding.

While multiple factors are likely contributors to the growing crisis of ED crowding, recent research suggests that ED crowding is not caused by the input factors (i.e., nonemergency ED patient visits), but rather by the output factors (i.e., the overall hospital throughput) [42, 53, 54]. Access block has been identified as a major cause to ED crowding [28, 42, 48, 55]. It refers to the situation where patients in the ED requiring inpatient care are unable to gain access to appropriate hospital beds due to a lack of available inpatient beds. In this circumstance, admitted patients remain in the ED until a hospital bed becomes available. This access block period can last from hours to days [18], limiting the patient's evaluation/treatment and causing ED crowding [50]. Schneider et al. [49] evaluated multiple trialed strategies to reduce ED crowding in Rochester, NY, USA in the last decade. They realized

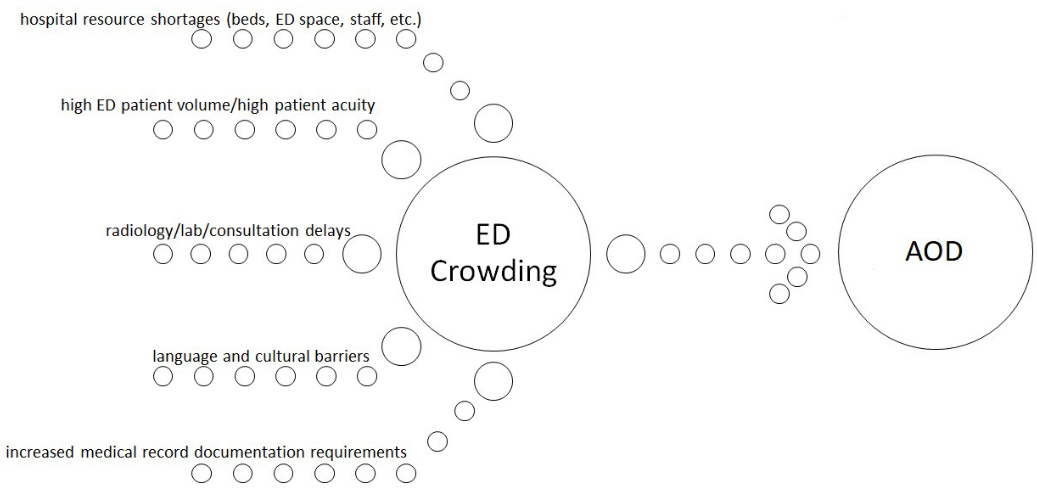

hospital resource shortages (beds, ED space, staff, etc.)
high ED patient volume/high patient acuity
radiology/lab/consultation delays
language and cultural barriers
increased medical record documentation requirements
ED Crowding
AOD

Figure 2.3: A summary diagram of the common causes of ED crowding, which leads to AOD.

that those strategies based from the ED were the ones with little effect; while the ones addressed factors external to the ED were more successful. Other researchers [42, 56, 57] supported this conclusion with a recommendation of finding the solutions in managing hospital bed stock and systemic patient capacity, including the use of primary care and community resources.

2.5 AOD Consequences

As stated by Cooney et al. [28], consequences of AOD can be categorized into two major headings: consequences to the patient and consequences to the EMS system. Consequences to patients include delay to definitive care, poor pain control, delayed time to treatment, etc., which may result in compromising patient care and safety. Consequences to the EMS system are negative impacts on the system status and resource availability. It may prolong the ambulance response time and time spent on task, resulting in decreased efficiency of the ambulance services, and the need for additional staffing [26, 58]. In addition, financial burdens and legal concerns regarding the AOD problem have also been reported [5, 29, 30].

2.5.1 AOD Impact on Patients

The time required to transfer patient care can be critical to ambulance patients upon arrival at the hospital. Any delay in this process (e.g., AOD) is a potential risk to patient safety [28]. Crilly et al. [20] conducted a study to describe and compare outcomes for ambulance patients arriving to EDs who experienced delays longer than 30 minutes with those who did not. This study was undertaken in Australia using 12 months of health data (September 2007-2008) from 40,783 patient visits to three EDs via ambulance. These visits made up about 30% of the total ED visits. Among these ambulance visits, 15% experienced an AOD longer than 30 minutes, and 63% of those had an ED length of stay (LOS) longer than 4 hours. This study confirmed that transport by ambulance to hospital does not guarantee timely access to medical care when there is AOD. The authors also reported that patients with an AOD shorter than 30 minutes had significantly better outcomes for almost all demographic and ED characteristics (i.e., time to triage, ED LOS) with the exception of in-hospital mortality. Similar conclusions were reported by Hitchcock et al. [27]. Their study was conducted to describe and compare patient outcomes between ambulance patients arriving to one ED in Australia (1 June - 31 August 2007) with (619 cases) and without (1,238 cases) experiencing AOD. The cases in the two groups were matched by age, gender, and presenting problem. Outcome measures included ED LOS and in-hospital mortality. The results indicated that patients who experienced AOD had significantly longer wait time to be triaged (10 minutes vs. 4 minutes), and comprised significantly higher proportions of those access blocked (43% vs. 34%). This study also reveals that the likelihood of having an ED LOS longer than 8 hours is 34% higher among patients who experienced an AOD. AOD is a contributing factor to prolonged ED LOS and adds additional strain on EDs. However, there was no significant difference identified in this study on the proportion of in-hospital mortality (2% vs. 3%) between the two patient groups, consistent with the previously discussed findings.

Kingswell et al. [59] also investigated the AOD experience from the perspective of patients. They carried out semi-structured interviews with seven patients who visited a regional ED in Queensland region, Australia via ambulance and experienced an AOD longer than 30 minutes. Most participants reported not understanding the causes of AOD, but understood some of the consequences. Though they felt safe waiting with paramedics, they expressed frustration with being kept 'in the dark' during AOD, due to the lack of communication regarding the availability of ED beds. This study provided in-depth patients experiences of AOD and indicated that improvements in communication with patients are required within the context of patient rights, health care safety and quality frameworks, to ensure quality care is delivered during AOD.

2.5.2 AOD Impact on EMS Resource Availability

AOD not only hinders the promptness of medical treatments for the patients, but also negatively affects the ability of EMS to provide consistent and timely care, due to the reduced ambulance availability [22, 60, 61]. It can affect response times and prolong time on task, resulting in decreased efficiency and the need for additional resources [26]. When ambulances are unavailable for future population calls due to AOD, there is potential to put the community and lives at risk due to the compromised availability of ambulance services [28].

The impact of AOD on EMS resource availability has seen less attention. Most research has been carried out by medical doctors and frontline personnel who try to understand the problem and highlight its importance and implications using observational studies. One of the early studies was reported by Cone et al. [10]. The group conducted a "time-motion prospective study" of the EMS turnaround interval by monitoring and recording the ambulance delivery and recovery activities (122 patients). They concluded that ambulance call report documentation required the greatest sub-interval of turnaround time in the observed system. AOD was not reported as a major concern in the study. However, in a later prospective longitudinal study conducted by Eckstein and Chan [26], the authors concluded that the decrease in ambulance availability may have a significant effect on an EMS systems' ability to provide timely response. Cooney et al. [40] conducted an observational study of a sample of 483 patients arriving via ambulance during a 12-month period to explore the relation between AOD and ED crowding. They reported that the median AOD time was significant and raised concerns related to patient care and EMS system resource availability.

In their position statement to the Canadian National Association of EMS Physicians, Cooney et al. [28] raised another concern regarding the impact of AOD on EMS resource availability, through an EMS operation practice called “mutual aid”. Mutual aid represents the EMS practice where free ambulances are drawn from outlying areas into another service area to assist with AOD and to maintain proper coverage in the problematic service area. This practice may result in ambulances being relocated away from their home service areas, possibly for the duration of their remaining shifts, and represents a potential decrease in surge capacity of the EMS system. Majedi [32] expressed a similar concern in his thesis that the mutual aid practice may result in ambulance shortage in the outlying areas, which put the communities at risk.

2.5.3 AOD Impact on Finance

It has also been reported that AOD adds costs to EMS providers. Majedi [32] argued that ambulance crews are likely to work overtime when AOD occurs, which can be costly. In 2006, the city of Toronto, ON, Canada spent $3,906,700 in EMS staff overtime expenditures alone [32]. The statistics provided by the Region of Waterloo Public Health (2007), ON, Canada revealed that the Waterloo region lost 13.25 ambulance days per month to AOD in 2005 and 12.36 ambulance days per month in 2006. That translated to a financial loss of approximately $840,000 in ambulance operations. To reduce the AOD time, the provincial government invested $96 million in its comprehensive action plan in 2006. However, AOD still costed the Toronto EMS approximately 180 ambulance hours per day in December 2007 [5]. Another province of Canada, Nova Scotia, is also experiencing the worsening AOD problem and its fiscal burden [30]. The EMS provider in Nova Scotia has estimated that the AOD problem results in about 2,900 ambulance hours per year, which equates to approximately $754,000 at the average paramedic salary. Smith [4] has reported that in England, AOD costs the National Health Service millions of pounds per year in the form of lost ambulance hours, which have risen from 37,000 hours in 2008/2009 to around 54,000 hours in 2010/2011.

2.5.4 AOD Legal Concerns

Some legal concerns are also rising for the AOD problem. A major regulatory issue is that paramedics cannot assume the role of ED staff [22, 29, 62]. In the USA, it is a federal regulatory expectation that "all EDs must have policies and procedures in place to immediately receive and assume care of the patient $\cdots$ A hospital's refusal to accept responsibility could be a violation of the Emergency Medical Treatment and Labor Act (EMTALA)". Delaying care of a patient could also be a violation of EMTALA [24]. Although EMTALA gives some clarity as to whom is responsible for the patient on the stretcher once arriving at the hospital, this issue has not been addressed by legislation nor tested in case law in Canada [62].

There has been some discussions and considerations related to paramedics' responsibility for patient care in EDs. Eckstein et al. [22] rationalized that paramedics should assist the ED staff to monitor their patients under ongoing disaster conditions (outstripped resources) of the ED. However, they also acknowledged that such behaviors may have detrimental impact on the EMS system if occurring on a regular basis. Schwartz [62] raised his concern as paramedics may not been trained to treat protracted conditions, considering that their primary goal is to provide initial patient care with limited resources during the patient transportation to the hospital. Furthermore, leaving patients with paramedics in EDs may offer a false sense of security to hospital staff, as the patients are not monitored at an ED level rather than that within the paramedic scope and skill set. This could lead to delayed detection of life-threatening conditions, as well as a debate of legal responsibility and liability for care within a hospital facility, in which only credentialed physicians are permitted to practice.

2.6 Interventions for the AOD Problem

Various interventions have been proposed, trialed, and evaluated to study their effects on reducing AOD, most target either EMS providers or hospital EDs. The following section reviews these interventions, and is divided into two major categories: hospital-based interventions and EMS-based interventions. OR models that are deployed in these intervention studies are summarized in Figure 2.4 at the end of this section.

2.6.1 Hospital Based Interventions on AOD

Various interventions have been proposed, trialed, and evaluated to study their effects on reducing AOD, most target either EMS providers or hospital EDs. The following section reviews these interventions, and is divided into two major categories: hospital-based interventions and EMS-based interventions. OR models that are deployed in these intervention studies are summarized in Table 2 at the end of this section.

Offload Programs

Two urban hospital EDs (the Queen Elizabeth II Health Science Centre and the Dartmouth General Hospital) in Nova Scotia, Canada have attempted to reduce AOD time by implementing an offload zone (OZ) concept, in collaboration with the local EMS provider [63]. An OZ is a monitored holding area in the hospital ED for patients who arrive by ambulance but cannot be admitted into the ED due to congestion. This practice frees the ambulance to return to service; while the patient is in the care of a dedicated nurse and paramedic waiting for an available ED bed [25]. With these two staff, the OZ can serve multiple patients (up to 6) at the same time, eliminating the need for one ambulance to wait with each patient. Two years after opening the two OZs, Carter et al. [30] completed a Health Care Failure Mode and Effect Analysis (HFMEA) study to identify risks to patient safety and process efficiency. They created a process map to provide a framework consisting of six major processes of the OZ, for understanding its function. They concluded that the OZ resulted in ED staff having little incentive to admit patients who were waiting in the OZ and instead admitted patients from the waiting room. This led to the OZ often being at capacity and unable to relieve AOD.

Motivated by this unexpected finding, Laan et al. [64] modeled the OZ using a continuous time Markov chain to investigate how this lack of incentive impacts AOD. The result suggested that, when the probability of "a patient admitted from the OZ when a patient of equal acuteness is waiting in the waiting room" is not greater than a certain threshold (0.35 in their case), implementing an OZ will result in even longer offload delay, as admission priority is disproportionately given to patients in the waiting room. This threshold is sensitive to the capacity of the OZ and the clinical load, meaning the ED's incentive to admit patients from the OZ has a smaller impact on AOD when there is a large OZ and when the ED is less busy. Therefore, certain OZ patient selection criteria need to be enforced to maintain the expected benefits of implementing the OZ – to reduce AOD.

The Ministry of Health and Long-Term Care in Ontario, Canada has funded a project involving hiring dedicated offload nurses to monitor low acuity ambulance patients while they wait for an available ED bed [65]. Over 10 performance measures related to the offload nurse program and AOD were collected and reported to the hospitals and the EMS providers bi-annually. These measures track offload bed utilization rate, as well as the LOS with the offload nurse, which allow the EMS provider and the EDs to monitor patient flow as a predictor of AOD. The result from this trial was unclear. Clarey et al. [66] designed a discrete event simulation model to assess the change on AOD in a scenario, where dedicated nurses were hired to assist with ambulance offloading patients. This study demonstrated a clear reduction in AOD when dedicated nursing levels were increased. However, the authors also raised their concern that using this as a sole method to reduce AOD would require unacceptably low staff utilization, which would cost hospitals both financially and in human resourcing. Job duties for these dedicated nurses need to be carefully designed so that additional work can be incorporated into their work yet still enable them to rapidly react to ambulance arrivals. From the perspective of patient and health services outcomes, Greaves et al. [67] investigated patient's waiting time to see a clinician in all ED visits ($n = 21,454$) 39 days before, during, and after an offload nurse was introduced in an Australian ED during July and November 2012. They concluded that the waiting time improved marginally during the trial period, but was not sustained when the role was removed.

Expanding ED Capacity

Expanding ED capacity has been explored multiple times by different research groups using different methods, yet fielded controversial results. Silvestri et al. [24] performed a 22-month longitudinal observational study between January 2003 and October 2004. The goal was to examine the impact of ED bed availability on AOD time in a regional EMS system with four receiving hospitals in Orlando, FL, USA. Two of these hospitals remained unchanged during the study period, while the other two hospitals implemented two different AOD mitigating strategies starting in 2004. One hospital introduced an offload time limit policy, and the other one expanded the ED capacity. The median offload time was then reported decreased in all hospitals collectively from 39.6 (in 2003) to 35.1 minutes (in 2004). The result suggests that an increase in the ED bed availability decreased AOD.

Majedi [32] modeled the interaction of an EMS and a hospital ED using queuing theory, and modeled the behavior of the system as a continuous time Markov chain. The tested scenarios included adding more ED beds, adding more ambulances, and reducing the ED LOS of patients. By evaluating various performance measures (such as the average number of ambulances in offload delay, the average AOD, and ambulance and ED bed utilization), Majedi concludes that adding more beds to the ED could have a positive impact on these performance measures. In particular, the average number of ambulances experiencing offload delay and the average AOD were decreased.

Almehdawe et al. [5] used a Markov chain queueing model to analyze the interface between a regional EMS provider and multiple EDs serving both ambulances and walk-in patients. By using matrix-analytic methods, they solved for the steady state probability distributions of queue lengths and waiting times for both ambulance and walk-in patients in all the studied EDs (AOD was measured using the waiting times of ambulance patients). They computed a variety of performance measures subject to different resource levels, particularly for assessing the AOD problem and its impact on the system resources. This study concludes that the priority based admitting policy has a great impact on patient waiting times. Assigning a higher priority to ambulance patients ensure minimal AOD at the cost of long waiting times for walk-in patients. When additional resources are considered for the system, the benefit of adding capacity is greater for EDs with higher utilization. The authors propose that this model can be used to assess the effect of adding more capacity to the system. It can also show where to add resources to improve the system performance the most.

Some other studies, on the contrary, concluded that expanding ED capacity does not show any improvement on mitigating the AOD problem. Han et al. [68] examined the effects of ED expansion (from 28 to 53 licensed beds) on a metric of EMS performance at an urban, academic Level 1 trauma center in Nashville, TN, USA.

Data was compared with a five-month pre-expansion period (November 1, 2004, to March 1, 2005) and a five-month post-expansion period (June 1, 2005, to October 31, 2005). An accelerated failure time model was performed to test if ED expansion was associated with a better EMS performance while adjusting for potential confounders. The study concludes that an increase in ED bed capacity did not affect the specific EMS performance. Therefore, ED expansion appears to be an insufficient solution without addressing other bottlenecks in the hospital.

Crilly et al. [69] investigated the impact of opening a new ED on patient and healthcare service outcomes using a 24-month deterministically linked data set from the ambulance service and three ED and hospital admission databases in Queensland, Australia. Total volume of ED visits was reported to increase 18%, while local population increased 3%. Healthcare service and patient outcomes at the two pre-existing hospitals (including ambulance offload time, ED LOS, and access block) did not improve. They concluded that the increase in the total volume of ED visits was at a far greater rate than local population growth, suggesting it either provided an unmet need or a shifting of activity from one sector to another. There was an inherent need to take a "whole of health service area" approach to solve crowding issues.

Later, Crilly et al. [70] conducted a retrospective comparative cohort study to identify predictors of admission and to describe outcomes for ambulance patients at three Australian public EDs, before and after the opening of 41 additional ED beds (from 81 to 122). Reported data included: AOD, time to see doctor, ED LOS, admission requirement, access block, hospital LOS, and in-hospital mortality. The authors reported that after the increase of emergency capacity, in-hospital mortality was the only outcome measure that improved during the study period; while all other time-related service outcomes, including ED LOS, time to see doctor, and AOD, did not show any improvement.

Increasing ED Patient Throughput

As previously discussed, ED crowding and access block is a widespread problem and often results in AOD. Actions to address ED crowding and Access Block include continuously monitoring ED patient throughput times, identifying any correctable areas of delay, and implementing effective triage and bed utilization strategies (i.e., the use of fast track, acute care clinics, observational units) [22].

Majedi [32] showed that reducing patients' ED LOS, which increases the ED patient throughput, can have a positive impact on EMS system performance, including the average number of ambulances in offload delay, average AOD, and ambulance utilization. Lee et al. [71] applied a high-turnover utility bed intervention at the ED of an urban tertiary hospital in Taipei, China to improve ED patient throughput and alleviate ED crowding. 14 utility beds were designated exclusively for ED patients with a strict 48-hour LOS limit for each patient. In the pre- and post- intervention period cohort study, the authors reported improved EMS performance and a shortened ED LOS from 9.7 hours to 8.0 hours. Furthermore, there was no difference in ED revisit within 72 hours and cardiac arrest management, when assessing the impact of this intervention on the patient outcomes.

Alberta Health Services [72] in Canada implemented a province-wide ED Overcapacity Protocol (OCP) in December 2010 to battle the growing ED crowding and AOD problems in the province. This OCP sets triggers such as: ED bed occupancy $> 110\%$, $\geq 35\%$ of ED care spaces blocked, no ED space available for high severity patients, etc. When these triggers were reached, immediate actions were executed by the varied ED staff (ED physicians, nurses, clerks, etc.) to reduce ED wait times and to improve the ability to move admitted patients out of EDs. These actions are on an urgent basis and can be escalated up to the CEO level, if impact on wait times is not timely. The OCP frees up ED care spaces by increasing patient throughput. Patients might be asked to share a room, to move to a different room or facility, to receive ongoing care in the community, or to be admitted to a hospital unit and given a stretcher or chair in a temporary location. A pre-/post- OCP comparison study was conducted by McRae et al. [73] using administrative data from February to October 2010 as the pre-OCP period and the data from February to October 2011 as the post-OCP period. The ED volume was increased by 7.0% while the ambulance service demand increased by 11.1% between the pre- and post-OCP periods. The authors reported that improvements in ED patient flow led to improvements in ambulance offload time. Preliminary evaluation on the mean AOD suggested a significant reduction before and after the implementation of OCP. Cooney et al. [28] have also emphasized the importance of improving patient throughput. They argued that decreasing AOD directly, without improving throughput, does not address the issue of ED crowding. Therefore, all components of the healthcare system must work together to improve throughput on all levels to ultimately result in decreases in AOD.

These hospital interventions take different approaches to tackle the AOD. The offload nurse program assigns dedicated hospital personnel to directly work on reducing AOD; while increasing ED capacity explores the possibility of reducing AOD with additional ED facility resources. The offload zone trial takes a cooperative approach to reduce AOD by bringing together the hospital EDs and the local EMS provider; while increasing ED patient throughput requires collaboration between the ED and other hospital departments. A common thread shared between these hospital-based interventions is that they require additional ED resources, either human resources (offload nurse program), facility resources (expand ED capacity), or even a combination of both (offload zone, increase ED patient throughput). Eckstein et al. [22] recommended that every hospital should create a system to provide rapid access to additional ED resources (i.e., stretchers) when needed as well as a written plan to address ED crowding. Such contingency plans are designed to release EMS personnel rapidly from hospitals, especially in a disaster situation, when resources are scarce. Eckstein et al. [22] also recommended that EDs should apply a mandatory nurse–patient ratio (the minimum staffing ratios for good patient care in critical care areas of the hospital) as an indicator of the ED status. Furthermore, hospital administrators should emphasize the importance of enabling paramedics to transfer care of patients to ED with minimal delay.

2.6.2 EMS Interventions on AOD

EMS systems have put forth efforts to minimize the impact of AOD through several different interventions. Some of them have been trialed and reported in literatures. Others only appear in gray literatures and work reports from stakeholders. In general, the effects of these innovative practices are not well studied, with the exception of ambulance diversion.

Ambulance Diversion

Ambulance diversion (AD), first described by Lagoe & Jastremski [74], is the practice where an ED diverts incoming ambulance patients to other facilities due to overcrowding [75]. This gives the ED staff time to recover and decreases the risk of adverse events occurring in overcrowded situations [43]. To reduce the growing problem of ED crowding, many hospitals and health care systems have implemented AD policies [22, 76]. Burt et al. [77] used the 2003 National Hospital Ambulatory Medical Care Survey (on 40,253 visits to 405 participating EDs) data to determine the frequency of AD. They reported that about 45% of EDs reported diverting ambulances at some point during the previous year. Among this 45%, approximately 3% of operating time was spent in diversion status. In 2003, an estimated 501,000 diversions occurred, equivalent to one per minute.

Research has been conducted to further study and evaluate AD. Warden et al. [78] investigated the potential predictive factors of AD. Kuruvilla [79] developed various causal models to determine the probability of a hospital going on diversion. Leegon et al. [79] evaluated the accuracy of using a Gaussian Process to predict AD. Hagtvedt et al. [80] used several tools, including a birth-death process, discrete event simulations, agent-based simulation model, and some game theory to examine the potential for cooperative strategies to reduce ambulance diversion. Ramirez-Nafarrate et al. [81, 82] explored optimal AD control policies using different methods, including a simulation-optimization approach [81] and a Markov Decision Process (MDP) formulation [82]. Lin et al. [83] developed a simulation model to quantitatively evaluate the effectiveness of various ambulance diversion strategies on relieving ED overcrowding by assessing the crowdedness index, the patient waiting time for service, and the percentage of adverse patients. The same research group (Kao et al. [84]) also utilized a patient flow queuing model for simulating AD among multiple EDs in a region to evaluate the impact of different AD strategies on the crowdedness of the EDs.

While appealing in theory, AD has yielded conflicting results, and the growing issue of ED crowding has brought this strategy into question. Scheulen et al. [85] investigated the impact of AD policies in urban, suburban, and rural areas of central Maryland, USA. They found that AD policy had a limited effect in preventing further patient volume in urban and suburban areas, and it had no impact in rural areas.

Therefore, the authors argued that “the impact and efficacy of AD policies should be evaluated to ensure they are having the intended effect”. Carter and Grierson [86] researched the impact of AD on the availability of ambulance resources, specifically transport time, hospital turnaround, and total out-of-service time. 1,563 instances of diversion and 1,403 controls were included in this study, showing an average 2-minute difference in turnaround time and no difference in transport, hospital turnaround, and total out-of-service times between diversion and control time periods. Therefore, it was concluded that the availability of EMS resources was maintained during the AD periods.

Numerous studies have also suggested a variety of problems that may be caused by AD, such as: delaying prompt and appropriate medical care for diverted patients [87–89], adversely affecting EMS system efficiency [90, 91], exacerbating crowding at other facilities [28, 92, 93], and generating financial burdens to hospitals and EMS systems [94–96] . Eckstein et al. [22] also argued that use of this temporizing methodology has created false expectations of relief and often results in adversarial relations between the two key groups - the EMS and ED staff, which may put the EMS system at risk of liability. Weaver [97] reported that AD has become less effective and more problematic with hospitals everywhere filling to capacity. In addition, AD may result in legal problems [97, 98], as well as ethical and logistical ones [97, 99–101]. Hence, decisions regarding AD should be made with careful consideration of patient preferences, local EMS laws, and institutional surge capacity. The American College of Emergency Physicians (ACEP) [102] has developed some guidelines for AD to ensure access to emergency care and suggested that “each EMS system, including all of its component agencies, must develop a cooperative diversion policy” and should only allow AD to occur “after the hospital has exhausted all internal mechanisms to avert a diversion”.

Due to the controversial results, many health care systems have adopted policies to limit or eliminate AD [92, 103, 104] and studies have been carried out to evaluate these policies. One example of such implementation was detailed by Patel et al. [105] in their study undertaken in 17 hospitals of the greater Sacramento region, CA, USA from January 2001 to December 2003. After successful implementation of a comprehensive reduction program, AD in the Sacramento region was reduced by 1,428 hours per month (a 74% reduction). Furthermore, such reduction occurred despite overall increases in ED census, hospital admissions from the ED, EMS arrivals to the ED, inpatient hospital census, and overall population. In the follow-up program, Patel and Vinson [106] sought to further reduce and eliminate AD by progressively decreasing the duration of each AD event from 3 to 1 hour. This decreased AD from 8,469 hours to 2,306 hours in approximately 3 years. The author suggested that with a collaborative and cooperative goal, urban regions can effectively reduce AD by "systematically and sequentially" limiting the duration of each AD event, as demonstrated in the greater Sacramento region. Another study by Friedman et al. [107] has drawn similar conclusions after a two-week moratorium on citywide diversion in October 2006 in a consortium of teaching hospitals in Boston, MA, USA.

Lagoe et al. [108] conducted a retrospective review on AOD procedures at the system and hospital levels in the metropolitan area of Syracuse, NY, USA, reporting a 33.6% reduction on diversion hours system-wide during the study period. They concluded that a combination of approaches at the community-wide and hospital-specific levels produced meaningful reductions of AD. Barthell et al. [109] used a collaborative approach to track and report AD and ED crowding in Milwaukee, WI, USA, and reported a reduction of AD after implementing this approach. Castillo et al. [110] described a state-wide initiative to reduce diversion in four regions of California, USA from September 2006 through August 2008. Hospitals developed and implemented several best practices to improve patients' input, throughput, and output during the study period, resulting in a significant AD decrease from an average of 1,468 hours to 1,176 hours monthly.

Vilke et al. [92] evaluated a voluntary community-wide intervention to reduce AD in a county of 2.8 million individuals in California, USA. This intervention consists three core rules, as detailed in a later report [111]: AD status is limited to a maximum one-hour duration; an ED must accept at least one patient after coming off and before declaring back on diversion; regardless of diversion status, hospitals must accept patients originally discharged from their facility. A significant decrease was reported in the number of patients who did not reach the requested facility due to AD for the trial period ($n = 322$) and post-trial period ($n = 449$), compared to the pre-trial period ($n = 1,320$). In the follow up study three years later, Vilke et al. [112] reported that this voluntary community-wide approach to attempt to decrease

AD was effective and sustainable with minimal intervention. Similar conclusion was drawn by Al Darrab et al. [113] after evaluating the impact of a city-wide voluntary intervention to reduce AD in Hamilton, ON, Canada. Massachusetts became the first state in the USA to successfully ban AD after implementing a statewide ban on AD initiated by the Massachusetts Department of Public Health in USA on January 1, 2009 [114]. The results were analyzed by multiple research groups [104, 115–117]. Lindstrom [104] argued that no adverse effect was found from stopping AD; therefore, hospitals should be forced to implement some improvements of protocols and streamline operations to eliminate AD. Similar results have been reported by Holley [118] when evaluating the no-ambulance-diversion policy adopted in the city of Memphis, TN, USA.

In addition to evaluating AD related policies, researchers have developed different strategies and methods to help health care decision makers avoid situations where AD is inevitable. Strear et al. [119] applied the theory of constraints to patient care workflow and achieved a 99.6% reduction of the AD time during a 12-month implementation period. McLeod et al. [120] reported the effects of a regional information dashboard on ED capacity, which took real-time information from all three tertiary EDs in the city of Calgary, AB, Canada and assigned a color code (green, yellow, orange, or red) to reflect receiving status for each individual ED. Central dispatch had the status of all three EDs and ambulances were advised to avoid the most overcrowded ED. The authors concluded that the implementation of this real-time surveillance system resulted in an increase in the proportion of total time region hospitals reported favorable status (green/yellow) (57.5% vs. 64.1%), while the AD fell from 198 to 27 hours. El-Masri and Saddik [121] proposed a new comprehensive emergency system to facilitate the communication process in emergency cases from ambulance dispatch to the transfer of patient's care to the ED staff. Such a system enhances communication in the clinical handover process, and contributes to reducing ED crowding and AD. Beechner [122] constructed a fuzzy inference system that performs as a decision support system to eliminate AD by diverting a certain percentage of lower acuity patients to outpatient clinics or primary care physicians.

The reduction of AD may have a negative effect on the EMS system, resulting in longer AOD time, when other factors causing ED crowding are not probably addressed or corrected [28]. Asamoah et al. [123] employed a strict limitation policy to reduce AD (restricting each hospital to 1 hour out of every 8) and reported an 82% reduction in AD. However, they also observed a side effect of the new policy on the system, as the mean AOD time increased by 32%. Pham et al. [124] reported that AD may be reduced by "adding more facility and human resources (usually at the hospital level)", which reduces ED crowding. Using AD as a surrogate marker for ED crowding, Schull et al. [53] demonstrated that AD time increased by 6.2 minutes per admitted patient boarded in the ED due to ED crowding.

Efforts to reduce AD are common (and mostly successful), but the question remains of how to best reduce AD without increasing ED crowding or worsening AOD [125]. Cooney et al. [28] emphasized that monitoring AD and AOD are important to health care systems, as they are both essential indicators for assessing ED status and identifying inefficiency in the system. Although AD may temporarily release pressure on ED staff; growing AOD may put it back to hospital personal to address root causes of ED crowding.

Patient Allocation Policy

As excess AD may cause negative impact on quality of patient care and EMS operation, a few studies have suggested other alternative patient allocation policies to alleviate ED crowding in a more controlled and centralized manner than AD.

Shah et al. [126] implemented a voluntary, physician-directed ambulance destination control program in Rochester, NY, USA (during July 2003) to directs ambulances to the ED that is most able to provide appropriate and timely care. EMS providers were asked to call a destination-control physician for patients requesting transport to either of the two participating hospitals. The physician determined the optimal patient destination by using patient and system variables as well as EMS providers' and patients' input. During the intervention month, 2,708 patients were transported to the participating hospitals. EMS providers contacted the destination-control physician for 1,866 (69%) patients. The original destination was changed for 253 (14%) patients with reasons such as system needs, patient needs, physician affiliation, recent ED or hospital care, patient wishes, and primary care physician wishes. During the intervention month, AD decreased 190 (41%) hours at the university hospital and 62

(61%) hours at the community hospital, as compared with the control month. The authors concluded that this type of program may be effective in reducing overcrowding and maximizing the availability of emergency health care resources.

In the greater Edmonton metropolitan region, AB, Canada, an ambulance destination determination system has been created jointly between the Capital Health Authority (the regional hospital organization) and Edmonton EMS (the primary ambulance provider) [127]. The system functions with staff from the Edmonton EMS and the Regional Patient Transport Office operating together to coordinate the distribution of ambulances to the various hospital EDs. The coordinators have access to the real-time information on both the Edmonton EMS ambulance trips records management system and the Capital Health Authority emergency status screen on the status of each ED in the region. Such information includes the number of ED beds both occupied and available, the number of emergency inpatients and waiting room patients, the number of ED patients in each category of the Canadian Emergency Department Triage and Acuity Scale (CTAS), the number of current active ambulance trips, and the ambulance activities (i.e., dispatch, response, arrival on scene, transport to and arrival at a designated hospital). With the available information, the coordinators make decisions about how to distribute ambulances to the various hospital EDs and provide that information to the EMS personnel. An overall positive response was reported after a 6-month pilot implementation. The three community hospitals had an increase in ambulance transports, with a corresponding decrease for the two major hospitals. The author thus suggested that this ambulance destination determination system helped to maximize available ED resources and was a valid alternative to AD.

An attempt to mitigate the ambulance at-hospital interval (turnaround time) in Baltimore, MD, USA, was conducted by Halliday et al. [128], to improve communication within the local EMS system. A senior EMS paramedic was assigned as the medical duty officer in the fire communication bureau of the Baltimore City Fire Department. The primary task of this position was to provide prospective management of city EMS resources through monitoring ambulance availability and hospital ED traffic, and suggesting alternative transport destinations in the event of ED crowding.

The authors compared a total of 13,921 EMS calls during in the post-intervention period with 15,567 during the pre-intervention period and 14,699 in the seasonal match control period one year earlier. They reported a 1.35-minute decrease of the average at-hospital time from pre- to post-intervention periods, and a 4.53-minute decrease from the seasonal match control to post-intervention periods, representing a statistically significant decrease. Furthermore, hospital alert time was also shown to have a statistically significant difference between the pre- and post-intervention periods in this study with an approximately 1,700-hour decrease. The decrease in ambulance response time was, however, not statistically significant. This study emphasized the importance of better coordination between EMS and hospital EDs as well as future intervention initiatives.

To mitigate the AOD problem, Almehdawe et al. [33] introduced a stylized queueing network model with blocking to investigate the effect of patient routing decisions on EMS offload delays. They constructed and solved an optimization problem to find the optimal allocation of ambulance patients to each ED in a region. The optimization model was tested to be robust under normal operating conditions as supported by the numerical analysis in the study. The authors suggested that this model can be used as a decision support tool to guide EMS dispatchers on how to allocate patients to hospital EDs when they make their dispatching decisions. In one of their earlier studies, Almehdawe et al. [5] analyzed two routing probability scenarios in a three-ED system. The imbalance scenario represented a system where heuristic routing policies were used by emergency control staff, while the balance scenario demonstrated a system where the routing probabilities were proportional to ED capacities. System performance measures were computed, particularly for assessing the AOD and its impact on system resources. The results show that when the system changes from the "imbalance" routing probability scenario to the "balance" one, the expected total number of ambulances in offload decreases by 14% and the total expected AOD decreased by 9.9%. This patient allocation policy can be seen as a proactive intervention to reduce the chance of ED crowding, thereby mitigate AOD. Similar practice has been implemented in some EMS systems as an intervention to cope with AOD / ED crowding in daily operations [63].

Redirecting Patients to Alternative Care Destinations

It is recognized that a substantial proportion of ambulance service calls are neither life threatening nor serious [129]. With constantly increasing demands, many EMS systems have explored the options to screen and divert potentially non-emergent patients from the system at their dispatch centers [130].

Shah et al. [131] reviewed dispatch data on 19,332 calls in Salt Lake City, UT, USA, to identify EMS dispatch codes associated with low illness acuity. A low-acuity dispatch code was defined as one in which at least 90% of coded patients required only basic life support care. 28 out of 118 dispatch codes or code groups, with 7,801 patients, met the definition of low acuity. The authors concluded that certain dispatch codes were associated with likely to be low acuity patients and further validated these codes in a later study [132] . That study concluded that 21 of the dispatch codes can be potentially used to identify low-acuity patients who do not require emergent response. A similar study by Woollard [133] sought expert consensus about which ambulance dispatch codes could be appropriate for a nonemergency response in Cardiff, UK. Using majority voting, the results indicated that 54 dispatch codes (22%) were recommended for a nonemergency response/referral, which equaled to 12.44% of annual emergency calls in a typical UK ambulance service system. Theoretically, the implementation of nonemergency responses could lead to improved response times for critically ill patients by freeing up resources. The author suggested that further research is required to validate the recommendations made by the experts using clinical outcome data. Villarreal et al. [134] investigated a new model of patient screening implemented in West Midlands, UK, where a partnership between general practitioners and ambulance services was formed to reduce conveyance rates to the Hospital EDs. Call handlers identified patients with needs that could be addressed by a general practitioner using pre-determined criteria. General practitioners supported the assessment of such patients either at scene or by telephone. Routine data were collected from October 2012 to November 2013, from the ambulance service computer-aided dispatch system. Logistic regression models were used to determine the likelihood for patients being transported to ED. Of 23,395 emergency contacts during the evaluation period, 1,903 (8.1%) patients were triaged to general practitioner supported assessment. 1,221 (64.2%) had face-to-face assessment with general practitioners and

682 (35.8%) via telephone. 1,500 (78%) of those who received general practitioner support were not transported to hospital. The authors concluded that support of the paramedic service by general practitioners enabled patients to avoid transfer to an ED, potentially avoiding subsequent hospital admission, reducing costs, and improving quality of care for patients that were not in need of hospital services. They also addressed that the overall impact and safety of this model required further evaluation.

However, the risk-management challenges associated with patient screening has made this "politically unpalatable" with occasional bad outcomes [22, 135]. Furthermore, this EMS strategy of discouraging communities from requiring ambulance services in "non-emergencies" has often backfired, with observed increases in calls [130]. In their literature review of addressing the ability of ambulance crews for patient screening, Snooks et al. [138] argued that not enough evidence has indicated that "there is a clinically safe approach to identify patients who call for an ambulance but do not need transportations to ED". Most of the previous work has been hypothetical only, with rare intervention studies, yet consistently showing the need for caution. Millin et al. [135] further addressed that "EMS systems that utilize these policies must have additional education for the providers, a quality improvement process, active physician oversight", and the determination of non-transport for a specific situation should be supported by peer-reviewed literature. Despite these challenges, Snooks et al. [136] suggested that further research in this area is urgently required due to the inefficiency of the current model of emergency care. Eckstein et al. [22] also insisted that EMS systems should continue to explore such patient screening concepts to reduce the demands and to achieve some relief for the system. They recommended some innovative strategies such as finding citizens alternative numbers to call (especially "after hours") or alternative places of appropriate medical care (i.e., shuttle transport to nearby clinics).

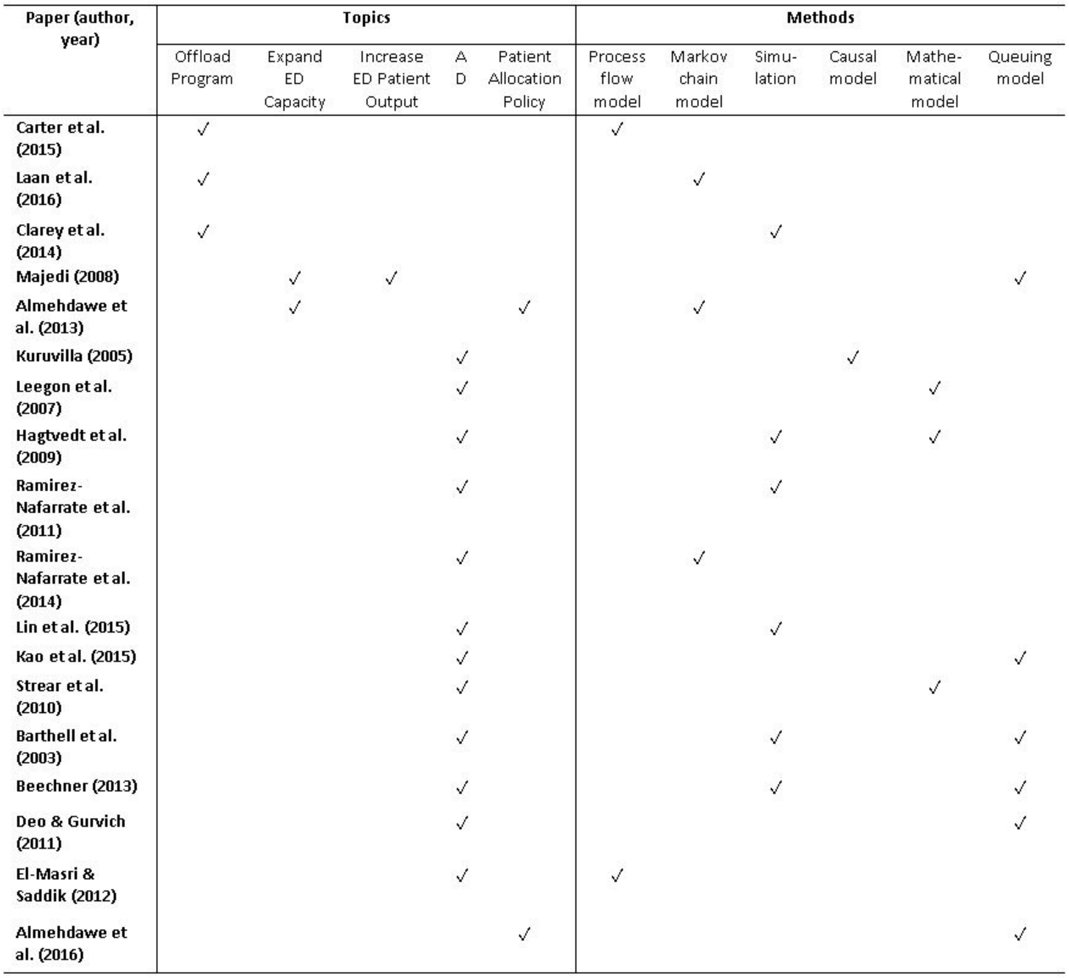
A similar EMS intervention to patient screening is to accept low acuity patients into the system and later redirecting them to alternative care destinations, other than EDs. The Ontario Ministry of Health and Long-Term Care in Canada established the hospital emergency department and ambulance effectiveness working group in 2005 to investigate the AOD problem and advise Ontario's Minister of Health on it. This group submitted a report to the Ministry of Health and Long-Term Care [46] with recommendations aiming to ensure the improvements of ambulance availability. One of their primary proposals was to consider transporting selected ambulance patients to destinations other than EDs, such as urgent care facilities. The group recommended an evaluation of the safety and effectiveness of such initiatives through pilot projects in urban regions in Ontario. The government of Ontario, Canada later followed this recommendation and initiated a demonstration project in the city of Toronto to redirect low acuity ambulance patients to Urgent Care Centres instead of EDs [60]. The motivation behind this intervention was to take advantage of the faster ambulance turnaround time at the Urgent Care Centres comparing to regular EDs. It could release ambulances to be back to the road sooner, therefore, increase ambulance availability in the city. In the evaluation of this project, Esensoy reported a total of 855 hours of reclaimed ambulance time over the two-year trial period. However, this result failed to show that the potential volume of these Urgent Care Centres was high enough to make a significant system-wide impact on the AOD problem with the current setup. The vagueness of the patient clinical criteria for redirecting to the Urgent Care Centres was suggested to be the primary driver for low paramedic uptake. The author argued that such decision-making intervention required extensive training up front and continuous change management activities to ensure a smooth implementation.

A similar intervention was launched in the UK, with the goal of avoiding the admission of minor patients to acute care hospital EDs [129, 137]. By developing and testing a protocol to identify specific low-acuity patients for transport and treatment at urgent care clinics rather than EDs, Schaefer et al. [137] reported a 15% relative decrease (51.8% vs. 44.6%) in the proportion of patients who received care in the ED when compared with a historical control group with similar diagnostic, acuity, and seasonal characteristics, by implementing this intervention of alternate care destinations. The referral was appropriate in 97% of cases, and that the patients transferred on from urgent care clinic to ED did not suffer any delay in resolution of their condition. The authors concluded that despite the low usage rate of alternative care locations, this intervention has time saving benefits to most patients and the ambulance service, therefore, should be continuously employed with improved training. In Snooks et al.' study [129], patients were then followed up and the outcomes of patients taken to alternative care locations were compared with those taken to EDs. The results indicated that patients taken to alternative care locations were 7.2 times as likely to rate their care as excellent. In addition, ambulance service also benefited from this intervention as ambulance on task time was shorter for patients taken to alternative care locations.

Other EMS Interventions

Newell et al. [65] reported a strategy that the ambulance services in Ottawa, ON, Canada has instituted to cope with AOD. Paramedics are not required to hand over their electronic paramedic care report to the receiving hospital ED of a patient who meets certain criteria (not CTAS 1 or 2). The paramedics can depart the hospital right after the transfer of care and complete the report while mobile. The completed report will be uploaded to the server over a secure Wi-Fi connection for the ED staff to view and download. During the eight-week trial period, the average ambulance turnaround time was reported dropped by 14 minutes per patient transported following the new protocol. Despite the success, this intervention has been met with some resistance due to patient safety concerns and hospitals having timely access to patient information. The author acknowledged the concerns and argued that the next potential ambulance patient can also be at significant risk if ambulance resources are tied up in AOD, considering that rapid turnover of patient care is critical to the EMS system. This intervention represents a significant change in workforce culture and needs to be recognized by both the EMS and the hospital EDs.

Eckstein et al. [22] recommended that EMS providers should have a contingency plan in place to approach and mitigate the AOD problem. The importance of collaboration between the EMS provider and the receiving hospital / ED staff was also highlighted in their recommendations to ensure that policies and procedures are in place and the team effort keeps the ambulance turnaround period brief. Another EMS intervention in practice is referred as “mutual aid”. A detailed description of it can be found in Section 2.5.2. Concerns have been raised against this practice regarding ambulance availability of outlying areas. Yet “mutual aid” is still employed by some EMS providers to cope with the AOD problem [28].



Paper (author, year)	Topics					Methods					
	Offload Program	Expand ED Capacity	Increase ED Patient Output	A D	Patient Allocation Policy	Process flow model	Markov chain model	Simu-lation	Causal model	Mathe-matical model	Queuing model
Carter et al. (2015)	√					√					
Laan et al. (2016)	√						√				
Clarey et al. (2014)	√							√			
Majedi (2008)		√	√								√
Almehdawe et al. (2013)		√			√		√				
Kuruvilla (2005)				√					√		
Leegon et al. (2007)				√						√	
Hagtvedt et al. (2009)				√				√		√	
Ramirez-Nafarrate et al. (2011)				√				√			
Ramirez-Nafarrate et al. (2014)				√			√				
Lin et al. (2015)				√				√			
Kao et al. (2015)				√							√
Strear et al. (2010)				√						√	
Barthell et al. (2003)				√				√			√
Beechner (2013)				√				√			√
Deo & Gurvich (2011)				√							√
El-Masri & Saddik (2012)				√		√					
Almehdawe et al. (2016)					√						√

Figure 2.4: The summary of OR models in AOD intervention studies.

2.7 Discussion

The causes and consequences of the growing AOD problem were first described to capture the complexity of this problem. Next, key measures that are used to assess system performance were listed. Then a literature review of related studies and models was carried out to summarize common features, including the data used for research, the methods and approaches, and the main results. 137 articles are reviewed (summarized in the electronic supplementary material), including studies of the causes, effects, and solutions of AOD. Different topics and methodologies are employed throughout these studies (as described in Figures 2.5, note that some papers do not fall under the described topics, while some papers utilize multiple methodologies).

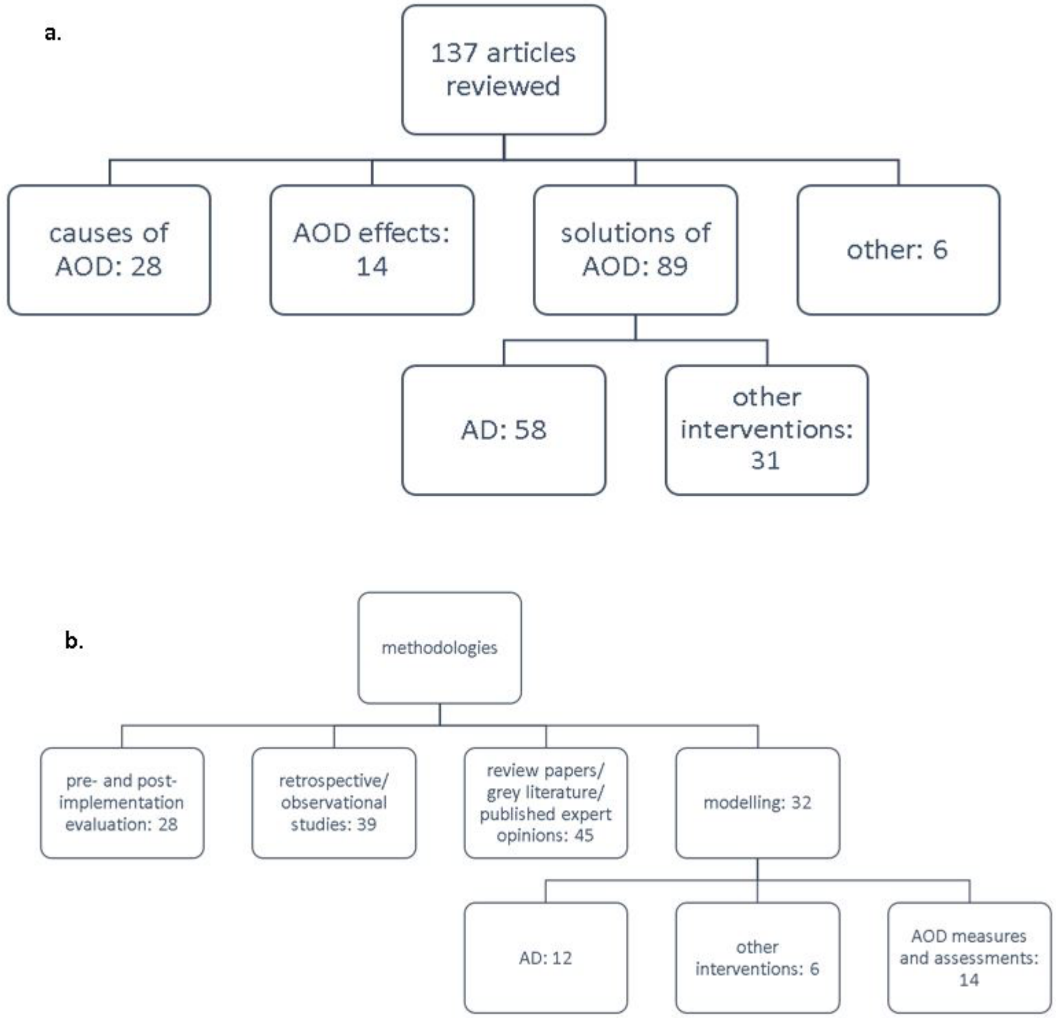

a.
137 articles reviewed
causes of AOD: 28
AOD effects: 14
solutions of AOD: 89
other: 6
AD: 58
other interventions: 31
b.
methodologies
pre- and post-implementation evaluation: 28
retrospective/ observational studies: 39
review papers/ grey literature/ published expert opinions: 45
modelling: 32
AD: 12
other interventions: 6
AOD measures and assessments: 14

Figure 2.5: Summary charts of the reviewed articles: **a**. topics, **b**. methodologies.

The analysis of the literature reveals that many researchers have investigated areas of ED crowding and ambulance diversion; however, there is limited research focused on AOD. Specifically, we found a lack of OR methodologies used in addressing AOD. For instance, of the 89 articles that studied solutions to AOD, only 18 (20.2%) of them introduced OR methodologies to test interventions in a virtual setting. One possible reason is that AOD is a relatively new problem and has not attracted a lot of attention from researchers in the OR field. Another reason could be that the complexity of modelling the interface between the EMS and ED services has deterred

OR researchers.

For the interventions to be effective and true to real-world situations, the measurement of AOD needs to be improved. Presently, no method has been reported to measure the ambulance offload time accurately and reliably. Most research uses ambulance turnaround time as a measure of AOD. Further study is required to standardize the definition and the measurements of the ambulance offload process. It is important that interventions to decrease AOD are based on a solid understanding of the main components of this process.

The empirical assessments of AOD (Section 2.3.1) show that AOD has become a problem in many EMS systems since early 2000. Some data are reported but there appears to be no standardized reporting structure in standard increments. For example, it is difficult to compare AOD times across EMS systems due to the vastly different characteristics of EMS systems (e.g., location, size, etc.); it would therefore be helpful if studies of AOD included characteristics of the EMS system and a data dictionary of standard definitions for these time intervals. Furthermore, historic trends in AOD would be insightful to show the evolution/acceleration of AOD. Several empirical studies report that patients with medium acuity level experience the most prolonged AOD. Yet, to the authors' best knowledge, no further study has been reported that investigates if there is a correlation between AOD and patients risk levels. To be more specific, is AOD more common or prolonged for patients with certain clinical conditions? If so, what are the impacts on the safety and outcomes of these patients?

There are also few studies describing the relationship between ED crowding, ambulance offload time, and EMS performance (such as EMS response times and resource availability). Some studies have suggested that AOD impacts EMS. However, most use anecdotal evidence as opposed to empirical analysis. It would be beneficial for future work to quantify the impact of AOD on EMS systems. Therefore, much remains to be learned to fully understand and assess the AOD problem to improve resource utilization, response time, and patient care.

Another aspect of AOD assessment, which is overlooked in the current literature, is its impact on the workload of paramedics and ED staff. As the ED crowding and AOD become a new norm, are there human resource consequences, such as burn-out, increase rates of human error, morale issues, etc.?

An interesting observation of the literature is that, while ED crowding is a worldwide phenomenon, the AOD problem has only been reported in Canada, the USA, the UK, and Australia. This raises the question of how other EMS systems with ED crowding have avoided AOD, or do they also experience AOD but have not studied it formally? Perhaps other EMS systems have developed and implemented efficient interventions to avoid AOD? Or perhaps AOD is inevitable at a certain level of ED crowding but that level of ED crowding has not yet been reached? Regardless, this can be an important direction for future research to provide some insight into the fundamental cause of AOD and why it appears to be more prevalent in some countries than in others.

Several mitigation interventions of AOD (e.g., AD, expending the ED capacity) have been reported with debatable results. This debate suggests that initiatives and efforts from one party (EMS or ED) alone may not be sufficient to solve this problem. The ability of paramedics to transfer patient care to an ED is determined by the status of the ED, namely, the number of available ED beds. This availability is directly related to hospital throughput and the availability of inpatient beds. Therefore, the AOD problem is a consequence of a much bigger problem, which is the lack of capacity in the healthcare system to treat hospital inpatients, leading to ED overcrowding and access block. The majority of hospital EDs are reported operating at or over their capacities in a typical day, as discussed in Section 2.4. As such, it is not surprising to see that EMS providers continually find themselves struggling with timely patient transfer at hospital EDs. Studies have shown that changing the ED's structure or function cannot address the underlying causes of ED crowding and, therefore, cannot alleviate AOD. The evidence suggests that solutions to ED crowding lie outside the ED and will require system-wide policy changes. EMS systems do not exist in a vacuum, isolated from the rest of the health care system. The AOD problem includes clinical, operational, and administrative perspectives; the efficiency and effectiveness of ambulance offload time must be addressed in a system-wide manner.

Establishing better collaboration between EMS and hospital EDs is the first step forward towards the goal of building a system-wide solution to AOD. Timely information sharing (e.g., ED/EMS status, patient clinical outcome data) between these two parties also allows proactive interventions to mitigate the AOD problem. All the interventions trialed with a collaborative approach have been reported to yield positive results; while voluntarily-based interventions initiated by individual parties often present mixed results. Therefore, it is recommended by many researchers that EMS and hospital EDs initiate dialogues at high management levels and work together to take appropriate steps to mitigate AOD.

The root causes of AOD likely lie outside the EMS system and to address it (like addressing ED crowding), will take significant time and effort. In the meantime, AOD has appeared as a new norm in some EMS operations and needs to be addressed more quickly. Therefore, research should continue to develop interventions, either through OR models or trials, to help EMS operate in this difficult environment and mitigate the negative impacts of AOD.

Healthcare is an area of growing importance and cost around the world, thus an important area for operations research. As an important element of the healthcare network, EMS system requires constant performance improvements to ensure overall capacity to adequately and efficiently respond to emergency needs of the public. To help better assess and mitigate the AOD problem, models need to be further developed to estimate the system performance in a more realistic and detailed environment. While the AOD problem presents itself as a challenging problem, it also represents an opportunity for public health, EMS, and hospitals, to come together to identify best practices and interventions. Ultimately, all key components of the health care system should work together to ensure the ED crowding problem is eliminated or minimized, thereby alleviating much of the AOD problem.

Chapter 3

AN EMPIRICAL ANALYSIS OF THE EFFECT OF AMBULANCE OFFLOAD DELAY ON THE EFFICIENCY OF THE AMBULANCE SYSTEM

3.1 Introduction

In a typical North American emergency medical services (EMS) system setting, when a hospital emergency department (ED) cannot accept the incoming ambulance patient immediately (often due to congestion), paramedics wait with their patient(s), and continue to provide patient care until an ED bed becomes available and the ED personnel assume responsibility for the patient(s). This delay in transfer of care is referred to as ambulance offload delay (AOD). Due to the increasing issue of ED crowding, AOD has become a growing concern for many health care providers [5, 23, 138].

The consequences of AOD can be significant [20]. AOD hinders the promptness of medical treatments for patients with potential negative consequences (e.g., delay to definitive care, poor pain control, delayed time to treatment, etc.), which may result in compromising patient care and safety [20, 27]. AOD can also negatively affect the ability of EMS to provide consistent and timely care, due to reduced ambulance availability [22, 60, 61]. During this delay, the ambulance and crew are unavailable to respond to future emergency calls. It may prolong the ambulance response time and time spent on calls, resulting in decreased efficiency of the EMS systems, and the need for additional staffing [26, 58, 135]. In addition, financial burdens and legal concerns regarding the AOD problem have been reported [5, 29, 30]. In England, AOD has been reported to cost the National Health Service "millions of pounds per year in the form of lost ambulance hours", which has risen from 37,000 hours in 2008/2009 to around 54,000 hours in 2010/2011 [4]. The Region of Waterloo, ON, Canada lost approximately $840,000 in ambulance operations in 2007 due to AOD

[5]. The city of Toronto, ON, Canada lost approximately 180 ambulance hours per day in December 2007 [5]. The EMS provider in Nova Scotia has estimated in 2015 that the AOD problem costs about 2,900 ambulance hours per year, which equates to approximately $754,000 at the average paramedic salary [30].

The AOD problem has only recently become an active research topic. To date, most current research on AOD is carried out by medical doctors and front-line personnel who try to understand the problem and highlight its importance by using observational studies [23, 26, 35–37, 40]. These studies observe the prolonged waiting time of ambulances at hospital EDs, and measure AOD by its mean or median values. Eckstein and Chan [26] analyzed AOD incidents in Los Angeles, CA, USA, which accounted for 1 out of every 8 ambulance transports (8.4% were in excess of 1 hour, a median of 27 minutes and an inter-quartile range of 20 to 40 minutes). Segal et al. [35] examined the ambulance turnaround time at a hospital ED in Montreal, QC, Canada, and found that the turnaround time represents 45% of the total call time (45.24 minutes and 101.06 minutes, respectively). The results show that the majority of the turnaround time occur after the completion of triage (with a mean time of 31.33 minutes), indicating that the ambulances experience difficulties in transferring patient care to the ED (i.e., AOD). Cone et al. [36] conducted a retrospective study to assess the common AOD problem in New South Wales, Australia. Of 141,381 transports, 12.5% of patients experience an AOD of 30 – 60 minutes, and 5% experience a delay of ≥ 60 minutes. Stewart et al. [139] used administrative data to study all high-acuity (Canadian Triage Acuity Scale 2–3) EMS arrivals to EDs from July 2013 to June 2016 in Calgary, AB, Canada. They reported that of 162,002 arrivals, 70,711 (43.65%) had offload delays $<$ 15 minutes and 41,032 (25.33%) had delays $>$ 60 minutes. Silvestri et al. [24] evaluated the AOD association with patient acuity levels. The average offload time is reported to be an overall 32.7 minutes, and 34,39, and 1.6 minutes, respectively, corresponding to patient acuity levels of low, medium, and high. Cooney et al. [140] conducted a similar study and observed significant differences between all five patient acuity levels when measuring the average AOD. The mid-level severity (level 3) is associated with the longest average AOD.

Presently, no method has been reported to measure the ambulance offload time reliably and accurately. Hammond et al. [38] identified significant inconsistencies in the practice and reporting of AOD across all EDs in Southeast Queensland, Australia. Taylor et al. [39] demonstrated a small but significant delay between the recorded ambulance arrival to an ED and the actual delivery of the patient to the clinical area of the ED in Bath, UK. As most EMS systems find it challenging to accurately measure the offload time [7], the majority of research uses ambulance turnaround time, which is the ambulance's total time at hospital, as a measure of AOD instead. Carter et al. [7] tested the validity of using the turnaround time as a surrogate for the offload time, which results in a good correlation (0.753). Other studies have focused on developing different methods to help measure AOD more accurately [7, 38, 39, 41]. Steer et al. [41] use radio frequency identification tags to record ambulance cot traffic throughout the ED ambulance entrance. Cooney et al. [37, 40] explore if the National Emergency Department Overcrowding Scale (NEDOCS) score could be used to predict AOD by assessing the problem in Syracuse, NY, USA. The authors find a positive correlation between the NEDOCS score and AOD.

All these studies represent progress in understanding the AOD problem, offer insight into the consequences of AOD and the potential solutions of it. However, there is a lack of studies exploring the relationship between AOD and EMS performance, such as ambulance response time, total call time, and ambulance availability. Ambulance availability affects overall system performance [7], and depends on many factors (e.g., hour of day, number of ambulances on shift, number of calls received, etc.) [135]. Anecdotally, AOD has been reported to have a significant negative effect on ambulance availability [22, 60, 61], when multiple ambulances are out of service due to AOD [20, 37, 135]. Some studies have suggested that AOD has a negative impact on EMS systems [22, 26, 58, 60, 61, 135]. However, most use anecdotal evidence and rationalizations as opposed to empirical measurements. Therefore, an empirical analysis based on scientific evidence can be beneficial for quantifying the effects of AOD on the EMS systems.

The goal of this study is to quantify the AOD problem occurring in the Halifax Regional Municipality (HRM) area in Nova Scotia, Canada, and to measure the effects of AOD on the provincial EMS system, with combination of urban and rural regions. This study measures EMS system with performance metrics, such as ambulance turnaround time, total call time, response time, and ambulance availability. It aims to show the effect of AOD on the EMS system province-wide, not only on the area where AOD occurs.

The remaining sections of the paper are organized as follows. Section 3.2 provides a description of the data sources, data structure and statistical analysis to analyze the effects of AOD. The results are presented in Section 3.3. In Section 3.4, we discuss the insights of this empirical analysis study. Finally, Section 3.5 includes some conclusions.

3.2 Methods

3.2.1 Study Setting

Emergency Health Services (EHS), the provincial ambulance service provider in Nova Scotia, Canada, serves the whole population of the province. The province is separated into four operational regions: the Western, Northern, Eastern, and Central regions. There are pre-determined numbers of ambulances operating in each region at a given time of a day, handled by different dispatchers at the central EHS communication centre. The ambulance service in each region operates independently most of the time, with the ability to collaborate when required. There is a total of 37 EDs in the province, and the Central Region is served by three of them [141]. The HRM is located within the Central Region. Being the most populated region and containing the only tertiary care trauma center for Nova Scotia, the Central Region often suffers the AOD problem and requires ambulance reinforcement from adjacent regions.

3.2.2 Study Design

The study subject is the AOD problem in the Central Region of Nova Scotia, Canada. The study period is between January 1st, 2016 and December 31st, 2016. The hypothesis is that: 1) AOD has a negative impact on the EMS performance in the region experiencing AOD, and 2) AOD also affects surrounding regions in one of two ways: ambulances transporting patients into the region experiencing AOD may be delayed, and ambulances from surrounding areas may be repositioned into the Central Region to cover the shortages of ambulances. This EHS practice is referred to "mutual aid". It results in ambulances being relocated away from their home service areas, possibly for the duration of their remaining shifts, and represents a potential decrease in surge capacity of the EMS system [28]. Anecdotally, it also has a cascade effect which may cause ambulance shortage in the outlying areas [16]. This study aims to demonstrate the effects of AOD on the system performance, and to gain knowledge of the current AOD problem in Nova Scotia, Canada.

3.2.3 Data Capture

The retrospective study queried data from two different primary data sources: the EHS computer aided dispatch (CAD) system and the electronic patient care reporting (ePCR) system.

The CAD system contains only ambulance operational data. There is no patient or ED information. With a geographic information system (GIS) tracker available for each ambulance, the CAD system monitors and records the location of each ambulance in the province in real time, as well as ambulance activities. The activities that are relevant to this study are included in the data query (see Table 3.1). Each data entry includes an ambulance location (latitude and longitude) and the ambulance's activity underway at that time. Basic clinical call information, such as patient's Canadian Triage and Acuity Scale (CTAS) [142] was queried from the ePCR database to understand the priority given to each call. CTAS is a tool that Canadian EMS systems and hospital EDs use to triage patients according the type and severity of their presenting signs and symptoms, and prioritize patient care requirements. The scale ranks from 1 to 5, where CTAS 1 patients are the ones with most severe medical conditions, while the CTAS 5 patients are the least severe ones.

Each emergency/urgent call for EHS ground ambulance is assigned with a unique identification number, known as the Master Incident Number (MIN), which is generated from the CAD system. Operation related data associated with each call are documented in this system and were extracted for this study, including ambulance locations, operational and transport disposition times (e.g. arrive scene, depart scene, arrival at hospital time, available time, etc.). The data containing personal health identifiers was kept separate from the main study dataset. No data were required from the personal health identifiers dataset.

The following data element categories (Table 3.2) are included in the data query

Ambulance Activity	**Source**
Assign to Post	CAD
At Destination	CAD
Available	CAD
Available Charting	CAD
Called Off Meal	CAD
Cancel Vehicle Assign	CAD
Depart Scene	CAD
Dispatched	CAD
Division Change	CAD
End Meal	CAD
End Shift	CAD
Enroute To Post	CAD
In Quarters	CAD
Late Start Shift	CAD
Late Vehicle	CAD
Local Area	CAD
On Scene	CAD
Out of Service	CAD
Reassign Vehicle	CAD
Remove Out of Service	CAD
Responding	CAD
Shift Add	CAD
Shift Edit	CAD
Staged	CAD
Start Shift	CAD
Start Meal Record	CAD

Table 3.1: The ambulance activities collected from the CAD system.

for each emergency/urgent call:

This data query was exported into a Microsoft® Excel file. Each row of this file represents a call and all the information associated with it. There was a total of 113,173 records during the study period. These records include all ground ambulance vehicles activities associated with calls, including special units such as supervisor vehicles, patient transfer units (PTUs), etc. These special units offer supports to ambulance fleet and crew, but cannot respond to emergency/urgent calls solely without an accompanying ambulance. Therefore, for the objectives of this study, the records associated with these special units were removed from the dataset. The CAD system prohibits deleting records at any circumstances. When a modification is made to a call, a new record is generated with the updated information, and creates a duplicate

Data Element	Source	Definition
Master Incident Number (MIN)	CAD	Unique master incident number assigned to each EHS ambulance call
Ambulance Radio Name	CAD	Ambulance radio name
Ambulance Radio Code	CAD	Ambulance radio code
Ambulance Location	CAD	The latitude and longitude of an ambulance
Incident location type	ePCR / CAD	Location type
Response Mode	CAD	Level of response to call (Level of the response priority to the scene)
Transport Mode	CAD	Level of the response priority to the scene
Date of Service (Request for Service)	CAD	Date of service identifier
Time of Day (Request for Service)	CAD	Time of service request
Arrive Scene Time	CAD	Time ambulance signals arrived on scene
Depart Scene Time	CAD	Time ambulance signals departed scene to go to hospital
Clear Scene	CAD	Time ambulance signals cleared scene
Arrival at Destination	CAD	Time ambulance signals arrived at hospital
Transports Location / Address	CAD	Hospital location that ambulance transfers the patient to
Transfer of Care	CAD	Time ambulance signals transferred patient to hospital
Available Time	CAD	Time crew indicates available for next call (patient care and charting complete)
Call Disposition	ePCR / CAD	Transport outcome (transported or not)
CTAS (First)	ePCR	First documented CTAS of an ambulance patient

Table 3.2: The data elements collected for each emergency/urgent call.

record associated with the same call. The two records share the same MIN. In this study, all records were sorted by the MINs and only the record with the latest updates was kept for each MIN. Any duplicate records were removed, so that each record in the dataset represents a unique call. A total of 100,126 records are remained after duplication removal.

In this study, non-emergency patient transfers are excluded in the analysis of the Central Region. This patient transfer service provides transportation services for patients who need to go from one hospital to another, or between their home and the hospital within Nova Scotia. In the Central Region, most patient transfer calls are handled by the PTUs. It normally does not interfere with the ambulance responses to emergency/urgent calls. Therefore, this additional responsibility of ambulances is neglected in this part of the analysis. However, in the other regions of the province, we acknowledge that not all non-emergency patient transfers are handled by PTUs. Some are fulfilled by utilizing ambulance resources (often when no PTU is available in that region). Therefore, when analyzing the AOD impact on ambulance availability provincially, we included these calls into the analysis with all EHS ground ambulance activities, as patient transfers may affect the number of available ambulances in a region.

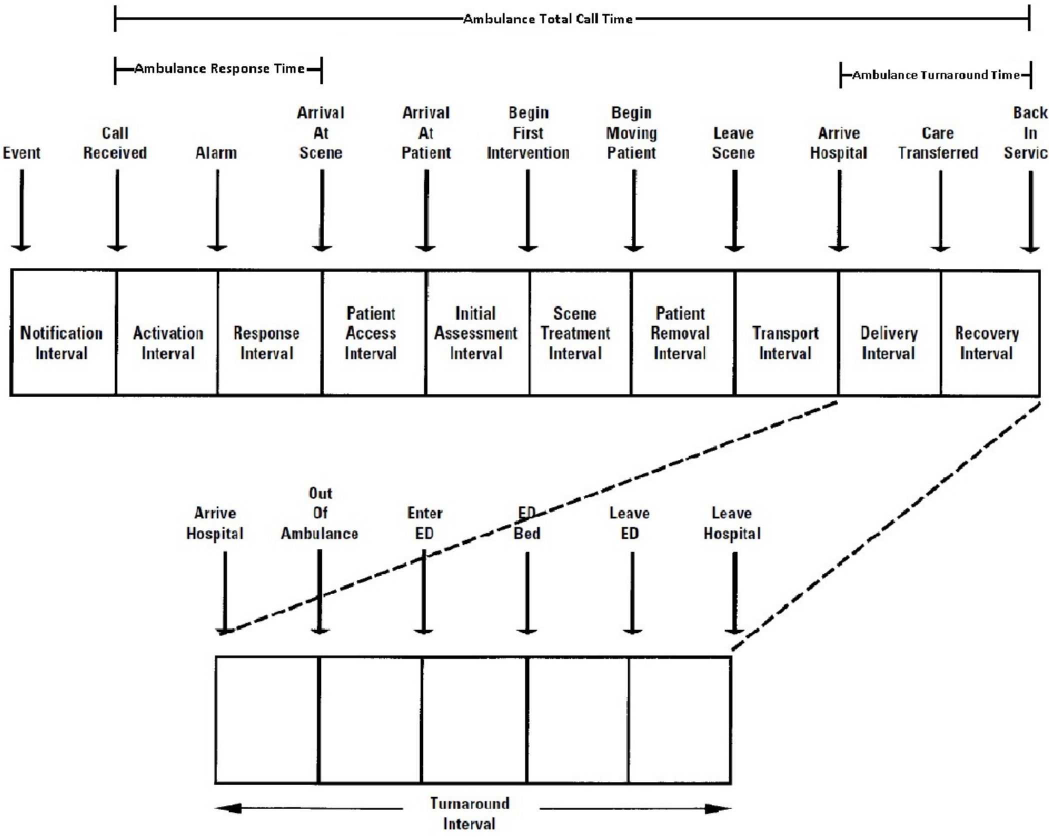


Figure 3.1: A summary of the time intervals of ambulance response events.

3.2.4 Outcomes

Figure 3.1 (adapted from Cone et al. [10] with modifications) summarizes all ambulance activities associated with calls. In this study, ambulance total call time is defined as the time that an ambulance spends to conduct all possible activities associated with a call (from responding to being available for the next call), which includes all the intervals shown in the figure, except the "notification interval". The ambulance response time of a call is defined as the elapsed time from when the call is received at the dispatch centre to when an ambulance arrives at the scene [1]. When an ambulance needs to transport a patient to an ED, the time it spends at the hospital is known as the "turnaround time", or "turnaround interval". It starts when the ambulance arrives at the ED, and ends when the unit is available for future calls. The turnaround interval can further be divided into delivery interval (the actual offload time) and recovery interval (cleaning, restocking the ambulance, completing patient care reports, etc.). When AOD occurs, the delivery interval will be prolonged, meaning that paramedics must wait extra time at the hospital to complete the transfer of patient care to the ED staff. The delivery interval is an accurate measure of potential AOD. However, in our dataset, "transfer of care" is not a mandatory log activity in the CAD system and often left blank. In its absence, this study relies on the ambulance turnaround interval as a proxy for AOD. Literature has shown that the correlation between the delivery and turnaround intervals is good, and the ambulance turnaround time can be used as a surrogate measure. Furthermore, conversations with EHS operation paramedic supervisors have confirmed that the recovery intervals are relatively consistent among all calls. The current EHS policy allows a 20-minute recovery interval and the ambulance will be marked as available at the end of that period, unless a notice has been given by the paramedics to extend that time.

3.2.5 Analysis

A geo-processing application, ArcMap$^{®}$ v.10.5, is used for the location analysis in this study. The purpose of this analysis is two folds. First, it allows us to track the location (region) of each call. These calls can then be separated into four subsets with calls originating in each region. We then reconstructed the queue of waiting ambulances at the EDs in each subset. This is possible since the data record when an ambulance arrives at and leaves from an ED. From these reconstructed queues, we calculate and aggregate the number of ambulances at the EDs in 30-minute increments. Other information was extracted and calculated from the dataset, including call volumes, ambulance response time, turnaround time, total call time, etc. This information was used for the results displayed in Section 3.3.1. The location analysis clarifies the ambulance location (region), instead of the pre-determined region of the ambulance. It is important to add this GIS component to the data analysis, especially in a system with "mutual aid" practice. For example, when an ambulance from the Northern Region comes into the Central Region for one shift due to an AOD-induced ambulance shortage, it is still identified as an ambulance from the Northern Region in the CAD system based on its radio name. However, since it operates in the Central Region for that shift, it is in fact an ambulance resource for the Central Region, not for the

Northern Region. With the GIS analysis, these ambulance activities across different regions can be accurately captured to reflect the true EMS system status in each region. This analysis generates a dataset with the number of ambulances in each region at any given time during the study period. This information is used for the results presented in Section 3.3.2. Queries were run by using Microsoft® Access, and data were exported into and analyzed by using Microsoft® Excel.

We then carried out a multiple regression analysis to assess the relationship between call volume, AOD (two independent variables), and the ambulance availability in the Central Region (dependent variable). R-statistical software, version 3.5.2 (http://www.Rproject.org/), was used for this regression analysis. A multiple linear regression model was built for each region by the following equation:

$$y_{availability} = \beta_0 + \beta_1 x_{calls} + \beta_2 x_{AOD} + \beta_3 x_{calls} \cdot x_{AOD} + \epsilon,$$

where $y_{availability}$ is the dependant variable, representing the hourly ambulance availability in a region, x_{calls} is representing the independent variable of hourly call volume in this region, and x_{AOD} is representing the number of ambulances at Central EDs in an hour. We extended the model to include an interaction term for interaction effects, $x_{calls} \cdot x_{AOD}$, as the *calls* may influence the relationship between *AOD* and *availability*, or vice versa. β_0 is the intercept term, β_1, β_2, and β_3 are the regression coefficients for the independent variables call volume, AOD, and the interaction term, respectively. ϵ is a mean-zero random error term.

To explore the effects of AOD on ambulance availability (Section 3.3.2), the data were integrated to summarize the system characteristics with hourly time intervals in each region, including the hourly call volume and the number of available ambulances. For example, calls received in the Eastern Region between 8 a.m. and 9 a.m. were counted, the value was then assigned to 8 a.m. for the Eastern Region. The number of ambulances at the Central EDs were aggregated the same way as an indicator of AOD in the Central Region. If an ambulance arrives at a Central ED at 8:28 a.m. and leaves at 9:40 a.m., this ambulance will be counted for both the 8 a.m. and 9 a.m. intervals. A total of 8784 (24 $hours/day \times 366$ $days$) data points were generated from the 12-month historical data for each variable. The multiple linear regression model was then fitted with the hourly data entries (366) by minimizing the sum of squared residuals. The results are reported in Section 3.3.2.

3.3 Results

There were 100,126 unique emergency/urgent calls received by EHS in 2016, of which 94,672 calls responded, and 66,169 resulted in transporting patients to an ED. Of these emergency/urgent call records, 23,214 were in the Western Region, 19,090 were in the Northern Region, 17,171 were in the Eastern Region, and 40,651 were in the Central Region. Further analysis of these call records is shown in Table 3.3.

	Emergency/ Urgent Calls	**Responded Calls**	**Calls resulted in patients transported to EDs**	**% of calls resulted in patients transported to EDs**
The Western Region	23,214	22,158	16,379	73.92%
The Northern Region	19,090	18,067	13,596	75.25%
The Eastern Region	17,171	16,289	12,314	75.60%
The Central Region	40,651	38,158	23,880	62.58%

Table 3.3: The summary of year 2016 emergency/urgent calls in each region (Western, Northern, Eastern, and Central) of Nova Scotia.

3.3.1 The Effects of AOD in the Central Region

To examine the effects of AOD in the Central Region, several EMS performance measures were analyzed, including the number of ambulances at EDs, ambulance turnaround time, ambulance total call time, and ambulance response time. To further demonstrate the impact of AOD in the Central Region, some of these performance measures from the other three regions (Western, Northern, and Eastern combined) were analyzed as the controls, because AOD has not been reported as an issue in these regions. Through comparison, readers can better observe and understand the differences of the EMS performances with or without AOD.

Number of Ambulances at EDs

During the 12-month study period, there were a total of 23,880 incidents in which ambulances transfer their patients to an ED in the Central Region. The numbers of ambulances at the Central EDs at any given time of the study period were summarized in Figure 3.2. Overall, there are three or more ambulances at the Central EDs approximately half of the time (46.52%) throughout the year of 2016. The data were further analyzed by being separated into two categories: non-busy hours (8 p.m.-8 a.m.) and busy hours (9 a.m.-7 p.m.), based on emergency call volumes. As expected, there are few ambulances at the Central EDs during the non-busy hours, compared to the busy hours, shown in Figure 3.2.

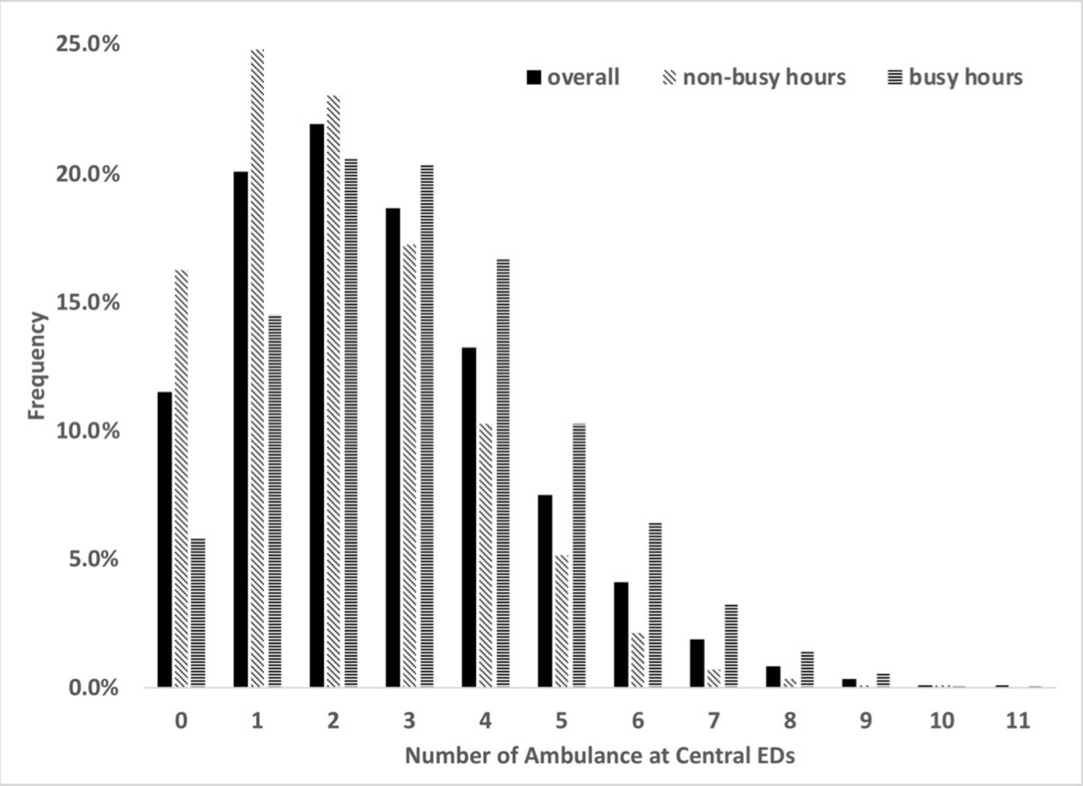


Figure 3.2: The frequency of ambulances held at EDs in the Central Region of Nova Scotia in 2016.

Ambulance Turnaround Time

Of these 23,880 incidents in the Central Region with patients transported to an ED by ambulance, the ambulance turnaround time averaged at 1h04’44”, with a median of 42’20”. Of the 42,289 similar incidents happened in the other three regions, the ambulance turnaround time averaged at 28’31”, with a median of 21’40”. These measures of the ambulance turnaround time were then investigated by stratifying the data by patient CTAS scores to evaluate the differences between the categories The result is reported using the average values, standard deviation (SD), and the 90th percentile, as shown in Table 3.4.

For patients who are categorized into CTAS 1 (most severe), it is crucial for

CTAS	Central Region			Other Regions		
	avg.	SD	90th percentile	avg.	SD	90th percentile
1	34’33"	26’40"	1h16’06"	34’13"	22’30"	1h05’18"
2	1h04’18"	1h03’11"	2h26’12"	33’27"	33’02"	1h02’42"
3	1h08’15"	1h09’02"	2h42’03"	29’25"	29’43"	56’04"
4	47’14"	55’12"	55’12"	25’20"	25’41"	48’30"
5	35’59"	46’43"	46’43"	21’20"	19’26"	41’18"

Table 3.4: The summary of the ambulance turnaround time in Central Region and other three regions combined with the averages, standard deviations, and 90th percentile.

them to receive timely medical attentions without delay during the EMS processes. Therefore, the NSHA policies ensure these patients receive treatments as soon as possible. The result indicates that these policies are implemented as expected. There is no significant difference on the average ambulance turnaround time between the Central Region and the other three regions of Nova Scotia, when transferring CTAS 1 patients to an ED. However, in all the other CTAS categories, patients with similar medical acuity experience significantly longer ambulance turnaround times in the Central Region. Under the assumption that the recovery intervals are similar in the two comparison groups, we conclude that the differences are caused by AOD in the Central Region. The delay affects patients with the mid-level acuity (CTAS 3) the most. One possible explanation is that there are policies implemented in Nova Scotia to allow ambulances offload low-level acuity patients (CTAS 4 & 5) who meet certain criteria to the waiting room of the ED, and thereby free the ambulances from any potential AOD. This effect is demonstrated in the results where low-level acuity patients (CTAS 4 & 5) experience shorter average ambulance turnaround time in both comparison groups. In general, CTAS 3 patients, are too ill to be left unattended in the waiting room, but still have a lower priority comparing to the higher-level acuity patients (CTAS 1 & 2), therefore experience the longest AOD on average. This result is consistent with results reported by Cooney et al. [140].

Ambulance Total Call Time

Ambulance total call time was calculated for each region of Nova Scotia and compared in Table 3.5 to demonstrate the impact of AOD. The calls that did not result in a patient transportation to an ED were excluded from this analysis, as no offload process (or AOD) occurs. The result shown in Table 3.5 indicates that there is a significant difference between the average ambulance total call times in the Central Region and the other regions. It suggests that the prolonged total call time in the Central Region is likely caused by AOD.

	The mean of the ambulance total call time	**The median of the ambulance total call time**	**The 90^{th} percentile of the ambulance total call time**
The Western Region	1h17’11”	1h13’06”	1h55’47"
The Northern Region	1h18’12”	1h10’45”	2h02’15"
The Eastern Region	1h17’59”	1h12’20”	1h59’23"
The Central Region	**1h54’40”**	**1h35’21”**	**3h15’00"**

Table 3.5: The means, the medians, and the 90^{th} percentiles of the ambulance on-task time of calls in the four regions of Nova Scotia in 2016.

For an emergency / urgent call that results in a patient transportation to an ED, the ambulance total call time can be considered as two parts: "prior to ED" and "after arrival at ED". In the Central Region, the time that an ambulance spends "after arrival at ED" on a call takes approximately 50.0% of the ambulance total call time on average; while in the other three regions, this component takes only approximately 10.7% of the ambulance total call time on average, as shown in Figure 3.3. Therefore, it suggests that there is a significant difference between the Central Region and other three regions, in terms of AOD.

Ambulance Response Time

Ambulance response time is a key indicator of the EMS system performance since time is vital in emergency situations. Many factors are associated with ambulance response time (e.g., dispatch logic, ambulance deployment and redeployment strategy, ambulance availability, etc.) [1]. In this study, the relationship between the ambulance response time and the AOD problem was examined to seek the effect of AOD on this key performance indicator in the Central Region of Nova Scotia.

The result in Figure 3.4 demonstrates the relationship between the average ambulance response time and the number of ambulance at the Central Region EDs. As more ambulances were delayed in the offload process, a longer response time was experienced in the Central Region. The values of the average response time are removed from the figure due to data sensitivity, but the scale and trend remain. A scale of

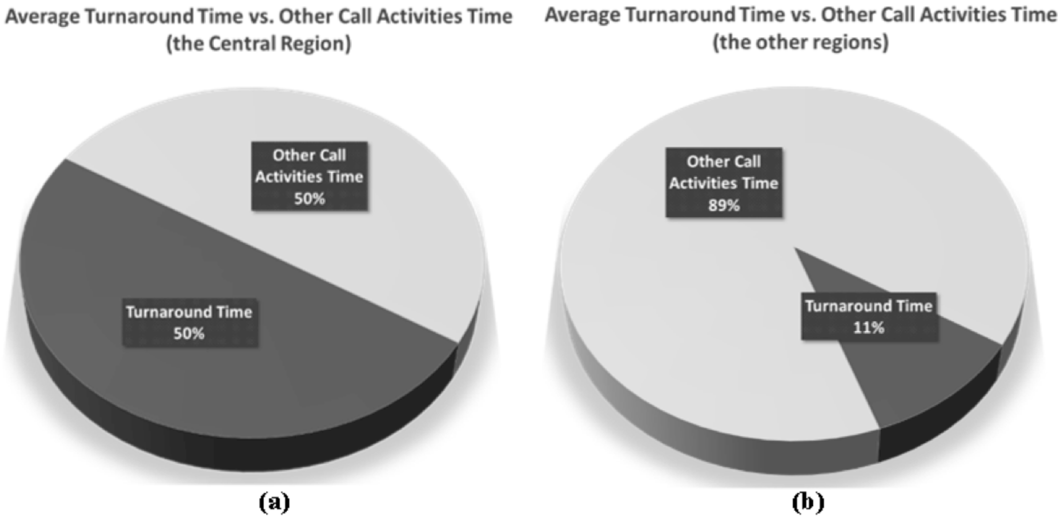


Figure 3.3: The percentages of the ambulance total call time that the ambulance turnaround time take in **(a)** the Central Region; **(b)** in the other three regions.

two minutes is given for the reader to better interpret the effects. The number of ambulances at EDs include all ambulances in the turnaround process, while some ambulances are experiencing AOD in their delivery intervals, and the others may be in their recovery intervals.

3.3.2 The Effects of AOD on the Provincial EMS System

In this section, we analyze the ambulance availability in each region of Nova Scotia to understand the potential effect of AOD occurring in the Central Region on the EMS network across the province. We choose the availability of ambulances as the performance measure to assess the AOD impact. For each region, the ambulance availability is calculated by dividing the number of available ambulances in the region by the total number of ambulances on shift in that region at any given time of a day. A value of ambulance availability of 100(%) means that all the ambulances on shift are available to future calls. The value decreases as some ambulances carry out activities associated with emergency/urgent calls, or other activities (e.g., re-positioning, meal break, etc.).

Figure 3.5 shows the average ambulances availability in each region as a function of hour of the day. The data variances (from the monthly averages) are also reported in

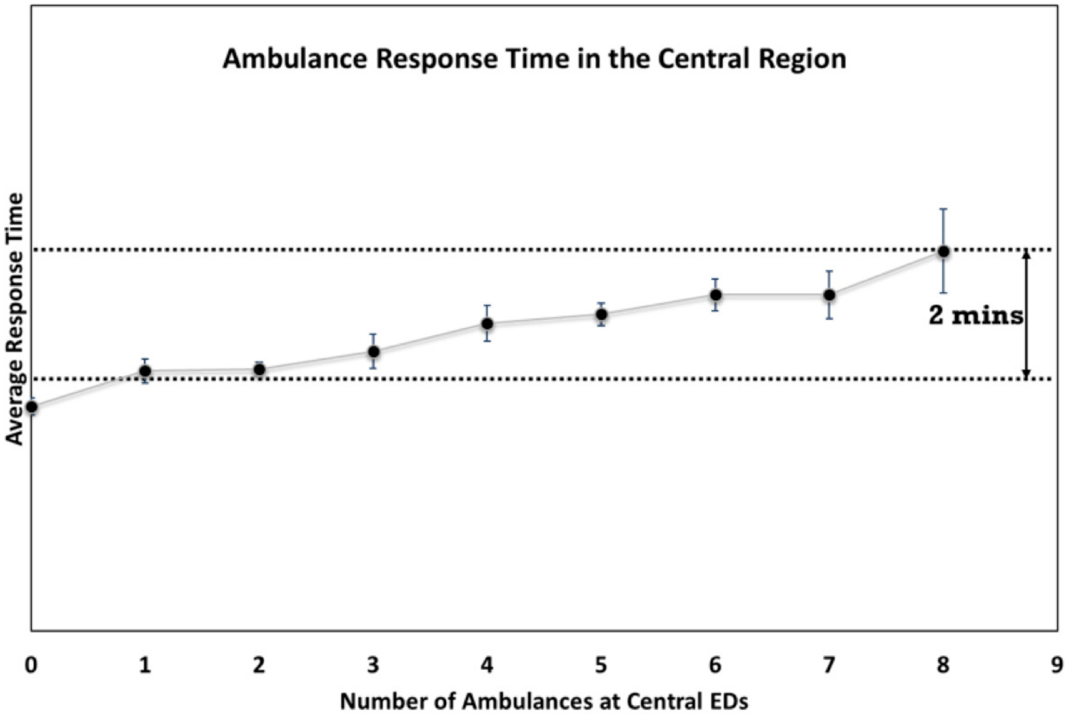


Figure 3.4: The average ambulance response time as a function of the number of ambulance at the EDs in the Central Region in 2016.

the figure as standard errors to demonstrate that there is no any significant monthly differences during the study period. The actual percentages of the availability are removed, but the trends of the curves remain. The result indicates that the average ambulances availability decreases significantly in each region during the day and slowly recover overnight. The steep dip between 11 a.m. and 2 p.m. are likely due to the stacked meal breaks paramedics take between calls.

Means (± SDs) were calculated for the call volume in the Central Region to find any statistical difference in different days of a week, weeks of the year prior to the multiple regression analysis. No significant difference was found. Therefore, the hourly call volumes from all the days during the study period (366 days) were obtained and used to analyze the distribution of the call volume of different hours of day. Figure 3.5 and further data analysis indicates that the relationships between call volumes, AOD, and the ambulance availability vary depending upon the time of a day. Therefore, we model them individually using the hourly aggregated data.

The relationship between the two independent variables (call volume and AOD)

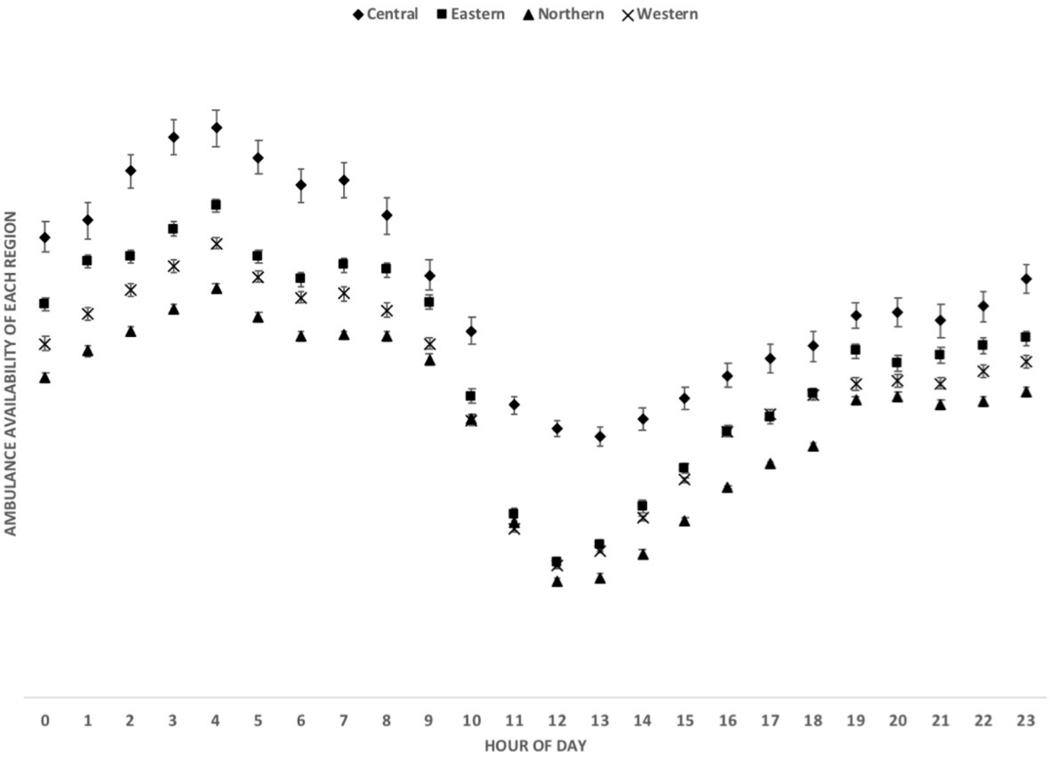


Figure 3.5: The average numbers of available ambulances in each region as a function of the hours of the day in 2016.

suggests that there is no significant correlation between these two independent variables in the data of Central Region. Figure 3.6 demonstrates that there is a lack of linear relationship between the call volume and AOD in any hour of the day, suggesting no strong correlation between these two variables. The distribution of the ambulance availability in the Central Region is approximately normally distributed from the historical data and the two independent variables each follows a Poisson distribution. The same analysis was conducted for the other three regions and similar results are found.

With these results, the multiple linear regression model introduced in Section 3.2.5 was built for the Central Region per hour of the day. Backward selection of the original model indicates that the interaction term can be eliminated from the model. Therefore, the regression model for the Central Region is simplified as:

$$y_{availability} = \beta_0 + \beta_1 x_{calls} + \beta_2 x_{AOD} + \epsilon.$$

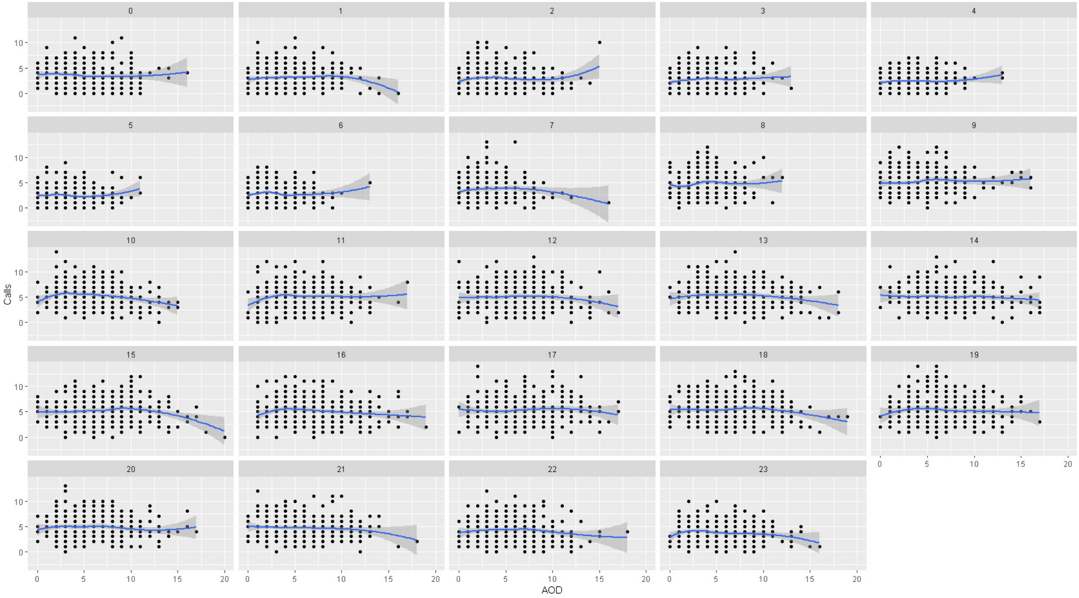


Figure 3.6: The relationship between the call volume and the number of ambulance at EDs in the Central at different hours of the day.

The model residuals, adjusted R^2, F value with the degree of freedom (df), and p value are reported in Table 3.6. There are a total of 24 regression models corresponding to each hour of a day. Both independent variables *Calls* and *AOD* constantly have significant effects on the dependent variable *Availability* among all models with one exception of the 13th hour (1 p.m.).

The coefficients of the independent variables in these regression models are presented in Table 3.7. Together with information presented in Table 3.6, we can summarize the regression equations found for the Central Region based on different hours of the day. For example, between 8 a.m. and 9 a.m., the regression equation is estimated as $Availability = 59.6095 - 1.5717(Calls) - 1.929(AOD) + \epsilon$, with a R^2 of 0.2757 ($F(2, 363) = 70.46, p < 0.001$). The p-values associated with call volume ($p < 0.001$), AOD ($p < 0.001$) are both statistically significant. Note that *Availability* is scaled from 0 to 100 (%) in the equation. The ambulances availability of the Central Region decreases by 1.929% for each ambulance added to the AOD queue, or decreases by 1.5717% for each new call received between 8 a.m. to 9 a.m. The p-values associated with call volume ($p < 0.001$), AOD ($p < 0.001$) are both statistically significant. The value of the adjusted R^2, however, is relatively low. In a complex and stochastic system like EMS, many factors are expected to influence the availability of the ambulances. We only consider two of these factors in our regression model, call volumes and AOD delay, hence, a low adjusted R^2 value is expected. Our goal is not to build a model to estimate the ambulance availability, but to merely explore if AOD delay would affect the ambulance availability.

Hour of Day	**Residuals**			$R^2_{adjusted}$	F	df	p
	min	median	max				
0	-29.576	-0.899	41.477	0.05363	11.34	363	<0.001
1	-27.402	-2.227	50.175	0.05905	12.45	363	<0.001
2	-34.891	-0.728	41.105	0.07197	15.15	363	<0.001
3	-31.202	-0.189	42.267	0.1088	23.29	363	<0.001
4	-36.792	0.449	41.544	0.0569	12.01	363	<0.001
5	-31.14	-0.152	36.29	0.09817	20.87	363	<0.001
6	-22.924	0.205	38.534	0.06738	14.18	363	<0.001
7	-27.801	-0.582	42.101	0.1313	28.58	363	<0.001
8	-24.489	-0.566	36.372	0.2757	70.46	363	<0.001
9	-22.295	-0.751	39.756	0.2386	58.18	363	<0.001
10	-20.957	-0.829	32.637	0.1251	27.09	363	<0.001
11	-15.9972	-0.2787	22.6831	0.06886	14.50	363	<0.001
12	-15.4992	-0.0189	25.0648	0.01821	4.386	363	0.01312
13	-14.6618	-0.7953	23.9343	0.00768	2.412	363	0.09104
14	-18.4533	-0.8495	25.9172	0.03214	7.061	363	<0.001
15	-17.369	-0.362	33.942	0.08851	18.72	363	<0.001
16	-24.9528	-0.2892	29.6604	0.1588	35.44	363	<0.001
17	-26.4553	-0.0702	25.9409	0.2449	60.20	363	<0.001
18	-25.854	-0.302	39.059	0.213	50.40	363	<0.001
19	-24.63	-0.892	35.482	0.1285	27.91	363	<0.001
20	-21.628	-1.577	44.582	0.08243	17.39	363	<0.001
21	-23.081	-1.739	42.584	0.09875	21.00	363	<0.001
22	-24.328	-0.977	33.813	0.08488	17.93	363	<0.001
23	-23.774	-0.914	41.163	0.07787	16.41	363	<0.001

Table 3.6: The summary of the regression models for the Central Region with residuals (min, median and max), the adjusted R^2, F value, df, and p value.

Hour of Day	**Variables**	**Coefficients**			
		Est.	Standard Error	t value	$Pr(> \lvert t \rvert)$
0	*AOD*	-0.6505	0.2207	-2.947	0.00342

	Calls	-1.1824	0.3004	-3.936	<0.001
1	*AOD*	-0.5068	0.2396	-2.115	0.0351
	Calls	-1.6380	0.3694	-4.434	<0.001
2	*AOD*	-0.7026	0.2511	-2.798	0.00542
	Calls	-1.6862	0.3536	-4.769	<0.001
3	*AOD*	-0.5773	0.2719	-2.123	0.0344
	Calls	-2.3876	0.382	-6.25	<0.001
4	*AOD*	-0.8417	0.2635	-3.194	0.001527
	Calls	-1.5528	0.438	-3.545	<0.001
5	*AOD*	-1.0787	0.2595	-4.157	<0.001
	Calls	-1.7159	0.352	-4.875	<0.001
6	*AOD*	-0.8525	0.2546	-3.348	<0.001
	Calls	-1.3542	0.3256	-4.16	<0.001
7	*AOD*	-1.4031	0.2343	-5.989	<0.001
	Calls	-1.1941	0.2627	-4.546	<0.001
8	*AOD*	-1.929	0.2114	-9.125	<0.001
	Calls	-1.5717	0.2278	-6.898	<0.001
9	*AOD*	-1.5568	0.1725	-9.026	<0.001
	Calls	-1.1945	0.2266	-5.271	<0.001
10	*AOD*	-1.072	0.1534	-6.99	<0.001
	Calls	-0.6457	0.2031	-3.18	0.0016
11	*AOD*	-0.5477	0.1158	-4.728	<0.001
	Calls	-0.3594	0.1481	-2.426	0.0157
12	*AOD*	-0.2592	0.114	-2.273	0.0236
	Calls	-0.3144	0.159	-1.978	0.0487
13	*AOD*	-0.1717	0.1073	-1.6	0.111
	Calls	-0.2595	0.1606	-1.616	0.107
14	*AOD*	-0.3782	0.1202	-3.148	0.00178
	Calls	-0.4115	0.1893	-2.173	0.03041
15	*AOD*	-0.48	0.1164	-4.123	<0.001
	Calls	-0.8508	0.1866	-4.56	<0.001
16	*AOD*	-0.7994	0.1258	-6.356	<0.001

	Calls	-1.2033	0.1979	-6.08	<0.001
17	*AOD*	-1.1637	0.1336	-8.713	<0.001
	Calls	-1.2015	0.1884	-6.378	<0.001
18	*AOD*	-1.1565	0.1217	-9.506	<0.001
	Calls	-0.7218	0.1933	-3.734	<0.001
19	*AOD*	-0.839	0.1495	-5.613	<0.001
	Calls	-1.0079	0.1967	-5.124	<0.001
20	*AOD*	-0.6512	0.1631	-3.993	<0.001
	Calls	-1.1451	0.2478	-4.62	<0.001
21	*AOD*	-0.587	0.164	-3.579	<0.001
	Calls	-1.4467	0.2557	-5.657	<0.001
22	*AOD*	-0.5733	0.1666	-3.442	<0.001
	Calls	-1.2257	0.239	-5.129	<0.001
23	*AOD*	-0.5736	0.1874	-3.061	0.00237
	Calls	-1.5314	0.297	-5.156	<0.001

Table 3.7: The coefficients of the independent variables in the regression models for the Central Region.

Generally speaking, all the estimated values in Table 3.7 suggest negative relationships between *Calls*, *AOD*, and *Availability*. Hence, the availability of ambulances in the Central Region can be expected to decrease when a new call originates, or when a new ambulance enters the AOD queue at the EDs. Analysis indicates that the independent variables, the call volume and the number of ambulances at EDs have a significant impact on the ambulance availability during most of the hours with the exception of 1 p.m., where it is likely some other factor not included in our model plays a dominant role to affect the ambulance availability, such as meal breaks, shift changes, etc.

Furthermore, it is recognized from anecdotal evidences that AOD may have some indirect impacts on the ambulance availability of other regions in an EMS system with shared resources. To proof this hypothesis, we used the proposed regression model to analyze the potential impacts of AOD on the ambulance availability in the other three regions (Western, Northern, and Eastern) of Nova Scotia. We assess the relationship between the ambulance availability of other regions and the call volume in each region, as well as the AOD in the Central Region. The procedure of conducting the regression model is similar. Before fitting the regression model, the relationship between the variable hourly call volume in each region and the number of ambulances at Central Region EDs within an hour was investigated and no strong correlation was found. The regression models for each region were then built with the equation described in Section 3.2 for each hour of the day. As the backward selection procedure suggests that the interaction term $calls \cdot AOD$ is a significant variable in some models, the term is kept for the analysis of the other regions.

The results show less consistency throughout the day compared to the results of the Central Region. The inconsistent pattern of the results is likely due to the data separated by hours of the day, causing a scattering effect. There are also differences between regions. While AOD in the Central Region shows some impact on the ambulance availability in the Northern and Western regions, there is no such impact on the Eastern Region. Geographically, the Eastern Region is the furthest region away from the Central Region and not adjacent to the Central Region. Therefore, the influence of AOD in the Central Region is understandably less prominent (if any) to the Eastern Region than to the other two regions. For the Northern and Western regions, we summarize the variables that show the significance in each model in Table 3.8, as a mark of $*$. The overall trend is that AOD in the Central Region has an impact on the ambulance availability of the other two regions primarily in the afternoons and evenings, not in the morning. This results aligns with the fact that AOD normally builds up during the day and reaches the peak later in the afternoon. As such, the result confirms our hypothesis that AOD in the Central Region has a negative impact on the ambulance availability not only in its own region, but also in adjacent regions province-wide.

3.4 Discussion

The results of this study paints a clearer picture of the effects of AOD on the ambulance performance in the region that experiencing it. The most commonly used performance metrics from literature are the number of ambulances waiting at the ED, and the ambulance turnaround time. In this study, we proposed a method to include other performance metrics such as ambulance total call time, ambulance response time, and ambulance availability. We use ambulance performance measures from regions with the same EMS setting but not experiencing AOD as a baseline scenario for the analysis. Through comparison, we demonstrate the significant differences in all these performance metrics between the region experiencing AOD and regions without it. These results clearly demonstrate the impact of AOD on ambulance operations. They can provide an sight to EMS decision makers for quantifying the impact of AOD from a more comprehensive perspective. Another approach of analyzing the AOD impact on ambulance availability in other regions is mixed effects model, where the hour of the day is considered as a random effect variable to be included in the model. The results from the mixed effects models suggest a similar but more general conclusion. It shows that AOD in the Central Region is a significant independent variable to the ambulance availability in the Northern and Western regions. Yet, there is no such significance to the Eastern Region. Alternatively, future work on the same analysis may investigate the potential to group different hours of the day into a few subset of data, based on the distributions of the variables. This grouping mechanism may benefit the regression model and show more definitive pattern of the AOD impact on ambulance availability.

In a complex and stochastic system like EMS, many factors can influence the ambulance availability. Our intention in this study is not to build a model to estimate the ambulance availability, but merely to test the hypothesis that if AOD would affect the ambulance availability. Intuitively, call volume is expected to have an effect on ambulance availability. Therefore, We consider AOD and call volume in our regression model to explore the relationship between these two independent variables and the dependent variable. The values of the adjusted R^2 of these found regression models are relatively low as expected. Future research aiming to develop an estimation model of ambulance availability will require to include other influential factors, such as meal breaks, shift schedules, etc.

3.5 Conclusion

This study provides a comprehensive depiction of the effect of AOD on the Nova Scotia's EMS system, with combination of urban and rural regions. In the Central

Region, AOD was frequent and took a sizable proportion of ambulances out of service in the year of 2016. This led to prolonged ambulance turnaround times, total call times, response times and negatively affects ambulance availability. The ambulance availability in two of other three regions of Nova Scotia is also affected by AOD in the Central Region as AOD causes a cascade effect on other regions. However, the effect is less pronounced and consistent. Any analysis or evaluation of the effects of AOD on EMS systems should take approaches to try to understand its impacts from a system level beyond the region where AOD is measured. The results of this study offer an insight into a more comprehensive understanding of the impacts of AOD on the EMS system. This approach can also be generalized to other EMS systems and regions to quantify AOD and measure its impacts on the EMS system. The AOD problem occurs at the interface of the EMS and the hospital EDs, and includes clinical, operational, and administrative perspectives. Therefore, it must be addressed in a system-wide manner. EMS providers and hospitals need to work collaboratively to implement interventions that can mitigate this problem to improve resource utilization and patient care.

Hour of Day	**Northern Region**			**Western Region**		
	AOD	$calls$	$calls \cdot AOD$	AOD	$calls$	$calls \cdot AOD$
0		*			*	
1	*					
2	*			*		
3	*			*	*	
4				*		
5				*	*	
6						
7				*		*
8						
9						
10						
11						
12						
13						
14					*	
15	*	*	*			
16						
17						
18						
19	*					
20	*		*			
21	*					
22		*	*			
23				*		

Table 3.8: The summary of the significant variables in the regression models for the Northern Region and Western Region.

Chapter 4

PREDICTING AMBULANCE OFFLOAD DELAY USING A HYBRID DECISION TREE ANALYSIS

In the last chapter, the negative effects of AOD in Nova Scotia, Canada have been measured and assessed via analysis of the historical EMS data of year 2016. The results indicate that AOD creates negative impacts in the urban Halifax region and also on the provincial EMS system as a whole. We conclude that AOD is a significant factor impacting ambulance availability together with call volume. As a result, it is critical for EMS providers to obtain information on AOD to plan their operations proactively. Therefore, in this chapter, we further investigate the properties of AOD and design a model to predict the AOD stats at the ED to provide a decision support tool for EMS.

4.1 Introduction

Emergency Medical Service (EMS) system, as a key component of the health care system, faces the challenge to organize its processes more effectively and efficiently to keep up with the increasing demands in aging societies. Researchers have shown great interest in analyzing a variety of EMS processes to make suggestions for improvements in: response time, dispatch time, deployment and redeployment, etc. [11–14]. However, the EMS interface with hospital emergency departments (EDs) has seen less attention.

In recent years, the ambulance offload delay (AOD) problem has been raised as a growing concern for health care providers in many countries [5, 7, 10, 22, 32, 33, 143]. Ambulance offload time is the time it takes to transfer a patient from an ambulance to an ED of a hospital [7]. If the ED cannot accept the incoming ambulance patient immediately due to congestion, a common course of action is to let paramedics continue to provide patient care until an ED bed becomes available. This delay period in transfer of care is referred to as AOD. It is typically caused by overcrowding in the

ED [5, 21–25]. AOD has been associated with negative patient outcomes and poor performances of EMS systems, affecting care quality, patient safety and the system's ability to respond to future calls [20, 27, 28, 58, 138]. As a direct consequence of ED crowding, AOD indicates a deterioration of the EMS system status in the affected area. There are indicators that may suggest that an EMS system is prone to AOD (e.g., high level of ED congestions, high numbers of calls, etc.). Theoretically, these indicators can be used to predict the severity of AOD.

In this study, we introduce a decision-support tool to predict the AOD problem occurring in the Halifax Regional Municipality (HRM) in Nova Scotia, Canada. This area is served by one EMS provider, Emergency Health Services (EHS), and three EDs. Being the most populated area and containing the only tertiary care trauma centre for Nova Scotia, the HRM often suffers from AOD. EHS has estimated in year 2015 that the AOD problem results in about 2,900 ambulance hours per year, which equates to approximately $754,000 at the average paramedic salary [30]. The primary objective of this study is to provide the EHS personnel with a decision-support model that can predict the AOD problem based on the current system status. This way, the decision makers can activate various proactive interventions at different states of the system to mitigate AOD.

Decision trees are popular prediction tools as they produce a model that is easy to interpret. Each leaf node can be presented as an if/then rule. The logical rules followed by a decision tree closely resemble human reasoning and are intuitively appealing to decision makers, who tend to feel more comfortable with models that they can understand [144]. Decision trees are also non-parametric, which can model a wide range of data distributions with no assumption that the data is drawn from one (or a mixture of) probability distributions of known form [145]. This feature is suitable in many cases as the nature of the relationship is unknown. Furthermore, decision trees can handle data of different types without requiring any transformation of the data. Most importantly, decision trees have the capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution that is often easier to interpret [146].

In this study, we use the hybrid decision tree algorithm proposed by Farid et al. [147] with some modifications. This algorithm employs a naïve Bayes classifier to remove noisy instances from the training set before the decision tree induction. It is selected because of its comprehensibility and its prediction accuracy as reported in [147]. The data of this study include indicators of the EMS system status (e.g., hour of day, day of week, number of calls, etc.) and indicators of the hospital EDs status (e.g., NEDOCS score, numbers of ambulance at an ED, etc.).

This chapter is organized as follows: a brief literature review on related work is presented in Section 4.2. In Section 4.3 we introduce the data collection and analysis in greater detail, and formulate the hybrid decision tree model for our case. We then present the results in Section 4.4, with a case study example to demonstrate the application of the prediction model. Section 4.5 discuss the potential benefits of such model in an EMS setting, as well as some suggestions for further improvements. Finally, Section 4.6 provides some general conclusions.

4.2 Related Works

There are two streams of literature related to our work. The first stream is the development of models for the AOD problem. It occurs at the EMS interface with hospital EDs and has seen less attention in the Operations Research (OR) field [7, 10, 135]. Furthermore, the consequences caused by AOD on the EMS system have not been well studied [135]. Only several OR models have been found which analyze this growing issue.

Majedi [32] constructs a system representing the interaction of EMS and ED using queuing theory, and models the behavior of the system as a continuous time Markov chain to evaluate various system performance measures (e.g., the average number of ambulances in offload delay, the average AOD, ambulance and bed utilization). Clarey et al. [66] design a discrete event simulation model to assess the change on AOD in a scenario, where dedicated nurses are hired to assist with ambulance offloading patients. This study demonstrates a clear reduction in AOD when dedicated nursing levels are increased. However, the authors also raise concerns that using this as a sole method to reduce AOD would require unacceptably low staff utilization, which would cost hospitals both financially and in human resourcing. Almehdawe et al. [5] uses a Markov chain queuing model to analyze the interface between an EMS provider and multiple EDs that serve both ambulances and walk-in patients. Matrix-analytic methods are used to solve the steady state probability distributions of queue lengths and waiting times. The study concludes that the priority based admitting policy had a great impact on patient waiting times. When additional resources are considered for the system, the benefit of adding capacity is greater for EDs with higher utilization. Almehdawe et al. later [33] introduce a stylized queuing network model with blocking to investigate the effect of patient routing decisions on EMS offload delays and to determine the optimal allocation of ambulance patients to each ED in a region.

The second stream of the literature relates to decision tree models in health care and popular research trends. A decision tree is a machine learning (ML) method for constructing prediction models from data, which can be used for both classification and regression [148]. Models where the target variable can take a discrete set of values are called classification trees; while models where the target variable can take continuous values (typically real numbers) are called regression trees [149]. A decision tree model logically combines a sequence of simple tests to partition the data and fit a prediction model within each partition. The results of the models can be represented graphically as a decision tree [145].

Numerous decision tree algorithms have been developed, for example, Classification and Regression Tree (CART) [150], Iterative Dichotomiser 3 (ID3) [151], and C4.5 [152]. A recent study by Tjen-Sien et al. [153] compares decision trees and other learning algorithms. The study has shown that these algorithms each have their own advantages and characteristics. Their accuracies are sufficiently similar. The differences are statistically insignificant and probably also insignificant in practical terms [153]. Therefore, all these decision tree algorithms can be found widely and almost evenly used in multiple fields [154, 155], tailored to specific research.

Decision trees have also been used extensively in the health care settings, including clinical diagnostics, drug development [156], medical predictions, and data analysis [157]. Handley et al. [158] used CART modeling to determine specific risk profiles and predictors of suicidal ideation in a community-based sample of older adults. Chen et al. [159] utilized a decision tree model (C4.5) empowered by the particle swarm optimization algorithm to achieve efficient gene selection from thousands of candidate genes that may contribute to the occurrence of cancers. Snousy et al. [160] used various decision tree methods (C4.5, CART, etc.) to determine genes that are highly expressed in cancer cells, and compared the classification accuracy among them. Patel et al. [161] developed a decision tree using the CART method to create risk strata (age, country, per capita government health expenditure, and delay from symptom onset to hospitalization) for mortality of human HPAI H5N1 reported in World Health Organization Global Alert and Response. Luk et al. [162] employed artificial neural network and decision tree (CART) data-mining methods to analyze the patient profiling data and to delineate significant patterns and trends for discriminating hepatocellular carcinoma from non-malignant liver tissues. Chang and Chen [157] used decision tree (C5.0, similar to C4.5 with improvements) combining with neural network classification methods to construct the best predictive model to increase the quality of dermatologic diagnosis.

Many fields benefit from using various ML methodologies to discover hidden patterns and properties of systems over the past decades. However, data sets with unique characteristics and properties may require different ML methods to generate robust and accurate predictive models. To better guide the selection of the ML methods, research has been carried out to apply various ML methods to a multitude of data sets to compare their performances and determine which outperforms the others under certain circumstances. Decision tree, as a widely accepted ML method, is still a popular classification approach for its ease of construction and its ability of interpretation. Demšar [163] theoretically and empirically examines several suitable tests (e.g., the Wilcoxon signed ranks test and the Friedman test) to compare classification algorithms on multiple data sets. Others propose and review different statistical tests to compare different ML algorithms. Alpaydin [164] proposed a 5x2 cv (five replications with two-fold cross validation) F test that combines multiple statistics to get a more robust test when comparing supervised classification learning algorithms. Brazdil and Soares [165] present three ranking methods to investigate the problem of using past performance information to select an algorithm for a given classification problem, including: average ranks, success rate ratios and significant wins. A combination of Friedman's test and Dunn's multiple comparison procedure is adopted to compare ranking methods.

The development of a decision tree model includes two phases: tree growing and tree pruning. Tree pruning is a crucial step to avoid over-fitting the model and ensure the accuracy of the model. Therefore, researchers and statisticians have expressed their interests in analyzing different pruning methods and reviewing them through performance comparisons of the decision tree models. Elomaa [166] analyzes the reduced error pruning method to clarify the different variants and to bring new insight to its algorithm properties. Esposito et al. [167] conduct a comparative study of six well-known pruning methods (reduced error pruning, pessimistic error pruning, minimum error pruning, critical value pruning, cost-complexity pruning, and error-based pruning) with the aim of understanding their theoretical foundations, their computational complexity, and the strengths and weaknesses of their formulation. Quinlan [168] discusses and compares four pruning techniques for simplifying decision trees while retaining their accuracy, including cost-complexity pruning, reduced error pruning, pessimistic pruning, and simplifying to production rules. Bradford et al. [169] describe an experimental study of pruning methods for decision tree models to minimize loss rather than error and conclude that no single pruning algorithm dominated over all data sets. The study revealed that using the Laplace correction at the leaves is beneficial and aids all pruning methods used.

A sufficient number of hybrid algorithms have been proposed to improve the decision tree algorithms by combining them with other algorithms. Balamurugan and Rajaram [170] proposes a method to resolve the tie that appears during the rule generation procedure in basic decision tree induction algorithms. The tie occurs in decision tree induction algorithms when the class prediction at a leaf node cannot be determined by majority voting. The improvement is demonstrated by experimental results on various data sets. Garofalakis et al. [171] construct "simple" decision trees with few nodes by specifying constraints on tree size or accuracy, so that they are easy for humans to interpret. Polat and Güneş [172] propose a novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems. Chandra and Varghese [173] present a fuzzy decision tree algorithm to fuzzify the decision boundary to avoid the problem that the traditional decision tree algorithms face: having sharp decision boundaries which may not be found in all classification problems. Aviad and Roy [174] introduce a decision tree construction method based on adjusted cluster analysis classification called classification by clustering (CbC). Li et al. [175] present a cluster-based logistic regression model with a regression and classification tree approach employed to split the source date to clusters at first. The clusters are further considered as the dummy variables for the logistic regression analysis. De Caigny et al. [176] propose a new hybrid algorithm, the logit leaf model (LLM), that enhances logistic regression and decision tree in two stages: a segmentation phase and a prediction phase. In the first stage customer segments are identified using decision rules and in the second stage a model is created for every leaf of this tree. Aitkenhead [177] addresses the problem that the decision tree structure can be vulnerable to changes in the training data set and presents an evolutionary method which allows decision tree flexibility through the use of co-evolving competition between the decision tree and the training data set. Llorà and Garrell [178] propose a fine-grained parallel evolutionary algorithm to induce a decision trees with an unified algorithm based on artificial evolution. Farid et al. [147] introduce two independent hybrid algorithms to improve the classification accuracy rates of decision tree and naïve Bayes classifiers for the classification of multi-class problems. In the first proposed hybrid decision tree algorithm, a naïve Bayes classifier is employed to remove the noisy instances from the training set before the decision tree induction; while in the second proposed hybrid naïve Bayes classifier, a decision tree induction is employed to select a comparatively more important subset of attributes for the production of naïve assumption of class conditional independence.

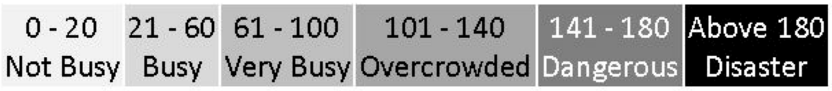
4.3 Methods

In this study, we aim to develop a robust and accurate model to predict the AOD states at a major ED in the HRM region of Nova Scotia, Canada. AOD is complex and stochastic, and can be affected by many factors. Data for this study originates from ambulance operation logs and basic measures of ED crowding. Such operational data are commonly available in health system but are prone to be noisy and inconsistent. Therefore, we searched for a sophisticated decision tree algorithm that can provide prediction accuracy while maintaining a simple structure of a tree, as the interpretability of the model is critical to convey the results to decision makers. Farid et al. [147] introduce a hybrid decision tree algorithm using a naïve Bayes classifier to remove the noisy instances from the training set before the decision tree induction. The naïve Bayes classifier removes misclassified observations by selecting the class that has the highest posterior probability as the final classification for the instance. After removing these instances, we subsequently built a decision tree model using the updated training data set with noise-free data. The model aims to predict the number of ambulances at the ED in X hours (where $X \in 1, 2, \ldots$). Three prediction models with various sets of AOD state classifications are defined later in this section, to fit the specific requests of the ambulance service provider.

4.3.1 Data Collection

The study population includes all emergency and urgent calls for EHS ground ambulance services between January 1^{st}, 2016 and December 31^{st}, 2016 in the HRM region. The data were collected from two different primary data sources: the EHS computer aided dispatch (CAD) system, and the Emergency Department Information System (EDIS) database reporting ED congestion in HRM.

Each emergency/urgent call for EHS ground ambulance is assigned with a unique identification number, known as the Master Incident Number (MIN), which is generated from the CAD system. All EMS responses for completing that call are documented in the CAD system, including operational and transport dispositions. With a geographic information system (GIS) tracker available for each ambulance, the CAD system also monitors and records the location of each ambulance in real time. The National Emergency Department Overcrowding Scale (NEDOCS) [179] from the EDIS database is shared with EHS regarding the status of the EDs in the HRM. NEDOCS is a performance measure (ranges between 0 and 200) implemented in most of the North American's EDs to assess the degree of crowding. These scores can be categorized into groups: "not busy" (0 − 20), "busy" (20 − 60), "very busy" (61 − 100), "overcrowding" (101 − 140), "dangerous" (141 − 180), and "disaster" ($>$ 180) . Figure 4.1 shows these NEDOCS categories.



0 - 20	21 - 60	61 - 100	101 - 140	141 - 180	Above 180
Not Busy	Busy	Very Busy	Overcrowded	Dangerous	Disaster

Figure 4.1: The NEDOCS categories.

Operational data associated with each call during the study period (e.g., arrive scene time, depart scene time, arrival at hospital time, available time, locations, etc.) were abstracted from the CAD system. The NEDOCS records from EDIS were collected for the study period and used to evaluate the level of ED congestion when a call originated.

The following data element categories are included in the query:

- MIN number (from CAD): this is provided as a call ID to link all ambulance activities associate with a specific call in the CAD system.

- Operational call data (from CAD): ambulance radio name, ambulance location (latitude and longitude), transport mode (response priority to hospital), date of service, time of day, ambulance activities (including arrive scene time, depart scene time, clear scene time, arrival at destination time, transports location/address, available time), call disposition.

- ED status (from EDIS Interval Report, HRM region): NEDOCS records (5 minutes interval) at the Queen Elizabeth II Health Science Centre ED.

The EMS system in this study can be viewed as a system that is only responsible for emergency/urgent calls, as the data includes information on emergency/urgent calls but not on other non-urgent functions of ambulances such as patient transfers.

4.3.2 Data Analysis

The ED at the Queen Elizabeth II Health Science Centre in Halifax, Nova Scotia is the major ED serving the HRM. Thus, in this study, the decision tree model was built by analyzing the data set associated with this ED. Among the available data that related to AOD, the following were identified and included as the predictor variables of the decision tree model: the day of week, the hour of day, the call volume, the clear rate of ambulances at the ED, the NEDOCS score of the ED, and the current number of ambulances at the ED. Table 4.1 summarizes these predictor variables and the rationale to include them in the model.

The call records resulting in patient transporting to the Queen Elizabeth II Health Science Centre ED (13,486) were sorted by arrival at ED time and the available time (ready to leave the ED) to restructure the queue at the ED. This queue was then used to:

Predictor Variable	**Motivation to Include**
Day of Week	To capture the variances of the system for different days of week
Hour of Day	To capture the variances of the system for different hours of day
Number of Calls per Hour	To incorporate the effect of call volumes
Ambulance Clear Rate at the ED per Hour	To incorporate the effect of ambulance clear rates at EDs
NEDOCS Score (in categories)	To incorporate the impact of the ED congestions
Number of Ambulances Currently at the ED	To incorporate the current status of AOD

Table 4.1: The summary of the predictor variables of the model and the rationale to include them in the model.

- determine the maximum number of ambulances at the ED each hour;
- determine the number of ambulances cleared from the ED each hour.

The average hourly NEDOCS scores were calculated and matched with the hourly call volume and clear rate at the ED by date and hour. The NEDOCS scores are categorized using "not busy", "busy", "overcrowding", etc. (see Figure 4.1). At this point, all data points of the predictor variables were obtained. Some data points were missing because no event was associated with certain date and hour combinations. For example, if no call was received between 2 a.m. and 3 a.m. on January 10th, 2016, then that field of "Number of Calls" would be empty. In this case, the value of that variable was set to zero. Similarly, when no ambulance arrived at or cleared from the ED in an hour, the corresponding fields were set to zero. The final data set of the predictor variables was thus a matrix of 8784 rows (24 hours/day × 366 days) and 6 columns (each predictor variable per column). We choose to aggregate the data hourly as it is sufficiently detailed for the decision makers and helps to reduce noise.

The response variable of the prediction model is the AOD status of the system at some future moment in time (e.g., in X hours). It varies based on the specific purposes of the prediction. These data points were obtained from the aggregated data set and generated a matrix of $(8784 - X)$ rows and 1 column. For example, when the focus is to predict the categorical AOD states, say, ≥ 9 ambulances in AOD in X hours, the classification groups can be defined as: class 1 is that there are 0 to 8 ambulances in AOD in X hours, while class 2 is the rest (≥ 9).

We developed three prediction models with various sets of AOD states of the system as the classifications to fit the specific requests of the ambulance service provider. A summary categories of these prediction models with the historical distributions of the instances can be found in Table 4.2.

Model A: this prediction model aims to predict the AOD states with a high level of precision. Each class includes approximately three different numbers of ambulances at the ED. The AOD states are defined as: good (0,1,2), bad (3,4,5), problematic (6,7,8), and excessive(≥ 9).

Model B: this prediction model considers historical probabilities (frequency) of different numbers of ambulances at EDs and defines three AOD states to evenly distribute these probabilities. The AOD states are defined as good (0-3), bad (4-6), and problematic (≥ 7).

Model C: this prediction model focuses on identifying the excessive AOD states. An excessive AOD problem may be the most problematic and requires a long recovery period for the system to be back to its normal status. Therefore, this model only consider two classes: normal (0-8) and excessive (≥ 9).

Model	Number of Classes	Class Name	Number of Ambulances at the ED in X hours	Historical Probability, %
A	4	Good	0,1,2	22.82
		Bad	3,4,5	35.33
		Problematic	6,7,8	31.77
		Excessive	≥ 9	10.08
B	3	Good	0,1,2,3	37.21
		Bad	4,5,6	33.54
		Problematic	≥ 7	29.25
C	2	Normal	0-8	89.92
		Excessive	≥ 9	10.08

Table 4.2: The three prediction models A, B, and C, with different classification categories.

4.3.3 Model Development

For each model defined in Table 4.2, we first randomly separate the data set into two sub-sets: the training set and the test set, with approximately 90% and 10% of the data, respectively. The training set, $D = \{x_1, x_2, \cdots, x_n\}$, consists of n observations. Each observation in the set is represented as x_i. The set of predictor variables of x_i is represented as A_i, contains the following attribute values$\{A_{i1}, A_{i2}, \cdots, A_{ij}\}$, where i is the number of training observations, and j is the number of different predictor variables. The response variable of x_i is represented as $C_m, (m = 1, 2, \cdots, k)$, where k is the number of different classes for x_i in D. We then apply a naïve Bayes classifier to each observation, $x_i \in D$. We calculate the prior probability $P(C_m)$ for each class in D, and the conditional probabilities $P(A_{ij} \mid C_m)$ for each predictor variable value in D. After classifying each observation, $x_i \in D$, using these probabilities, the class, C_m, with the highest posterior probability $P(C_m \mid x_i)$ is selected as the final classification of that observation. All observations with lower posterior probabilities are removed. The remaining data in the training set, which includes sufficient training observations, was used for the decision tree induction. This was carried out by using the standard CART algorithm [150] built in Matlab® R2018b. The tree was fully grown first, and then the post-pruning procedure was conducted by using a 10-fold cross validation to obtain the smallest tree whose cost is within one standard error of the minimum cost. The pruning procedure was not included in the work of Farid et al. [147]. We feel, however, the goal of this model is to encapsulate the training data in the smallest possible tree, as the simplest possible explanation for a set of phenomena is preferred over other explanations. A simpler tree often avoids over-fitting. Also, small trees produce decisions faster than large trees, and they are much easier to understand. Therefore, we introduced this modification to the hybrid decision tree model in this study. Algorithm 1 outlines the hybrid decision tree algorithm.

```
**Hybrid Decision Tree Algorithm**

**Input**
$D = \{x_1, x_2, \cdots, x_n\}$ - Training set that containing a set of observations and their associated classes
**Output**
$T$ - Decision tree
**Method**
*1: Naïve Bayes Algorithm*
  **for** each class, $C_m \in D$, **do**
    Find the prior probabilities, $P(C_m)$.
  **end for**
```

```
    for each predictor variable value, $A_{ij} \in D$, do
      Find the class conditional probabilities, $P(A_{ij} \mid C_m)$
    end for
2: Remove noisy observations
    for each training observation, $x_i \in D$, do
      Find the posterior probabilities, $P(C_m \mid x_i)$.
      if $x_i$ is misclassified, do
        Remove $x_i$ from $D$
      end if
    end for
3: Build a decision tree using the purified training data
    $T = 0$
    for each predictor variable, $A_i \in D$, do
      Determine best splitting attribute using Gini Diversity Index:

            $1 - \sum_i [p(i)]^2,$

      where the $p(i)$ is the observed fraction of classes with class $i$ that reach the node.
      T = Create a node and label it with the splitting attribute
      T = Add arc to the node for each split predicate and label
      D = Dataset created by applying splitting predicate to D
      if stopping point reached for this path, do
        $T'$ = Create a leaf node and label it with an appropriate class
      else
        Repeat the for loop
      end if
    $T$ = Add $T'$ to $T$
    end for
4: Prune the full grown decision tree using using a 10-fold cross validation to obtain
the smallest tree whose cost is within one standard error of the minimum cost
```

This model was then evaluated by comparing the predicted class with the target class (true class) of each observation in the test set of the data. The results are reported in Section 4.4.

4.4 Results

The case study was conducted using a Toshiba Portege R30-C computer with an Intel Core i5 processor and 16 GB RAM. Algorithms were coded and executed in Matlab® R2018b. We programmed the hybrid decision tree algorithm as well as a basic CART decision tree for comparison. In this section, we denote these two algorithms as *DT* for the basic CART decision tree algorithm and *NBTree* for the hybrid decision tree algorithm, respectively. The prediction period X is selected to be $X = \{1, 2, 3, 4, 5\}$ hours for each Model **A**, **B**, and **C**, respectively. These models aim to predict AOD states one to five hours into the future. Therefore, a total of 30 ($2 \times 3 \times 5$) prediction models are built in this study: two algorithms (*DT* and *NBTree*), three classification settings (Model **A**, **B**, and **C**), and five prediction periods (1-5 hours). The results cover both immediate and short-term time scales for EHS.

4.4.1 Model Comparison

Historical Data

To compare the two proposed methods *DT* and *NBTree* for our case study, we have used the classification accuracy, precision, and sensitivity-specificity analysis. The classification accuracy is evaluated by the data in the test set. The accuracy of the prediction model is the total number of correctly classified points divided by the total number of data points in the test data set:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Table 4.3 summarizes the classification accuracy rates of *DT* and *NBTree* for each of the 30 data sets. Generally speaking, the accuracy increases as the model becomes less precise. The ranges of accuracy (1-5 hours of prediction period) are approximately 60%-75%, 69%-83%, and 91%-95%, for Model **A**, **B**, and **C**, respectively.

The results in Table 4.3 and Figure 4.2 indicate that *NBTree* outperforms *DT* in most of the cases with only one exception (Model **A**, $X = 5$ hours). The improvements range from -0.18% to 6.32% with an average of 2.44%. This result is consistent with those reported in [147]. The *NBTree* algorithm is capable of identifying the noisy instances from each dataset before the decision tree induction. This prediction model generated from the updated training set is less likely to become overfitting and thus able to carry more generalization capabilities comparing to the *DT* algorithm generated directly from the original training set.

From Model **A** to **C**, the focus of the model shifts from predicting more detailed AOD states to merely predicting the troublesome AOD states in the near future (1-5 hours). As a result, one can expect the accuracy of the classification model increases from Model **A** to **C**. Also as expected, the accuracy tends to decrease gradually while the prediction period increases from 1 to 5 hours, shown in both algorithms (Figure 4.2). *NBTree* performs less consistently comparing to *DT*, with more aggressive changes in accuracy associated with the prediction period. When the prediction period is longer, the difference in classification accuracy between the two algorithms is smaller. It suggests that the accuracy of the prediction model may be affected by other factors than data noise or overfitting (e.g., limitation of the training set).

Model	**Prediction Period (hours)**	***DT*** (%)	***NBTree*** (%)	**Difference** (%)
A	1	70.97	75.35	4.37
	2	61.94	68.26	6.32
	3	59.66	64.95	5.29
	4	58.97	62.33	3.36
	5	60.57	60.39	-0.18
B	1	82.63	82.65	0.02
	2	72.57	76.60	4.03
	3	71.20	73.86	2.66
	4	69.94	70.78	0.84
	5	69.37	69.52	0.15
C	1	93.03	95.21	2.18
	2	90.97	94.06	3.09
	3	89.83	91.12	2.29
	4	90.74	92.12	1.38
	5	90.86	91.55	0.69

Table 4.3: The classification accuracy rates of *DT* and *NBTree*.

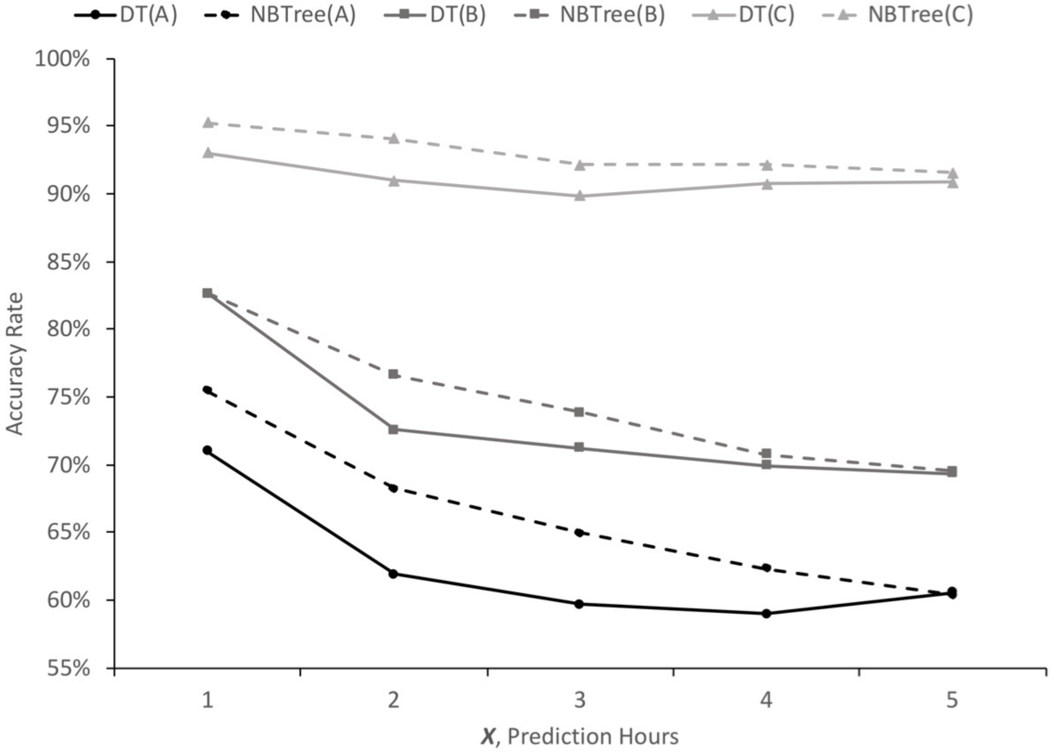


Figure 4.2: The comparison of the classification accuracy rates of *DT* and *NBTree*.

Furthermore, we have calculated the classification precision, sensitivity, and specificity for each model to compare the performances of *DT* and *NBTree*. We reported these values as the weighted average values, which are calculated by using the following equations:

$$precision = \frac{\sum_{m=1}^{k} \frac{(TP)_m}{(TP)_m+(FP)_m} \cdot N_m}{\sum_{m=1}^{k} N_m}$$

$$sensitivity = \frac{\sum_{m=1}^{k} \frac{(TP)_m}{(TP)_m+(FN)_m} \cdot N_m}{\sum_{m=1}^{k} N_m}$$

$$specificity = \frac{\sum_{m=1}^{k} \frac{(TN)_m}{(TN)_m+(FP)_m} \cdot N_m}{\sum_{m=1}^{k} N_m}$$

where k is the number of classes and N_m is the number of observations in the m^{th} classes, $m = 1, 2, \cdots, k$. Instead of each data point contributes equally to the final average, the weighted average is calculated using the number of instances belonging to one class divided by the total number of instances in one dataset. The values of the weighted average precision, sensitivity and specificity are presented in Tables 4.4, 4.5, and 4.6, respectively. The comparison of the prediction accuracy, precision, sensitivity, and specificity confirms that, for this case study, *NBTree* outperforms *DT* in most cases. Furthermore, the results from *NBTree* can still be presented in an easily interpretable form for the decision makers as it maintains a decision tree structure.

Model	**Prediction Period** (hours)	**DT Precision** (weighted avg.,%)	**NBTree Precision** (weighted avg., %)	**Diff.** (%)
A	1	71.04	76.19	5.15
	2	62.84	70.76	7.92
	3	62.54	66.64	4.11
	4	61.13	64.80	3.67
	5	63.73	65.38	1.64
B	1	83.37	82.75	-0.62
	2	75.28	77.33	2.06
	3	73.76	75.32	1.56
	4	73.91	74.08	0.17
	5	74.15	72.43	-1.73
C	1	93.29	95.34	2.05
	2	92.67	95.46	2.80
	3	96.79	94.45	-2.34
	4	93.97	93.64	-0.33
	5	98.36	95.63	-2.73

Table 4.4: The classification precision values of *DT* and *NBTree*.

Synthetic Data

In this section the prediction accuracy of the proposed method is further examined and compared with the CART decision tree algorithms using synthetic data. Distributions of call volume and clear rate of ambulances at EDs are determined from real-world historic data. Synthetic data are then generated following the same distributions.

Model	**Prediction Period** (hours)	**DT Sensitivity** (weighted avg.,%)	**NBTree Sensitivity** (weighted avg., %)	**Diff.** (%)
A	1	70.97	75.43	4.46
	2	61.94	68.26	6.32
	3	59.66	64.95	5.30
	4	58.97	62.33	3.36
	5	60.57	60.39	-0.18
B	1	82.63	82.65	0.02
	2	72.57	76.60	4.03
	3	71.20	73.86	2.66
	4	69.94	70.78	0.83
	5	69.37	69.52	0.15
C	1	93.03	95.21	2.18
	2	90.97	94.06	3.09
	3	89.83	92.12	2.29
	4	90.74	92.12	1.38
	5	90.86	91.55	0.70

Table 4.5: The classification sensitivity values of *DT* and *NBTree*.

The data of numbers of ambulances were obtained through queue reconstruction and simple calculation. Furthermore, other predictor variables, such as day of week, hour of day, and NEDOCS score, were kept consistent with the historical data. A total of additional 20 years of data were generated and used to train the decision tree models. For each year's data, 30 decision tree models were constructed in the same way as detailed in Section 4.3. The result is reported in the Figure 4.3. The standard errors of the prediction accuracy are also shown in the figure as error bars. According to these values, the hybrid decision tree algorithm (*NBTree*) shows consistent improvement on prediction accuracy in these synthetic data sets, comparing to the CART decision tree algorithm (*DT)*. Therefore, this method is confirmed to be suitable to analyze the data in this application as a preferred algorithm.

4.4.2 Case Study

The motivation for this study is to provide EHS personnel with a decision-support model that can predict AOD problems in advance allowing management to activate

Model	**Prediction Period** (hours)	**DT Specificity** (weighted avg.,%)	**NBTree Specificity** (weighted avg., %)	**Diff.** (%)
A	1	88.11	90.04	1.93
	2	84.36	87.95	3.59
	3	83.95	86.07	2.12
	4	83.29	85.08	1.80
	5	84.31	84.76	0.45
B	1	91.65	91.22	-0.62
	2	87.79	88.63	0.84
	3	87.22	88.22	1.00
	4	87.16	87.73	0.57
	5	87.31	86.57	-0.73
C	1	71.27	74.41	3.14
	2	61.87	72.18	10.31
	3	57.95	68.24	10.29
	4	50.84	58.89	8.05
	5	85.80	66.79	-19.01

Table 4.6: The classification specificity values of *DT* and *NBTree*.

proactive interventions. For this study, different models with various prediction focuses are developed (Model **A**, **B**, and **C**), as well as for different prediction periods. Given the nature of the EMS system under study, it represents a good trade-off between the accuracy of the prediction model and its practical purpose. In this section, we demonstrate an example of the results of a prediction model summarized in a table format. We selected a prediction model built by the hybrid decision tree algorithm, as this method generally generates models with improved performances, while still maintains the easy-to-interpret tree structures.

The Model **B** with a prediction period of $X = 4$ hours is chosen for the case study. This model should be able to provide a relatively accurate (approx. 70%) prediction of AOD in four hours, while reserving enough time for the EMS personnel to put interventions in action to be proactive. The results of the prediction model are summarized in Table 4.7, with only the predictor variables present in the final decision tree structure and its predictions. Each row of the table represents a scenario of the system. For example, row 1 in the table suggests that the model predicts the AOD states will continue to be *Good* (0-3 ambulances at the ED) in 4 hours if the current number of ambulances at the ED is between 0 and 3 and the emergency calls

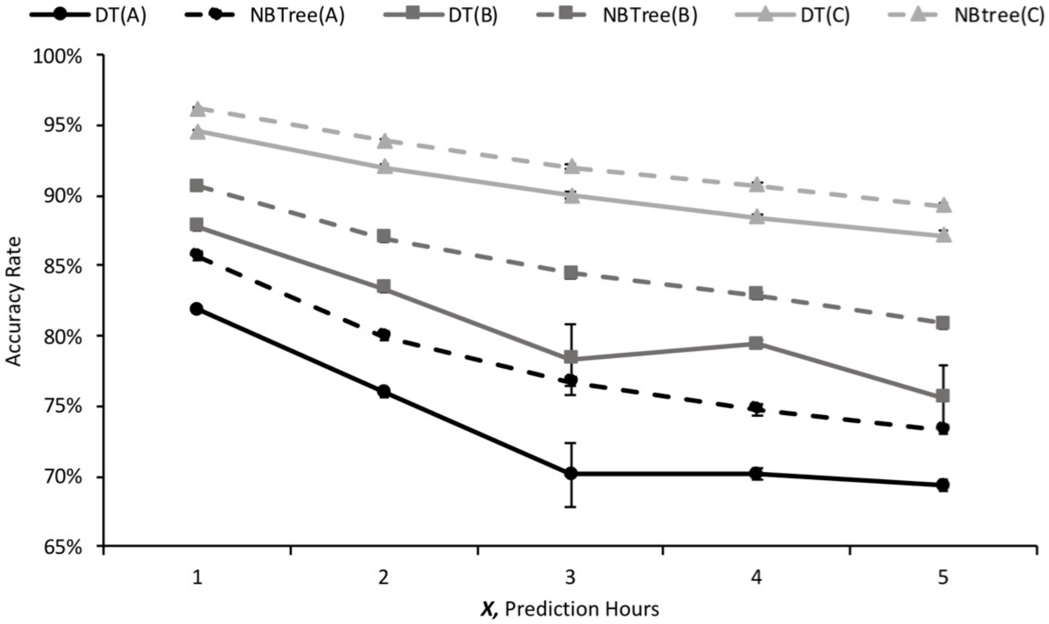


Figure 4.3: The comparison of the classification accuracy rates of *DT* and *NBTree* using 20-years of synthetic data.

received in the last hour are less than 7. However, if the call volume was greater than 7, the AOD state will deteriorate to be *Bad* (4-6 ambulances at the ED) (according to row 4 in the table).

This model suggests that the number of ambulance at EDs changes modestly during a four-hour period. The predictor variable, the number of ambulances currently at the ED, has a dominate impact on the prediction of AOD in the system in four hours. Number of calls (EHS received), the NEDOCS of the ED, and the hour of day are also important variables for the prediction. The rest of the predictor variables, ambulance clear rate at the ED and the day of week, show minor or no impact on the prediction of AOD in this prediction model.

The model can predict the AOD states of the system relatively well across different classes, as shown in the model's confusion matrix (Figure 4.4). In most cases, the AOD states (class) stays the same over a 4-hour period; while occasionally, the classification changes by one class but never by two. This observation is consistent throughout all the prediction models in this study.

Number of Ambulances at the ED	Number of Calls in last hour	NEDOCS	Hour of Day	Clear Rate	Prediction Class
0-3	0-6	-	-	-	Good (0-3)
4	0-3	Busy, Very Busy, Overcrowded, Dangerous	-	0-3	Good (0-3)
4	0-3	Busy, Very Busy, Overcrowded, Dangerous	0a.m. - 7a.m.	≥ 4	Good (0-3)
0-3	≥ 7	-	-	-	Bad (4-6)
4	0-3	Disaster	7a.m.-11p.m.	-	Bad (4-6)
4	0-3	Busy, Very Busy, Overcrowded, Dangerous	8a.m.-11p.m.	≥ 4	Bad (4-6)
4	≥ 4	-	-	-	Bad (4-6)
5-6	0-2	-	-	-	Bad (4-6)
5	≥ 4	-	-	-	Bad (4-6)
6	3-4	Busy, Very Busy, Overcrowded	-	-	Bad (4-6)
7	-	Busy, Very Busy	0a.m.-6a.m.	-	Bad (4-6)
7	0	Overcrowded	0a.m.-4a.m.	-	Bad (4-6)
6	≥ 3	Dangerous, Disaster	-	-	Problematic (≥ 7)
6	≥ 5	Busy, Very Busy, Overcrowded	-	-	Problematic (≥ 7)
7	0	Dangerous, Disaster	0a.m.-4a.m.	-	Problematic (≥ 7)
7	0	Not Busy, Overcrowded, Dangerous, Disaster	5a.m.-6a.m.	-	Problematic (≥ 7)
7	≥ 1	Not Busy, Overcrowded, Dangerous, Disaster	0a.m.-6a.m.	-	Problematic (≥ 7)
7	-	-	7a.m.-11p.m.	-	Problematic (≥ 7)
≥ 8	-	-	-	-	Problematic (≥ 7)

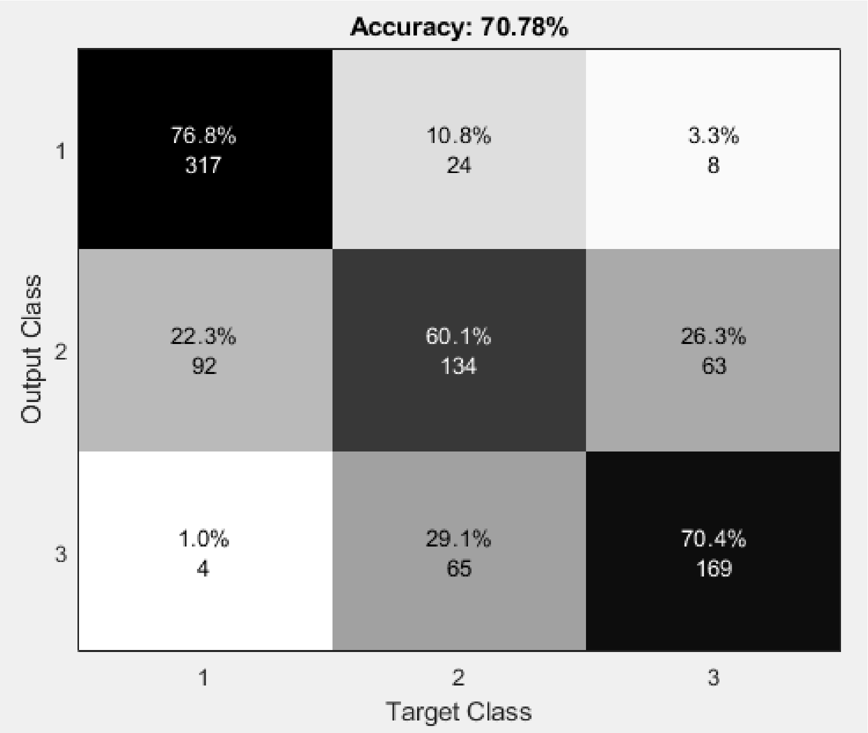
Table 4.7: The results of the hybrid decision tree model for predicting three different AOD states (good, bad, and problematic) in four hours.

4.5 Discussion

The worsening of AOD states does not happen suddenly allows EMS personnel to act proactively to avoid worst cases. With the prediction model presented in this study, they will have knowledge on the expected AOD state ahead of time. If the situation is predicted to be worse, certain operations can be activated. For example, establishing communications with paramedics and ED staffs to expedite the offload process, redirecting ambulances to less busy EDs, reallocating ambulances, etc.

AOD can be costly to an EMS system. Take EHS (in Nova Scotia, Canada) as an example, our preliminary analysis of AOD in year 2016 indicates that the time that ambulances spend offloading patients increases 7 minutes per vehicle on average with every additional ambulance added to the AOD queue. This can lead to significant loss of ambulance hours when considering multiple vehicles in AOD at multiple sites, let alone concerns regarding patient safety and quality of care. The prediction model does not directly mitigate AOD, but it provides a forecast on AOD which offers the potential to initiate practices that may help prevent AOD from worsening.

The demonstration of a table-format prediction rules (such as shown in Table 4.7) can be extended to all 30 prediction models generated in this study. Each tables offers a set of easy-to-understand rules for AOD predictions based on different prediction focuses and periods. EMS providers have the flexibility to choose the most suitable



	1	2	3
1	76.8% 317	10.8% 24	3.3% 8
2	22.3% 92	60.1% 134	26.3% 63
3	1.0% 4	29.1% 65	70.4% 169

Figure 4.4: The confusion matrix of the prediction model in the example (Model **B** with a prediction period of $X = 4$ hours.

models for their daily operations. For example, when predicting 1-2 hours into the future, it may be more beneficial to have more detailed information to plan the immediate ambulance operation. In which case, they may follow the results from Model **A**, where the most refined classifications are provided with relatively accurate predictions (approx. 70%). However, if the goal is to only predict the worst AOD states in advance so that there is enough time to initiate proactive actions to mitigate (e.g., schedule additional ambulances, communicate with hospital EDs, etc.), it may make more sense to refer to Model **C**, where the focus is to predict such AOD states with accuracy (above 90%) for a long prediction period (up to 5 hours in this study).

The *NBTree* algorithm provides an easy-to-interpret tree structured model with improved prediction accuracy comparing to traditional DT algorithm. However, the improvement of the model performances such as accuracy and other characteristics is not as remarkable as expected in this study. It may be due to the fact that the EMS system is a complex and stochastic system and the available data for the prediction model are limited. Some potential contributors are not included in the model (such as weather, traffic, holidays/events, schedules, staffing, other ambulance activities, etc.). These data are either not available to us or they are not collected, which makes it difficult to develop a model with greater prediction power. Some of these missing predictor variables may have great impacts on the model performance. Without them, some characteristics of the training observations are overlooked, leading to a reduction of prediction accuracy. These missing factors can not be compensated by reducing noises in the available data, which is what the *NBTree* algorithm essentially does. Therefore, if an AOD prediction model with higher accuracy is desired, more predictor variables would need to be added to future models.

4.6 Conclusion

In this study, we developed a framework to predict the ambulance offload delay states at an ED based on the current state of the EMS system. We have adapted a hybrid decision tree algorithm that uses a Naïve Bayes classifier to remove the noisy training observations before the decision tree induction. In this study, the performances of the model generated by this algorithm showed improvements on the classification accuracy rates in most cases. Improvements were also found in the classification precision, sensitivity and specificity analysis.

No significant change of the AOD states of the system appears in any prediction models in this study. This implies that the AOD states of the system may be robust and any variables that can cause a significant change of the state may take more than several hours to be reflected to the system. From the other perspective, it may also be difficult for any mitigation intervention to improve the AOD state in a short period of time.

Both the EMS and the hospital EDs are complex health care systems with random demands. As a problem occurring at the interface of these two systems, AOD involves different aspects of these systems, and can be affected by many factors. Therefore, significant variances can be expected from the real-world historical data. These variances can significantly impact the accuracy of a model that predicts the AOD states of the system. We selected six to be the predictor variables of the model in this study, based on our knowledge to the AOD problem and the availability of the data. By utilizing data of ED congestion and EMS operation from both the EMS provider and the hospital EDs, this study defines the thresholds of the EMS system in terms of AOD for the future development of mitigation interventions. It also provides insights for all involved parties to move from the current reactive practice to proactive response when coping with AOD. This may encourage improved communications and share of information between the two parties and inspire future collaboration on AOD related research projects.

Chapter 5

DETERMINING AMBULANCE DESTINATIONS WHEN FACING OFFLOAD DELAYS USING A MARKOV DECISION PROCESS MODEL

In Chapter 4, we demonstrate a hybrid decision tree approach to develop a prediction model for AOD states at the ED based on the current status of the EMS system. This information has the potential to benefit the operation of EMS if proactive actions are set in place to prevent problematic AOD states. To evaluate this potential benefit for EMS, we identify several mitigation interventions of AOD and investigate a selected intervention in this chapter.

5.1 Introduction

In this chapter, we aim to develop an EMS intervention to cope with AOD, and to measure the success and impact of its potential benefits to the EMS system when implemented. Anecdotally, there have been different interventions implemented by the hospitals and the EMS providers to mitigate the AOD problem, as previously discussed in Chapter 2. EMS and hospital staff may possess innovative intervention ideas and recommendations from their own experiences. Therefore, to mitigate the AOD problem, it is important to consult with the frontline personnel for potential interventions and their feasibilities.

Therefore, two focus group discussions were held on September 28th, 2017 and October 2nd, 2017 with key EHS personnel. In the focus groups, the attendees were asked to brainstorm and list EMS interventions that have (or may) mitigate the AOD problem. The goal of these focus group discussions is to gather intervention ideas regarding the mitigation of AOD, and determine which to further investigate with OR models.

Six paramedic supervisors from the Central Region and one paramedic supervisor from each of the other three regions (Western, Northern, and Eastern) participated in this exercise. As frontline personnel, paramedic supervisors possess a great deal of knowledge and experiences on how to cope with the AOD problem in EHS daily operations, thus, have valuable insights on feasible and creative interventions which may have the potential to mitigate the AOD problem.

Both focus group discussions were well received, and many intervention ideas were discussed. Some interventions had been tried before and some were new ideas. A total of 50 interventions were identifies during the discussions. These ideas are summarized and categorized into seven different themes, as presented in Table 5.1. Readers can find more detailed descriptions on each intervention in Appendix B. The majority of these interventions are EMS focused (three out of the seven themes), including: EMS processes based on patients' conditions, EMS processes based on system status, and general EMS processes. Since the AOD problem occurs at the interface of the EMS systems and hospital EDs, there are two themes focusing on collaborative practices including: offload programs, and communication. Another theme describes hospital interventions and the last one includes interventions that were indirect to mitigate the AOD problem but noteworthy.

Theme	Intervention Idea	Comment
(based on patients' conditions) **EMS Processes**	Extended Care Paramedic (ECP) Program	Expand (provincial)
	Palliative Care Program	Expand
	Bypass ED for patients with certain conditions (trauma, stroke, stemi, etc.)	Expand for more patient types
	Bypass ED for low acuity patients (define the medical necessity for an ambulance)	New idea
	Bypass ED for EMS super users and create special response protocols	New idea
	Direct to Chairs Policy	Continue & formalize with NSHA
(based on system status) **EMS Processes**	Utilize the emergency department information system (EDIS)	Continue
	Ambulance smoothing	Expand in/out of district; provincial
	Grant EHS supervisors' ability to redirect ambulances when see fit	New idea
	EHS communication centre escalation plans	New idea

	Double up	Continue
	Bed Swap	Continue & expand
(general) **EMS Processes**	Triage EMS response times	New idea
	Separate call and transfer service	New idea
	More PTUs hours during peak demand time	New idea
	Refuse ED-to-ED transfers when patient cannot be placed into a bed within a certain amount of time	New idea
	Allow PTUs to perform ED transfers	New idea
	Reduce EHS charting requirements and address work flow issues	New idea
Offload Programs	ED “hallway medicine” for non-complex cases when needed	New idea
	Re-implement offload zones (OZs) at EDs	New idea
	Holding areas for ambulance patients at EDs	New idea
	Discharge lounges for patients at EDs	New idea
	Double EHS team to operate OZs at EDs	New idea
	Independent personnel in charge of placing patients into ED beds	New idea
	Bed swap between ED and OZ beds when the patient in ED bed waiting to be processed	New idea
	Have hospital supervise patients who do not need to be overseen by paramedic	New idea
Communications	Check for ED bed availability to initiate conversation with charge nurse	Continue
	Communication between paramedics and ED staff	Continue
	Direct EHS supervisors and ED charge nurse interaction	Continue
	Direct EHS manager & NSHA director interaction	Continue
	Bring EHS representative to the ED executive table	New idea
	Create visual real-time measures	New idea
	Better define and measure TOC time	New idea
	NSHA access to EHS system status	New idea
	Define areas of responsibility and link that to performance	New idea

	Share patient care plans between EHS/NSHA	New idea
	Communicate best practices to all EHS staff	New idea
	Joint policy development between EHS and NSHA	New idea
Hospital Processes	Add more hours/resources at EDs	Explore
	Redefine concept of "bed count = patient care"	New idea
	Push patients through the system rather than pulling	New idea
	Enhanced ED outflow (early discharge with facilitated follow up)	New idea
	Define hospital escalation plans	Expand
	Separate charge nurse for internal and external processes	New idea
	Make charge nurse easily identifiable to paramedics	New idea
	Have specialized services at different hospital facilities	New idea
	Improve patient triage and registration processes	New idea
Indirect but Noteworthy	Better address AOD in EHS reports	New idea
	Bring food/supplies for paramedics in offload delay (morale)	Continue
	Public awareness of AOD problem	New idea

Table 5.1: The summary of interventions obtained from the two focus group discussions with key EHS personnel.

The intervention chosen for this study is to find the optimal ambulance destination policy to mitigate AOD, which is derived from the EHS current practice "ambulance smoothing". We develop this policy to provide guidelines to EMS on where to transport patients with consideration of AOD, patient acuity level, and travel distance. To generate the policy, we formulate a discrete time, infinite-horizon, discounted Markov Decision Process (MDP) model that determines how to optimally direct ambulances.

We propose two independent objectives: one is to minimize the time that ambulance crews spend transporting patients. The other is to minimize time-to-ED bed for patients. A computational study is conducted with real-world data from an EMS provider which currently experiences AOD regularly.

This study has two main contributions to the AOD and OR literature. First, our paper discusses an MDP approach to find optimal ambulance destination policy for an EMS system which considers AOD. Many systems have used ambulance diversion as a method to counter AOD. Ambulance diversion occurs when an ED restricts incoming ambulance traffic due to crowding, and ambulances are therefore routed elsewhere. The system under study does not use ambulance diversion but instead EMS dispatchers have a tool to route ambulances to less busy EDs to mitigate AOD. Second, it demonstrates a method to incorporate a large amount of real-world data into the MDP model design, and to solve the numerical case for a relatively large EMS region. Although the literature related to EMS systems includes various studies that use MDP models, their designs often use theoretical distributions rather than actual administrative data.

The remaining sections of the paper are organized as follows. Section 5.2 provides a literature review on various OR models applied to the AOD problem and applications of the MDP models to EMS. The proposed MDP model is formulated in Section 5.3. The real-world data used in the computational study are described in Section 5.4. The results for the model applied to this study are presented in Section 5.5. Finally, we include conclusions and discussions in Section 5.6.

5.2 Literature Review

There are two streams of literature that are related to our work. The first stream is the literature on MDPs. MDPs have been widely used to model and solve dynamic decision-making problems with multi-states under stochastic circumstances [180]. It has been applied to many areas including finance, agriculture, logistics, maintenance, manufacturing, and recently in health care [181, 182]. However, to our best knowledge, there is currently no application of MDP models in the AOD literature.

There are numerous MDP applications related to the EMS system, mainly focusing on optimizing the dispatch policy. Bandara et al. [183] examine the optimal dispatch policy within the EMS system while focusing on the urgency level of an emergency call. They develop an MDP model to identify how to optimally dispatch ambulances to maximize patient survivability. Building on this study, there are several papers that considers the ambulance dispatch problem with priority levels of patients [184, 185]. McLay and Mayorga [185] formulate an MDP model to determine the optimal dispatch policy while considering that dispatchers make classification errors in assessing the true customer priorities. To shorten response time for the urgent patients, their model allows increased response times for the non-urgent patients. They later extend the modeling framework [184] and examine the optimal EMS dispatch policy while considering the issue of balancing equity and efficiency. Four vehicles and four demand locations are included in their numerical work [184, 185].

Jarvis [186] addresses the problem of determining the dispatch policy by minimizing the average cost of assignment in an MDP model, while considering that individual vehicle may be unavailable due to previous assignments. Keneally et al. [187] develop an MDP model based on simulation data to examine aerial military medical evacuation dispatch policies in a combat environment. Some papers use MDPs to model the ambulance redeployment problem. Alanis et al. [188] propose and analyze a two-dimensional Markov chain model to identify a near-optimal compliance table policy. Berman [189, 190] uses this approach to examine optimal repositioning of emergency units for small systems. The same method was revisited by Zhang et al. [191] to solve a single-ambulance repositioning problem optimally. These models provide important insights for simplified models involving a few ambulances. Maxwell et al. [192, 193] construct an approximate dynamic programming model to find solutions for larger systems with fewer assumptions, generating optimal or near-optimal repositioning policies.

The second stream of literature related to our work is the literature on the AOD problem. It has only become an active research topic recently and there is limited research from the OR field that specifically focused on this problem [31]. Majedi [32] constructs a system representing the interaction of EMS and ED using queuing theory and models the behavior of the system as a continuous time Markov chain to evaluate various system performance measures (e.g., the average number of ambulances in offload delay, average AOD, and ambulance and bed utilization). Clarey et al. [66] design a discrete event simulation model to assess the change on AOD in a scenario, where dedicated nurses are hired to assist with ambulance offloading patients. This study demonstrates a clear reduction in AOD when dedicated nursing levels are increased. However, the authors also raise concerns that using this as a sole method to reduce AOD would require unacceptably low staff utilization, which would cost hospitals both financially and in human resourcing. Almehdawe et al. [5] uses a Markov chain queuing model to analyze the interface between an EMS provider and multiple EDs that serve both ambulances and walk-in patients. Matrix-analytic methods are used to solve the steady state probability distributions of queue lengths and waiting times. The study concludes that the priority based admitting policy had a great impact on patient waiting times. When additional resources are considered for the system, the benefit of adding capacity is greater for EDs with higher utilization. Almehdawe et al. later [33] introduce a stylized queuing network model with blocking to investigate the effect of patient routing decisions on EMS offload delays and to determine the optimal allocation of patients to each ED in a region.

Two urban hospital EDs in Nova Scotia, Canada have attempted to reduce AOD by implementing an offload zone (OZ), in collaboration with the local EMS provider [63]. This OZ is a holding area in the ED monitored by a dedicated nurse and paramedic team for patients who arrive by ambulance but cannot be admitted into the ED due to congestion. This practice eliminates the need for one ambulance crew (two paramedics) to wait with each patient, and thus frees the ambulance to return to service more quickly [30]. Two years after opening the two OZs, Carter et al. [30] completed a Health Care Failure Mode and Effect Analysis study to identify risks to patient safety and process efficiency. They conclude that the OZ results in ED staff having little incentive to admit patients who are waiting in the OZ and instead admit patients from the waiting room. This leads to the OZ often being at capacity and unable to relieve AOD. Motivated by this unexpected finding, Laan et al. [64] model the OZ using a continuous time Markov chain to investigate how this lack of incentive impacts AOD. The result suggests that, when the probability of "a patient admitted from the OZ when a patient of equal acuteness is waiting in the waiting room" is not greater than a certain threshold (0.35 in their case), implementing an OZ will result in even longer offload delay, as admission priority is disproportionately given to patients in the waiting room.

5.3 Methods

This section presents the MDP model for determining the optimal ambulance destination policy in an EMS system suffering from AOD. When there is no AOD, the destination ED for an ambulance patient is typically the closest ED appropriate for the patient's condition. However, when the number of queued ambulances is high at the urban EDs, it may become more time-efficient for ambulances to travel further distances to a community ED where there are typically no queued ambulances. Currently, ambulance destination decisions within the urban region in our case study system consider ED crowding, the number of queued ambulances, and other measures, but this practice does not extend to surrounding areas. The model described in this section is used to determine when it is advantageous to send patients to the further community EDs, given the number of queueing ambulances, patient acuity, and travel distance. We begin with a short introduction to discrete time, infinite-horizon, discounted MDP models, before presenting our model.

5.3.1 General Overview of MDP Modeling Framework

A discrete time, infinite-horizon, discounted MDP model is characterized by a set of five quantities, expressed as $\langle S, A, T(s,a,s'), R(s,a,s'), \gamma \rangle$ [194], where S is the finite set of all states of the model, A is the finite set of all available actions, $T(s,a,s')$ is the transition probability for reaching state s' when taking action a from state s, $R(s,a,s')$ is the reward function to receive a reward (or penalty) when getting from state s to state s' by taking action a, and γ is the discount factor ($0 < \gamma < 1$) to discount future rewards to the present time. A reward n steps away from the current state s is discounted by γ^n. The discount factor is necessary for the reward function to converge in an infinite horizon MDP model.

A decision rule prescribes a procedure for assigning an action a to each possible state s in S. A policy $\pi(s)$ is a sequence of decision rules to be used at all decision epochs. A state-value function $V_\pi(s)$ represents the expected objective value obtained following policy $\pi(s)$ from state s in S. It is defined as the expected value of all future rewards, which is the immediate reward of reaching state s as well as the rewards of subsequent states under the policy $\pi(s)$.

$$V_\pi(s) = R(s, \pi(s), s') + \gamma \sum_{s' \in S} T(s, \pi(s), s') V_\pi(s')$$

The action-value function $Q_\pi(s, a)$ is the expected objective value starting from state s, taking action a, while following policy π. It specifies how valuable state s is under the policy $\pi(s)$ for different actions a.

$$Q_\pi(s, a) = R(s, a, s') + \gamma \sum_{s' \in S} T(s, a, s') V_\pi(s')$$

The MDP algorithms are aimed at calculating or estimating value functions to determine useful actions and find the optimal policy. Solving an MDP over an infinite horizon results in deriving an optimal policy $\pi^*(s)$. It is defined as the policy which maximizes the expected reward (value) for each state with the discount factor, γ (0 $< \gamma <$ 1). Thus, if we denote the maximal value of the action-value function as

$$Q^*(s, a) = \max_\pi Q_\pi(s, a),$$

the optimal policy is the policy that maximizes the expected reward,

$$\pi^*(s) = \arg\max_\pi Q_\pi(s, a).$$

When the state and action spaces have finite cardinalities, the optimal policy takes on a stationary form as there is no reason to behave differently in the same state at different times, no matter how long the agent has run or will run in the future.

Several standard algorithms are available to compute the optimal policy $\pi^*(s)$ with total expected discounted rewards. These methods are linear programming, the policy iteration algorithm, and the value iteration algorithm [182, 194, 195]. We choose the policy iteration algorithm to solve our MDP model in this study. The algorithm and our motivation for this choice are detailed in Section 5.3.3.

5.3.2 Our Model

Each decision epoch represents a new ambulance call requiring a patient transportation and an ambulance destination decision: with probability $\mathbb{P}(B = b)$ the call is for a patient type b ($b = \{R, H, M, L\}$ for resuscitation, high, medium, and low acuity levels), with probability $\mathbb{P}(D_u = d_u, D_c = d_c)$ the call is d_u kilometers away from the closest urban ED, and d_c kilometers away from the closest community ED, and with $\mathbb{P}(N = n)$ there are n ambulances queued at the urban EDs when the destination decision is to be made. When taking the action of transporting a patient to an urban ED, n will increase by 1 if no ambulances are released between decision epochs. When taking the action of transporting a patient to a community ED, n does not increase. For both actions, n decreases by d with $\mathbb{P}(D = d)$ where D is the number of ambulances released from the urban EDs between decision epochs. We therefore define the state space as $S = S_N,\ S_{D_u,D_c},\ S_B$ where S_N is the state representing the number of ambulances queuing at the urban EDs, S_{D_u,D_c} is the call location state defined by a pair of travel distances to the closest urban ED and the closest community ED respectively, and S_B is the patient acuity state.

Actions

The decision at hand is to determine to which ED to send the patient. Rather than considering each ED individually, we aggregate the EDs within and outside the urban region into two groups as urban EDs and community EDs, respectively. This way, the decision is whether to send the patient to *an* urban ED or *a* community ED. The model is formulated such that only the closest of each type are considered as possible destinations. This is further appropriate because the urban EDs, although not located at the same place, share a virtual queue when busy. In other words, dispatchers consider the number of queued ambulances and ED crowding when determining the destination ED in the urban region. There is only one of two actions that may be taken when making the decision.

$$A = \begin{cases} \text{transport patient to an urban ED, } k = 1 \\ \text{transport patient to a community ED, } k = 2 \end{cases}$$

Penalty functions

Action $a_k \in A$ is chosen when the process is in state s, and the process then makes a transition to state s', and an immediate penalty is assigned. The penalty reflects the change of value to the objective function of the action selected. In this study, we consider two different penalty functions independently. The first is the time that an ambulance crew spends transporting a patient, including the time to return to the urban region (if necessary), $r^{a_k}_{Am}$, and the second is the time to an ED bed for a patient $r^{a_k}_{Pt}$.

The penalty function contains three time components: the inbound travel time (the time that an ambulance travels to transport a patient from the call location to an ED), the turnaround time (the time that the ambulance crews spends at an ED waiting to transfer the care of the patient to the ED staff, time for clean up or paperwork, and recovery time), and the outbound travel time (the time that an ambulance travels to their next posting location). In practice, the ambulance is actually "in service" during the outbound travel time and can be called upon to respond to a call in an EMS system with dynamic deployment. Despite this, we penalize by the outbound travel time because the ambulance is not in the urban region during this time. This ensures the consequence of sending an ambulance out of the urban region is completely accounted for in the model. In fact, in systems with dynamic redeployment it is possible that an ambulance crew sent to a community ED would remain in that community after delivering the patient. We choose to ignore this in the model and instead focusing on the urban system, which means we may be slightly over penalizing ambulances sent to community EDs. Let T^{in}_u and T^{out}_u respectively be the inbound and outbound travel time for patient being transported to an urban ED. Similarly, T^{in}_c and T^{out}_c are respectively the inbound and outbound travel time for patients being transported to a community ED. Let T^n_u be the turnaround time experienced by an ambulance at an urban ED when there are n ambulances in the queue.

To compute the time to ED bed for a patient, the turnaround time interval requires more explanation. “The turnaround time starts when the paramedics report to the dispatcher that they have arrived at the ED, and ends when the dispatcher is notified that the paramedics are available for another call” [7]. This is made up of multiple sub-intervals including the delivery or offload interval, defined as arrival at ED time to transfer of care time, and the recovery interval, defined as transfer of care time to clear at ED time (i.e. time for paramedics to recover after delivering the patient). We define our turnaround time variable in the urban ED such that T^n_u is the complete interval, T^0_u is the sum of all time sub-intervals except the queueing time sub-interval, and T^r_u is the length of time in the recovery sub-interval. At the community EDs, ambulance queueing is negligible. Therefore, the complete turnaround time interval is modeled with T_c^0 and the recovery time interval is T_c^r. The time to ED bed for patients is the sum of the inbound travel time and part of the turnaround time excluding the paramedic's recovery interval. At the urban ED, this is $T_u^{in} + T_u^n - T_u^r$; while at the community ED, this is $T_c^{in} + T_c^0 - T_c^r$. Figure 5.1 demonstrates the different time intervals (not necessarily to scale) included in our penalty functions and the difference between $r_{Am}^{a_k}$ and $r_{Pt}^{a_k}$.

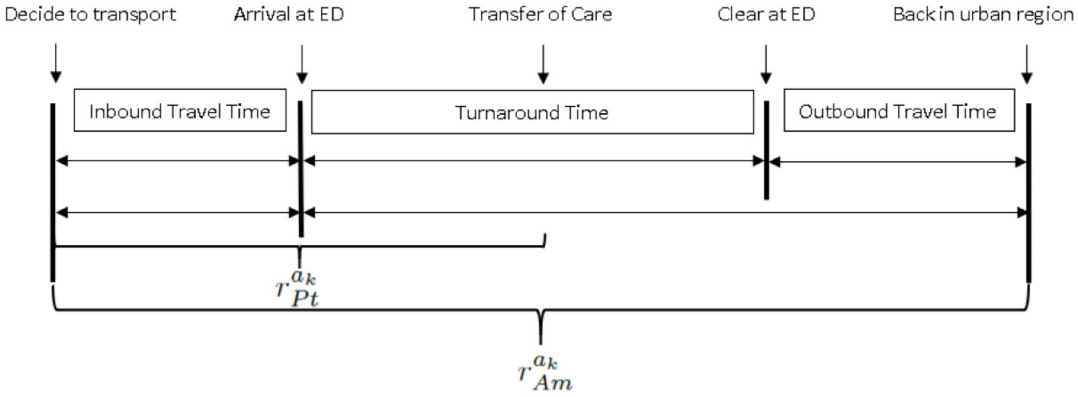


Figure 5.1: The ambulance operation processes included in the penalty functions $r_{Am}^{a_k}$ and $r_{Pt}^{a_k}$.

For each patient acuity level, different ED destinations are appropriate, and this is reflected in the penalty functions. Major trauma patients (resuscitation acuity) will always be treated immediately regardless of the status of the ED. Therefore they do not queue and the turnaround time for the ambulance is T_u^0. Low acuity patients, in this study, are defined as ambulance patients who can be safely offloaded directly to the ED waiting room (based on clinical impressions according to local EMS policy). Therefore, they also do not queue upon arrival. As such, the turnaround time for an ambulance with a low acuity patient is also T_u^0. High acuity patients will also be transported to the closest appropriate ED due to their severe conditions, however, they may be delayed at the ED due to AOD. Therefore, the turnaround time for an ambulance with a high acuity patient is T_u^n. Medium acuity patients are candidates for transporting to a community ED because their acuity allows it and they may experience delays at an urban ED waiting to be offloaded.

Minimize the ambulance transportation time

The penalty function of the ambulance transportation time, $r^{a_k}_{Am}$ can be determined as follows:

$$
r^{a_k}_{Am} = \begin{cases} T^{in}_u + T^0_u + T^{out}_u, & if\ k = 1, b = R\ or\ L \qquad (1) \\ T^{in}_u + T^n_u + T^{out}_u, & if\ k = 1, b = H\ or\ M \qquad (2) \\ Z, & if\ k = 2, b = R\ or\ H\ or\ L(3) \\ T^{in}_c + T^0_c + T^{out}_c, & if\ k = 2, b = M \qquad (4) \end{cases}
$$

where Eq.(1) and (2) describe the situation of sending an ambulance patient to an urban ED, while Eq.(3) and (4) describe the situation of sending an ambulance patient to a community ED. We denote Z as a large enough number to penalize the model from sending resuscitation, high or low acuity patients to a community ED, as shown in Eq.(3).

Minimize the time to ED bed for ambulance patients

The penalty function for the time to ED bed for ambulance patients, $r^{a_k}_{Pt}$, is constructed very similarly, except that it includes only part of the turnaround time and does not include the ambulance outbound travel time.

$$
r^{a_k}_{Pt} = \begin{cases} T^{in}_u + T^0_u - T^r_u, & if\ k = 1, b = R\ or\ L \\ T^{in}_u + T^n_u - T^r_u, & if\ k = 1, b = H\ or\ M \\ Z, & if\ k = 2, b = R\ or\ H\ or\ L \\ T^{in}_c + T^0_c - T^r_c, & if\ k = 2, b = M \end{cases}
$$

Both objectives are minimization problems. By converting time to negative values, the penalty functions are converted to a maximization function.

Transition probabilities

Denote the probability of moving from state s to s' given that action a_k is chosen, as $\mathbb{P}(s, a_k, s')$. Then, $\mathbb{P}(s, a_k, s')$ can be defined as:

$$
\mathbb{P}(s, a_k, s') = \mathbb{P}(N' = n | N = n, a_k, D = d) \times \mathbb{P}(B = b) \times \mathbb{P}(D_u = d_u, D_c = d_c)
$$

where $\mathbb{P}(N' = n | N = n, a_k, D = d)$ is the probability $N' = n$ in state s', given $N = n$ in state s. Note that $\mathbb{P}(B = b)$ and $\mathbb{P}(D_u = d_u, D_c = d_c)$ are independent of the system state and action and are determined from historical data.

When the action a_k is to send a patient to an urban ED ($k = 1$):

$$\mathbb{P}(N' = n|a_1) = \begin{cases} \mathbb{P}(D = 0)/\sum\limits_{d=0}^{n} \mathbb{P}(D = d), & when\ N = n - 1, n > 0 \\ \mathbb{P}(D = 1)/\sum\limits_{d=0}^{n+1} \mathbb{P}(D = d), & when\ N = n \\ \mathbb{P}(D = 2)/\sum\limits_{d=0}^{n+2} \mathbb{P}(D = d), & when\ N = n + 1 \\ \dots \\ \mathbb{P}(D = c - n + 1)/\sum\limits_{d=0}^{c+1} \mathbb{P}(D = d), & when\ N = c. \\ 0, & otherwise \end{cases}$$

where c is the total number of ambulances operating in the city area. For instance, when three ambulances are queueing at an ED ($N = 3$), the future S_N states can only be 4, 3, 2, and 1 with the first corresponding to no discharges ($D = 0$) and the latter corresponding to 1, 2, and 3 discharges respectively. The probability of each corresponding future S_N state depends on random variable D. Continuing with this example, $N' = N + 1$ when none of the 4 ambulances (the 3 existing plus the 1 newly arriving) are discharged between calls. Therefore, this occurs with $\mathbb{P}(D = 0|N' = 4, N = 3)$ and is computed by dividing $\mathbb{P}(D = 0)$ by the sum of all possible departure probabilities, in this case $\sum_{d=0}^{4} \mathbb{P}(D = d)$.

Similarly, when the action a_k is to send a patient to a community ED ($k = 2$):

$$\mathbb{P}(N' = n|a_2) = \begin{cases} \mathbb{P}(D = 0)/\sum\limits_{d=0}^{n} \mathbb{P}(D = d), & when\ N = n \\ \mathbb{P}(D = 1)/\sum\limits_{d=0}^{n+1} \mathbb{P}(D = d), & when\ N = n + 1 \\ \dots \\ \mathbb{P}(D = c - n + 1)/\sum\limits_{d=0}^{c} \mathbb{P}(D = d), & when\ N = c. \\ 0, & otherwise \end{cases}$$

Assumptions

The model reflects the typical EMS practice when responding to calls and delivering patients to the most appropriate ED. However, a number of assumptions are implicit in the model. First, EDs in this EMS network are categorized into two groups (urban and community) instead of individual EDs to avoid the "curse of dimensionality". This is a reasonable assumption in this study because of the following two reasons. 1) All three urban EDs share a virtual queue. EMS dispatchers know the level of congestion at these EDs and distribute ambulances accordingly. It is therefore unlikely for one of the EDs to be overwhelmingly busy while the other two are not busy. 2) For the three potential community EDs, the distances from the urban area to them are relatively similar, resulting in no significant difference in transportation time to any community ED. Furthermore, since community EDs experience low patient volumes and rarely experience AOD, it is reasonable to assume no ambulance queueing (AOD) at community hospitals in our model. Therefore, we can treat these community EDs as a group due to these similarities. The potential impact and appropriateness of the assumption of no AOD at community EDs are discussed with more details later in Section 5.5.

Second, the historical call volume data (detailed in Section 5.4) shows very little effect on day of week or seasonality, therefore, they are ignored in our model. Alternatively, call volumes is not stationary with respect to time of day. Two levels can be observed: 1). high level (busy hours from 9 a.m. to 7 p.m.) and 2). low level (non-busy hours from 8 p.m. to 8 a.m.). This feature is ignored in the basic model. In Section 5.5.2, two MDP models are solved to evaluate the optimal policies with the high/low levels of calls found at different times of the day. The analysis of the historical data shows that there is no significant differences of the probability distributions of patient acuity or call locations through out a day. Therefore, the probability distributions of patient acuity and call locations are considered time invariant. Other minor assumptions include, one ambulance is assigned to each patient call and calls occur sequentially. This of course ignores multiple patient calls which are uncommon.

5.3.3 Policy Iteration Algorithm

Policy Iteration is a fundamental algorithm in the study of MDPs. It manipulates the policy directly to find the optimal policy. It starts by evaluating an initial policy, and then uses the value function of that policy to find better policies. This is done by considering taking an action a in state s that is different from the one according to $\pi(s)$. If this change results in a better new policy (that selecting a in s and thereafter following the existing policy), we have successfully improved the policy. Once a policy

$\pi(s)$ has been improved using $V_\pi(s)$ to yield a better policy $\pi'(s)$, we can compute its value function $V_{\pi'}(s)$ and improve it again to yield an even better policy $\pi''(s)$. This procedure is repeated to consider all actions in all states, evaluate each action in each state and select the actions that yield the highest rewards.

This algorithm alternates between two steps, which are outlined:

```
Initialization: choose an initial policy
Repeat until policy is stable {
1.Policy evaluation
    Repeat until values converge {
      For each state {
        Calculate the value function when taking action according to the current policy;
        Update estimate of the optimal value function.
      } each state
    } value convergence
2.Policy improvement
    Find a new policy according to equation
        $\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_{\pi_i}(s')\right].$
}policy stable.
```

The state of the MDP is finite and therefore the number of possible stationary deterministic policies is also finite. The policy iteration algorithm is able to compute an optimal stationary policy in this situation. It is chosen to be used in this study as it is generally faster and less computationally heavy compared to the Value Iteration algorithm [195].

5.4 Data

This section demonstrates how to apply this MDP model to a relatively large EMS region. It describes the key procedures and methods to abstract required model parameters from the data. It also clarifies some situations and considerations that may be unique to the specific EMS provider in this study.

This EMS provider handles all emergency/urgent calls in Nova Scotia, Canada. This study includes emergency/urgent calls in the Halifax Regional Municipality (HRM) (i.e. the urban region) and its adjacent regions (i.e. the surrounding communities). AOD is commonly reported as a problem in HRM [30]. We analyze 12 months of computer-aided dispatch (CAD) data from January 1$^{\text{st}}$, 2016 to December 31$^{\text{st}}$, 2016. The dataset contains 22,243 EMS emergency calls originating in the urban area that are associated with a patient transportation to an ED. Each call record includes information about the patient acuity level and all the time stamps of the ambulance responses for completing that call, such as call time, arrival at scene time, departure from scene time, arrival at hospital time, leaving from hospital time, etc. With a geographic information system (GIS) tracker available for each ambulance, the locations of each ambulance in real time is also recorded and available for this study.

We first restructure the ambulance patient transportation events (new call / arrival at ED / departure from ED) in chronological order. From this we can compute $\mathbb{P}(D = d)$ and the turnaround times (and subintervals) as a function of the number of ambulances at urban EDs n. In other words, when an ambulance arrives at an ED, we can estimate how long the turnaround time will be given the number of ambulances queueing from the historical data. Since we observe that the number of queued ambulances rarely exceeded 9 in our data (763 out of 22,243 incidents, approximately 3.43%), the S_N state space is truncated to be $S_N = \{0, 1, \cdots, 9\}$ in the computational study, where at least 100 calls are recorded in the historical data for each state of S_N. The truncated historical distribution of $\mathbb{P}(D = d)$ is shown in Figure 5.2.

Each patient who requires transportation has a Canadian Triage and Acuity Scale (CTAS) score assigned. The CTAS score ranks the patients by severity from 1 to 5 (1 being highest acuity). It is known by the time that the paramedics evaluate the patient and make the decision to transport the patient to an ED. The probability distribution of CTAS scores is determined from historical data. In this study, we categorize patients who require a transportation into four acuity levels: resuscitation, high, medium, and low. These categories are based on the CTAS scores and discussions with content experts. Resuscitation acuity level patients include all CTAS score

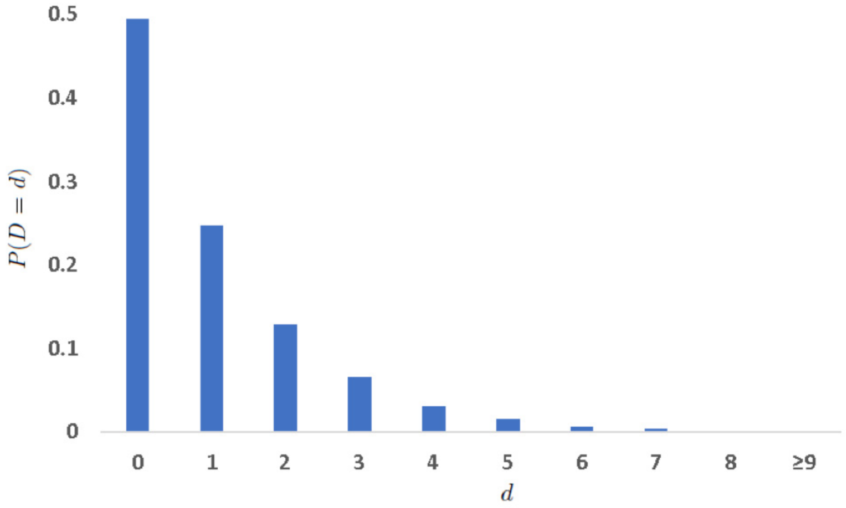


Figure 5.2: The historical distribution of the probability of d numbers of ambulances being released from the urban EDs between decision epochs.

1 patients plus 20% of CTAS score 2 patients. High acuity level patients include the rest of the 80% CTAS score 2 patients. Low acuity level patients include 50% CTAS score 4-5 patients, who can be direct offloaded to an ED waiting room based on local EMS policy. The rest of the patients are categorized as the medium acuity level patients, including all CTAS 3 patients plus the rest of the 50% CTAS score 4-5 patients. According to the 12-month historical data, the probabilities of patient acuity levels is 0.0975, 0.3233, 0.5131, and 0.0661 for resuscitation, high, medium, and low, respectively. This is an appropriate distribution based on our conversations with paramedics and supervisors. In other words, 51.31% of ambulance patients (medium acuity) are candidates to be sent to a community ED in our model.

To compute the penalty function, we need the turnaround time and the inbound/outbound travel time of the ambulances. Based on the local government benchmark with minimal offload time [196], we define the standard turnaround time without AOD T_u^0 and T_c^0 as 30 minutes, and the recovery intervals T_u^r and T_c^r as 20 minutes. We obtain the T_u^n values from the historical data as shown in Table 5.2.

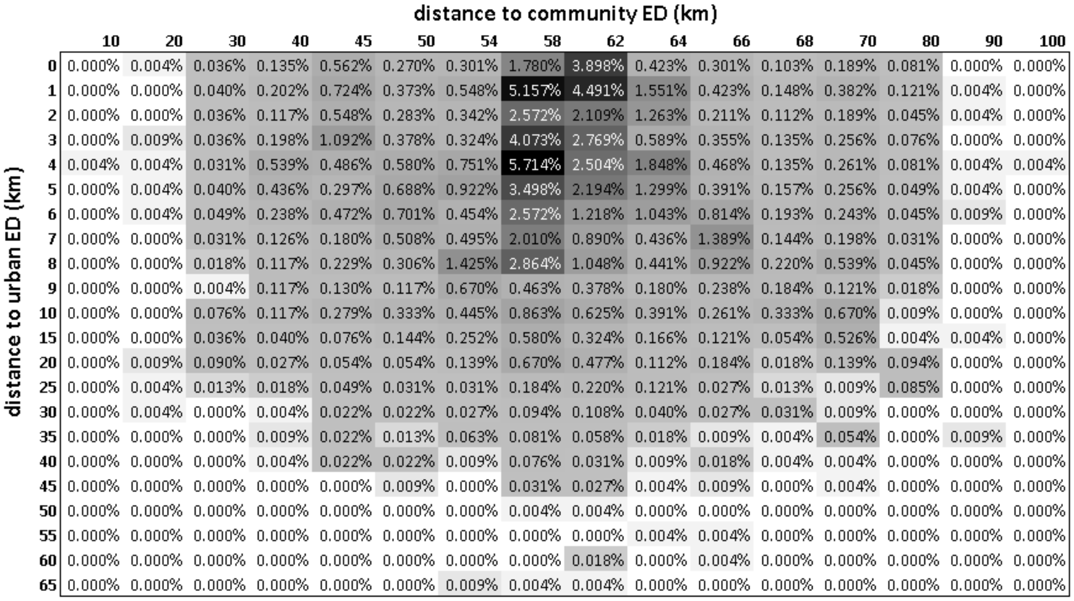
We use ArcMap® v10.5 to find the travel distance to the closest urban ED and

Number of ambulances at urban EDs, N	Average ambulance turnaround time, T_u^n (minutes)
0	30
1	49
2	56
3	63
4	70
5	78
6	82
7	83
8	92
≥ 9	93

Table 5.2: The summary of the average ambulance turnaround time at the urban EDs from the historical data of 2016.

the closest community ED, respectively, from each call in the data set. There are three urban EDs and three community EDs included in the computational study. We separate the distance to the closest urban ED into 22 bins and the distance to the closest community ED into 16 bins by analyzing the histogram of the data. Based on the data frequency, the range in each of the bins varies from 1 to 10 kilometers. A heat map (22×16) is generated to show the call frequencies of each location pair (Figure 5.3). To reduce the problem size, only the call locations (D_u, D_c) with a positive probability (greater than 0) from the historical data are included in the distance-state space matrix (S_{D_u,D_c}). After excluding the cells with probability equal to 0 (shown as the lightest cells in Figure 5.3), the number of categories is reduced to 230. When the model is generalized for other EMS systems, this value is subjected to change. This approach balances the amount of detail in the system representation (which typically results in a better outcome) with the computational difficulties.

The travel time varies based on the travel distance between the call location and the destination ED, and the speed at which the ambulance travels. To calculate the corresponding travel time, we need to estimate the ambulance travel speed. A common modeling practice employs a constant speed for simplicity purpose, which may result in overestimation of EMS system performance [197]. Therefore, in this study, we utilize the KWH model proposed by Kolesar et al. [198] to estimate the ambulance travel time instead. This model has been further validated by Budge et al. [197] and reported to be a reasonable approximation of the median travel time of ambulances. Furthermore, it distinguishes between short and long travel distances which is particularly useful in our study.



	distance to community ED (km)															
distance to urban ED (km)	10	20	30	40	45	50	54	58	62	64	66	68	70	80	90	100
0	0.000%	0.004%	0.036%	0.135%	0.562%	0.270%	0.301%	1.780%	3.898%	0.423%	0.301%	0.103%	0.189%	0.081%	0.000%	0.000%
1	0.000%	0.000%	0.040%	0.202%	0.724%	0.373%	0.548%	5.157%	4.491%	1.551%	0.423%	0.148%	0.382%	0.121%	0.004%	0.000%
2	0.000%	0.000%	0.036%	0.117%	0.548%	0.283%	0.342%	2.572%	2.109%	1.263%	0.211%	0.112%	0.189%	0.045%	0.004%	0.000%
3	0.000%	0.009%	0.036%	0.198%	1.092%	0.378%	0.324%	4.073%	2.769%	0.589%	0.355%	0.135%	0.256%	0.076%	0.000%	0.000%
4	0.004%	0.004%	0.031%	0.539%	0.486%	0.580%	0.751%	5.714%	2.504%	1.848%	0.468%	0.135%	0.261%	0.081%	0.004%	0.004%
5	0.000%	0.004%	0.040%	0.436%	0.297%	0.688%	0.922%	3.498%	2.194%	1.299%	0.391%	0.157%	0.256%	0.049%	0.004%	0.000%
6	0.000%	0.004%	0.049%	0.238%	0.472%	0.701%	0.454%	2.572%	1.218%	1.043%	0.814%	0.193%	0.243%	0.045%	0.009%	0.000%
7	0.000%	0.000%	0.031%	0.126%	0.180%	0.508%	0.495%	2.010%	0.890%	0.436%	1.389%	0.144%	0.198%	0.031%	0.000%	0.000%
8	0.000%	0.000%	0.018%	0.117%	0.229%	0.306%	1.425%	2.864%	1.048%	0.441%	0.922%	0.220%	0.539%	0.045%	0.000%	0.000%
9	0.000%	0.000%	0.004%	0.117%	0.130%	0.117%	0.670%	0.463%	0.378%	0.180%	0.238%	0.184%	0.121%	0.018%	0.000%	0.000%
10	0.000%	0.000%	0.076%	0.117%	0.279%	0.333%	0.445%	0.863%	0.625%	0.391%	0.261%	0.333%	0.670%	0.009%	0.000%	0.000%
15	0.000%	0.000%	0.036%	0.040%	0.076%	0.144%	0.252%	0.580%	0.324%	0.166%	0.121%	0.054%	0.526%	0.004%	0.004%	0.000%
20	0.000%	0.009%	0.090%	0.027%	0.054%	0.054%	0.139%	0.670%	0.477%	0.112%	0.184%	0.018%	0.139%	0.094%	0.000%	0.000%
25	0.000%	0.004%	0.013%	0.018%	0.049%	0.031%	0.031%	0.184%	0.220%	0.121%	0.027%	0.013%	0.009%	0.085%	0.000%	0.000%
30	0.000%	0.004%	0.000%	0.004%	0.022%	0.022%	0.027%	0.094%	0.108%	0.040%	0.027%	0.031%	0.009%	0.000%	0.000%	0.000%
35	0.000%	0.000%	0.000%	0.009%	0.022%	0.013%	0.063%	0.081%	0.058%	0.018%	0.009%	0.004%	0.054%	0.000%	0.009%	0.000%
40	0.000%	0.000%	0.000%	0.004%	0.022%	0.022%	0.009%	0.076%	0.031%	0.009%	0.018%	0.004%	0.004%	0.000%	0.000%	0.000%
45	0.000%	0.000%	0.000%	0.000%	0.000%	0.009%	0.000%	0.031%	0.027%	0.004%	0.009%	0.000%	0.004%	0.000%	0.000%	0.000%
50	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.004%	0.004%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
55	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.004%	0.004%	0.000%	0.000%	0.000%	0.000%	0.000%
60	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.018%	0.000%	0.004%	0.000%	0.000%	0.000%	0.000%	0.000%
65	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.009%	0.004%	0.004%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%

Figure 5.3: The heat map showing the call frequencies of each category of call locations.

The KWH model [198] is defined as follows:

$$T(L) = \begin{cases} 2\sqrt{(L/a)}, & L \leq 2d_a \\ v_c/a + L/v_c, D > 2d_a \end{cases}$$

where T is the estimated travel time, a is the acceleration rate, L is the travel distance, v_c is the cruising velocity, and d_a is the distance required to achieve the cruising velocity ($d_a = v_c^2/2a$).

We analyze the historical data to obtain the proper values of the parameters a and v_c. The actual travel time is calculated by using the time stamps for events of "depart scene" and "arrive destination" of each call. With the travel distances, we can then calculate the ambulance average travel speed during each patient transportation. To eliminate the outliers, only call records with an ambulance travel speed within two standard deviations from the mean travel speeds are kept. This resulted in 1.27% of data points being eliminated. The estimated travel time can be calculated using the KWH model with pre-determined values of a and v_c for each remaining call. By minimizing the sum of squared errors between the actual travel time and the estimated travel time for the 21,527 calls, we determined the value of a in this EMS system is

0.03 m/s^2 and the value of v_c is 85.86 km/h. With the best-fitting parameter values, we then use the KWH model to estimate the travel time (inbound and outbound) in the penalty functions.

T_c^{out} is the travel time from the community ED back to the urban region where the ambulance originated. To computer this we first compute the travel distance from each of the three community EDs to the urban boundary using the road network analyst package in ArcMap® v10.5. We then determine the frequency which each community ED is the closest to a call and use this to compute the weighted average distance as our outbound travel distance (25.02 km) in the KWH model to determine the outbound travel time T_c^{out}. Ambulances leaving the urban EDs are already in the urban region, therefore $T_u^{out} = 0$.

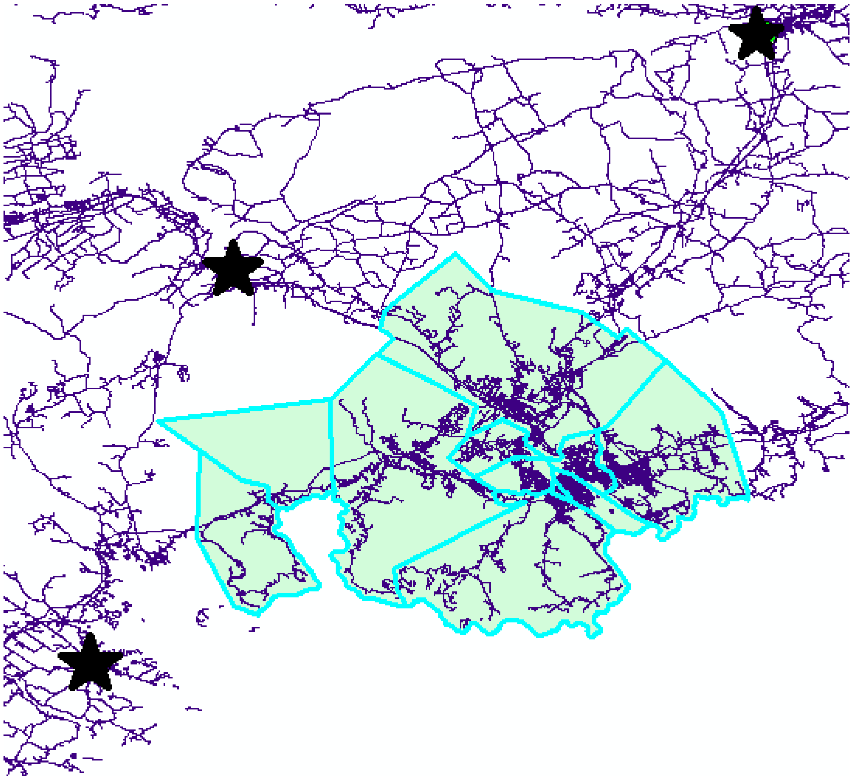


Figure 5.4: The locations of the three community EDs (stars) and the urban region boundaries (highlighted polygons) shown in ArcMap® v10.5.

5.5 Results

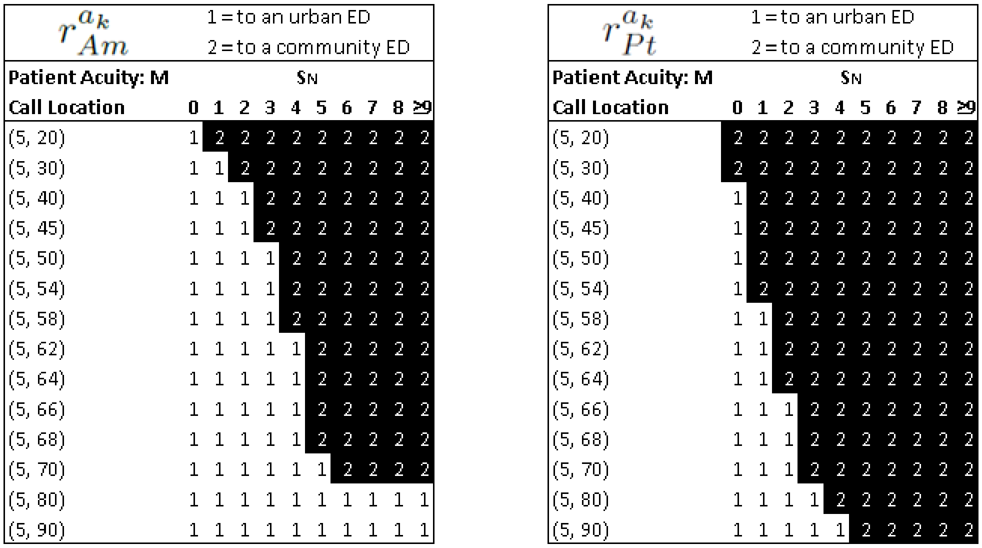
The transition matrices are generated using Microsoft® Excel with a program coded in Visual Basic for Applications. This procedure takes approximate 60 hours to complete, using a Toshiba Portege R30-C computer with an Intel Core i5 processor and 16 GB RAM. The policy iteration algorithm is utilized to solve the optimal policy of the computational study in MATLAB® R2018b. Convergence is reached after approximately 140-170 seconds using the same computer. The output of the MDP is a list of 9200 (10 x 4 x 230) states, and the optimal action associated with each state, for each penalty function, $r_{Am}^{a_k}$ and $r_{Pt}^{a_k}$. Various value of the discount factor γ (from 0.90 to 0.99 with an interval of 0.01) were used to test the robustness of the resulting policies. The results were found to be relatively insensitive to γ. For the remaining sections of the paper, we chose $\gamma = 0.95$ to report the results from the computational study.

5.5.1 Optimal Policies

As expected, both optimal policies suggest sending patients with an acuity level of resuscitation, high, or low to an urban ED. Only patients with a medium acuity level are the candidates for a potential transportation to a community ED. The result indicates that it is not always best to send a medium acuity level patient to the urban ED, especially when many ambulances are already queued there.

We show part of the summary table of each optimal policy in Figure 5.5 as an example to demonstrate the decisions made in some states for the medium acuity level patients. Each table includes 14×10 system states where the travel distance from the call location to the closest urban ED are the same (5 kms), while the travel distance to the closest community ED varies. The decisions from the two optimal policies under different call locations and S_N states are presented in the tables. Number 1 (marked in light colour) represents the decision of sending the patient to an urban ED, while number 2 (marked in dark colour) represents the decision of sending the patient to a community ED. For both penalty functions, the policy sends more patients to a community ED when S_N becomes larger. The decision is also impacted by call locations. When the closest community ED is further away compared to the urban

ED, the policy suggests sending fewer patients to a community ED. As expected, the policy sends more medium acuity level patients to a community ED when the objective is to minimize the time to ED bed for patients ($r^{a_k}_{Pt}$), compared to minimizing the ambulance transportation time ($r^{a_k}_{Am}$).



$r^{a_k}_{Am}$	1 = to an urban ED 2 = to a community ED									
Patient Acuity: M					**S_N**					
Call Location	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**≥9**
(5, 20)	1	2	2	2	2	2	2	2	2	2
(5, 30)	1	1	2	2	2	2	2	2	2	2
(5, 40)	1	1	1	2	2	2	2	2	2	2
(5, 45)	1	1	1	2	2	2	2	2	2	2
(5, 50)	1	1	1	1	2	2	2	2	2	2
(5, 54)	1	1	1	1	2	2	2	2	2	2
(5, 58)	1	1	1	1	2	2	2	2	2	2
(5, 62)	1	1	1	1	1	2	2	2	2	2
(5, 64)	1	1	1	1	1	2	2	2	2	2
(5, 66)	1	1	1	1	1	2	2	2	2	2
(5, 68)	1	1	1	1	1	2	2	2	2	2
(5, 70)	1	1	1	1	1	1	2	2	2	2
(5, 80)	1	1	1	1	1	1	1	1	1	1
(5, 90)	1	1	1	1	1	1	1	1	1	1

(a) $r^{a_k}_{Am}$

$r^{a_k}_{Pt}$	1 = to an urban ED 2 = to a community ED									
Patient Acuity: M					**S_N**					
Call Location	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**≥9**
(5, 20)	2	2	2	2	2	2	2	2	2	2
(5, 30)	2	2	2	2	2	2	2	2	2	2
(5, 40)	1	2	2	2	2	2	2	2	2	2
(5, 45)	1	2	2	2	2	2	2	2	2	2
(5, 50)	1	2	2	2	2	2	2	2	2	2
(5, 54)	1	2	2	2	2	2	2	2	2	2
(5, 58)	1	1	2	2	2	2	2	2	2	2
(5, 62)	1	1	2	2	2	2	2	2	2	2
(5, 64)	1	1	2	2	2	2	2	2	2	2
(5, 66)	1	1	1	2	2	2	2	2	2	2
(5, 68)	1	1	1	2	2	2	2	2	2	2
(5, 70)	1	1	1	2	2	2	2	2	2	2
(5, 80)	1	1	1	1	2	2	2	2	2	2
(5, 90)	1	1	1	1	1	2	2	2	2	2

(b) $r^{a_k}_{Pt}$

Figure 5.5: A sample of the optimal policy with penalty functions $r^{a_k}_{Am}$ and $r^{a_k}_{Pt}$.

There is a minimum number of queued ambulances to trigger the policy to start sending patient to community EDs for each state. In other words, the policy suggests sending medium acuity level patients to a community ED instead of an urban ED once the number of queued ambulances reaches a certain threshold level. Therefore, we use a matrix of the S_N state thresholds with distance variables to present the detailed policy for medium acuity level patients for each penalty function (Figure 5.6). Each policy is presented using the minimum number of ambulances in AOD to trigger the policy to start sending patient to community EDs for each state. For example, when a call is 4 km and 50 km respectively from the urban and community EDs, send patients to the community ED when $S_n \geq 4$ (with penalty function $r^{a_k}_{Am}$).

When the penalty function is $r^{a_k}_{Am}$, the optimal policy suggests to send 16.3% of all patients to a community ED (or 28.9% of medium acuity patients). When the penalty function is $r^{a_k}_{Pt}$, this percentage increases to 31.6% of all patients (or 61.6%

of medium acuity patients). In both policies, this occurs more often when the urban EDs experience severe AOD, which is intuitive and matches our expectation of the result (Table 5.3).

penalty function	**% of patients sent to community EDs**				
	Resuscitation	High	Medium	Low	**Overall**
$r^{a_k}_{Am}$	0.0	0.0	28.9	0.0	**16.3**
$r^{a_k}_{Pt}$	0.0	0.0	61.6	0.0	**31.6**

Table 5.3: The percentages of patients allocated to various community EDs.

To demonstrate the advantages of the optimal policies, we calculate the stationary probability of the optimal policies, as well as the current practice (which is always sending patients to an urban ED). The first performance measure is to determine the improvement in AOD at the urban EDs when following the optimal policies. Figure 5.7 illustrates the probabilities of the system in each S_N state for penalty functions of $r^{a_k}_{Am}$, $r^{a_k}_{Pt}$, and the current practice of sending all patients to the urban EDs. The result shows a decrease in the number of queued ambulances when following either of the optimal policies compared to the current practice.

With the trends shown in Figure 5.7, a significant reduction of frequency of high AOD system state occurrence can be observed. Fox instance, the probability of being in a state with no less than five ambulances queueing at the ED decrease from 61.80% in the current policy to 17.94% and 7.52% in $r^{a_k}_{Am}$ and $r^{a_k}_{Pt}$ optimal policies, respectively. Furthermore, one can expect that the average ambulance turnaround time at the urban EDs should also be reduced with the optimal policies, due to the reduction of the number of queued ambulances. We estimate the average ambulance turnaround time at the urban EDs under each policy by using the historical data and the calculated stationary AOD probabilities. We found a reduction from 75.81 minutes (current) to 61.48 minutes and 54.68 minutes, respectively, for penalty functions $r^{a_k}_{Am}$ and $r^{a_k}_{Pt}$. Based on the historical data from 2016, with a total number of 22,243 patient transportation requests originated in the urban region, these two optimal polices $r^{a_k}_{Am}$ and $r^{a_k}_{Pt}$ would save 5,312.68 and 7,832.70 ambulance hours annually due to the reduction of AOD.

It is noteworthy that the system gains these performance improvements at the sacrifice of ambulance travel distance. Instead of travelling to the urban EDs which are closer to most call locations in the urban region, the new policy requires ambulances to occasionally travel further distances to transport patients. We compute and compare the expected average ambulance travel distance for patient transportation under each policy. The current policy has an average travel distance to an ED as 5.28 km. The optimal polices would increase that value to 13.29 km and 21.79 km with penalty functions of $r_{Am}^{a_k}$ and $r_{Pt}^{a_k}$, respectively.

In the MDP model we do not have a state for queued ambulances at the community EDs. In our study, AOD at community EDs is negligible and therefore data on waiting times and ambulance turnaround times are not available. Should this be required in other applications, the method used to account for queued ambulances at the urban hospitals can be applied. To test if AOD at community EDs will be problematic when following the policies of the MDP, we compute the expected number of additional ambulance transportations that will be sent to the community EDs. We found a mean of 9.9 and 19.0 additional transportations per day would be sent to the community EDs. In discussions with content experts, it was determined that this increased volume would not cause congestion in any community ED. However, the ambulance patient volume is likely to occur in peaks during periods of time when the urban EDs are overcrowded due to the nature of the decision rules. Further analysis are required to ensure sufficient resources are available at the community EDs during the busy time of days when considering implementing the optimal policies.

5.5.2 Sensitivity Analysis

Sensitivity analysis is an approach that can help understand the relationships of model attributes and outcomes by analyzing how the outcomes changes with different variable values. There are two major objectives in the sensitivity analysis of this study. We aim to consider and understand the influences of time of day and increasing AOD times.

Time of day

As previously discussed, the call volume in the studied region is not stationary throughout the day. Data analysis of the historical data suggests that the busy hours of the day is from 9 a.m. to 7 p.m. (11 hours), where the hourly average call volume reaches a relatively constant high level, while the rest of the day (8 p.m. to 8 a.m.) are non-busy hours (13 hours), where the hourly average call volume drops to a lower level. Between the two different levels, the actual time span between calls (decision epochs) may vary considerably, which affects the empirical distribution of D, denoting the number of ambulances released from urban EDs between decision epochs. Therefore, we separate the historical data into a high (busy hours) and low (non-busy hours) call level periods and solve the MDP model for each scenario. The results of these two scenarios are included in the sensitivity analysis.

Figure 5.8 shows the empirical distributions of D in the high and low call level of periods, as well as the overall distribution. The distributions are surprisingly similar with only a slightly lower discharge rate during the busy hours. One possible explanation is that despite the shorter time between calls during the busy hours, the ED capacities are also likely higher at these hours. Further data analysis reveals that the empirical distributions of patient acuity level and call location are time invariant. To model these two time periods, new transition probabilities are calculated with separate datasets of busy and non-busy hours. The results are presented in Table 5.4 as a form of percentages of patients being sent to community EDs during different periods. The optimal policies send more patients to the community EDs during the busy hours and less patients to these EDs during non-busy hours, comparing to the overall scenarios. Based on the call volumes, we also compute the expected number of additional ambulance transportations that will be sent to the community EDs during these two periods. During the busy hours, the community EDs can be expected to receive additional 7.6 and 12.2 ambulance transportations when following optimal policies with the penalty function $r_{Am}^{a_k}$ and $r_{Pt}^{a_k}$, respectively. Similarly, the increases are 3.8 and 8.0 for non-busy hours. In the base scenario, a mean of 9.9 and 19.0 additional transportations would be sent to the community EDs per day, with the penalty function $r_{Am}^{a_k}$ and $r_{Pt}^{a_k}$, respectively.

Generally speaking, the optimal policies developed with separated datasets (busy and non-busy hours) do not indicate significantly different outcomes within the EMS system under study. Yet, an EMS system may have different distributions of ambulance discharge rates during busy/non-busy hours, such that larger differences can be expected in the optimal policies at different periods. In this case, policies can be developed based on the time of the day to provide EMS personals with more precised ambulance destination instructions to follow.

penalty function	**% of patients sent to community EDs**		
	Overall Scenario	Busy Hours	Non-busy Hours
$r^{a_k}_{Am}$	16.3	21.7	14.8
$r^{a_k}_{Pt}$	31.6	35.0	30.8

Table 5.4: The percentages of patients allocated to various community EDs in the non-busy hours and busy hours scenarios.

Increasing AOD time

To further explore model sensitivity, we increase the AOD time to observe its effect on the optimal ambulance destination policies and the total percentage of patients being sent to community EDs. We use the results of Section 5.5.1 as the base scenario, and consider the scenarios when the AOD time increases by 5%, 10%, and 20%. The results are reported in Table 5.5.

penalty function	**Overall % of patients sent to community EDs**			
	Base Scenario	AOD 5%	AOD 10%	AOD 20%
$r^{a_k}_{Am}$	16.3	20.7	22.0	24.8
$r^{a_k}_{Pt}$	31.6	35.2	37.1	41.5

Table 5.5: The percentages of patients allocated to community EDs when the AOD time increases by 5%, 10%, and 20%, respectively.

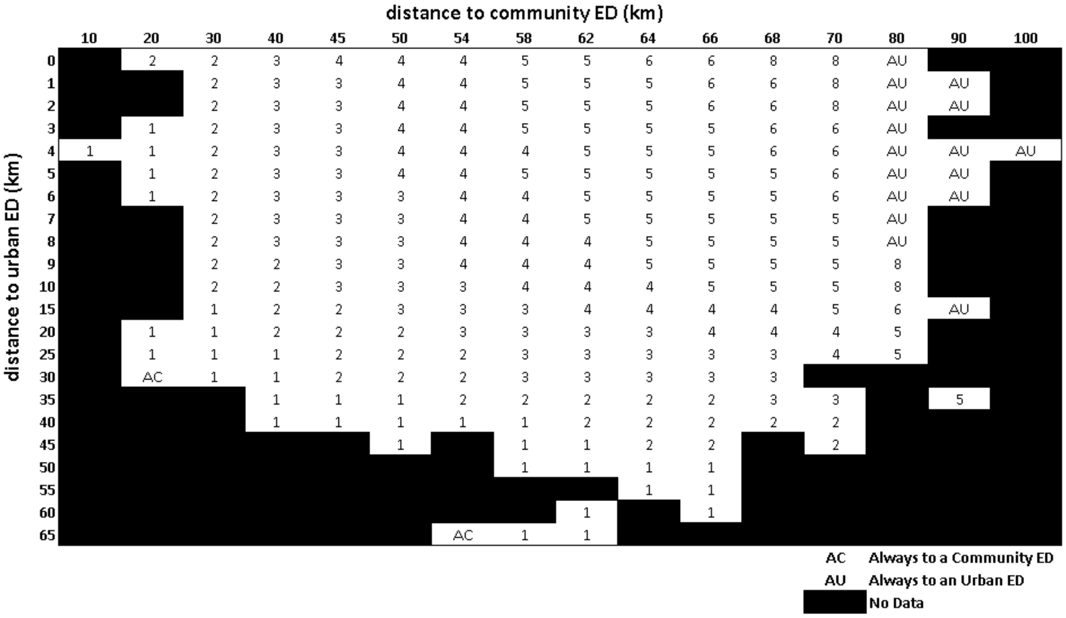
As expected, we observe that the optimal policies change gradually and send more patients to community EDs as the AOD in the city gets worse. This is an important observation as the model assumption of no AOD in any community ED may becomes inappropriate as a large amount of patients are sent to it. This can be overcome as discussed in Section 5.5.1.

5.6 Conclusion and Discussion

To find long-term solutions that minimize the effects of AOD and improve performance, we develop an MDP model to assist EMS dispatchers in determining the best ambulance destinations for their patients. The computational study indicates that we can make ambulance destination decisions using a robust policy based on the current number of queued ambulances, call location, and patient acuity level. According to the results, both the EMS systems and patients benefit from the improved policy. In addition to the metrics considered by the model, patient risks, outcomes, and preferences are factors which are important for future study and which should be part of an implementation plan.

Instead of using theoretical distributions, our model demonstrates a method to incorporate large amounts of administrative data. Such data is typically available in modern EMS systems and allows us to represent the real system with fewer limiting assumptions. Furthermore, we compute ambulance travel time using the KWH model instead of with an assumed constant speed. This is particularly important in this study since ambulances are being routed long distances in some cases. Our model also provides considerations of patient acuity levels that can influence the optimal destination decision.

The model is suitable for use in decision support systems that allow EMS dispatchers to quickly evaluate the situation and make decisions on which destination ED to send the incoming ambulances. The model also provides accurate estimates of the number of queued ambulances, the average ambulance turnaround time at EDs, and the average travel distance of the ambulances. It is sufficiently general to be used by EMS systems to mitigate the impact of AOD on their operations.

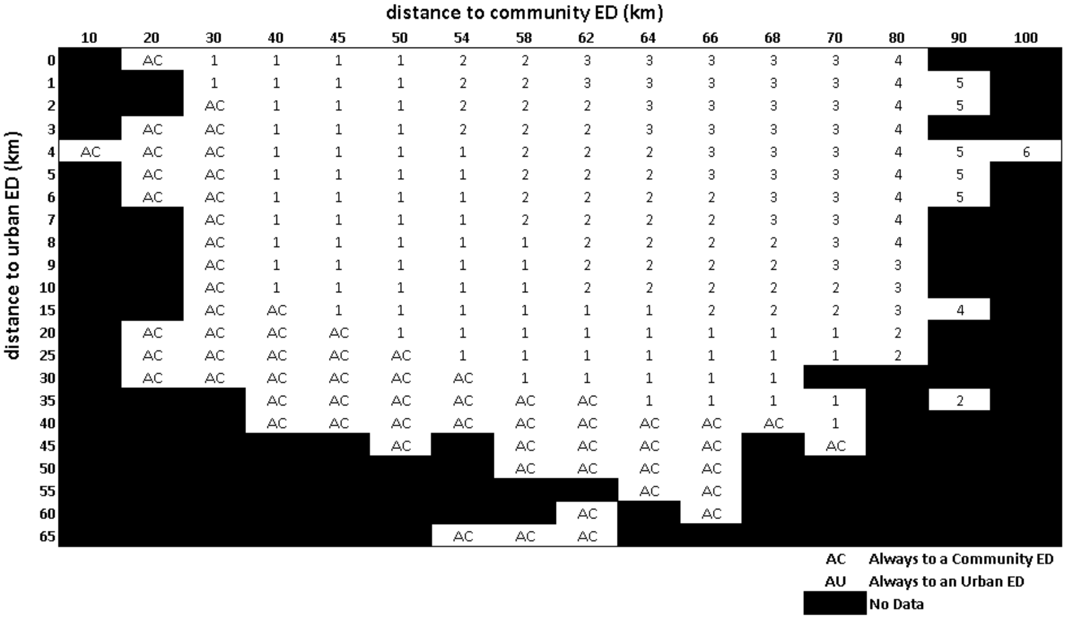


distance to community ED (km)

distance to urban ED (km)	10	20	30	40	45	50	54	58	62	64	66	68	70	80	90	100
0	■	2	2	3	4	4	4	5	5	6	6	8	8	AU	■	■
1	■	■	2	3	3	4	4	5	5	5	6	6	8	AU	AU	■
2	■	■	2	3	3	4	4	5	5	5	6	6	8	AU	AU	■
3	■	1	2	3	3	4	4	5	5	5	5	6	6	AU	■	■
4	1	1	2	3	3	4	4	4	5	5	5	6	6	AU	AU	AU
5	■	1	2	3	3	4	4	5	5	5	5	5	6	AU	AU	■
6	■	1	2	3	3	3	4	4	5	5	5	5	6	AU	AU	■
7	■	■	2	3	3	3	4	4	5	5	5	5	5	AU	■	■
8	■	■	2	3	3	3	4	4	4	5	5	5	5	AU	■	■
9	■	■	2	2	3	3	4	4	4	5	5	5	5	8	■	■
10	■	■	2	2	3	3	3	4	4	4	5	5	5	8	■	■
15	■	■	1	2	2	3	3	3	4	4	4	4	5	6	AU	■
20	■	1	1	2	2	2	3	3	3	3	4	4	4	5	■	■
25	■	1	1	1	2	2	2	3	3	3	3	3	4	5	■	■
30	■	AC	1	1	2	2	2	3	3	3	3	3	■	■	■	■
35	■	■	■	1	1	1	2	2	2	2	2	3	3	■	5	■
40	■	■	■	1	1	1	1	1	2	2	2	2	2	■	■	■
45	■	■	■	■	■	1	■	1	1	2	2	■	2	■	■	■
50	■	■	■	■	■	■	■	1	1	1	1	■	■	■	■	■
55	■	■	■	■	■	■	■	■	■	1	1	■	■	■	■	■
60	■	■	■	■	■	■	■	■	1	■	1	■	■	■	■	■
65	■	■	■	■	■	■	AC	1	1	■	■	■	■	■	■	■

AC	Always to a Community ED
AU	Always to an Urban ED
■	No Data

(a) $r_{Am}^{a_k}$



distance to community ED (km)

distance to urban ED (km)	10	20	30	40	45	50	54	58	62	64	66	68	70	80	90	100
0	■	AC	1	1	1	1	2	2	3	3	3	3	3	4	■	■
1	■	■	1	1	1	1	2	2	3	3	3	3	3	4	5	■
2	■	■	AC	1	1	1	2	2	2	3	3	3	3	4	5	■
3	■	AC	AC	1	1	1	2	2	2	3	3	3	3	4	■	■
4	AC	AC	AC	1	1	1	1	2	2	2	3	3	3	4	5	6
5	■	AC	AC	1	1	1	1	2	2	2	3	3	3	4	5	■
6	■	AC	AC	1	1	1	1	2	2	2	2	3	3	4	5	■
7	■	■	AC	1	1	1	1	2	2	2	2	3	3	4	■	■
8	■	■	AC	1	1	1	1	1	2	2	2	2	3	4	■	■
9	■	■	AC	1	1	1	1	1	2	2	2	2	3	3	■	■
10	■	■	AC	1	1	1	1	1	2	2	2	2	2	3	■	■
15	■	■	AC	AC	1	1	1	1	1	1	2	2	2	3	4	■
20	■	AC	AC	AC	AC	1	1	1	1	1	1	1	1	2	■	■
25	■	AC	AC	AC	AC	AC	1	1	1	1	1	1	1	2	■	■
30	■	AC	AC	AC	AC	AC	AC	1	1	1	1	1	■	■	■	■
35	■	■	■	AC	AC	AC	AC	AC	AC	1	1	1	1	■	2	■
40	■	■	■	AC	AC	AC	AC	AC	AC	AC	AC	AC	1	■	■	■
45	■	■	■	■	■	AC	■	AC	AC	AC	AC	■	AC	■	■	■
50	■	■	■	■	■	■	■	AC	AC	AC	AC	■	■	■	■	■
55	■	■	■	■	■	■	■	■	■	AC	AC	■	■	■	■	■
60	■	■	■	■	■	■	■	■	AC	■	AC	■	■	■	■	■
65	■	■	■	■	■	■	AC	AC	AC	■	■	■	■	■	■	■

AC	Always to a Community ED
AU	Always to an Urban ED
■	No Data

(b) $r_{Pt}^{a_k}$

Figure 5.6: The detailed policy of the medium acuity level patients for each penalty function, (a) $r_{Am}^{a_k}$ and (b) $r_{Pt}^{a_k}$.

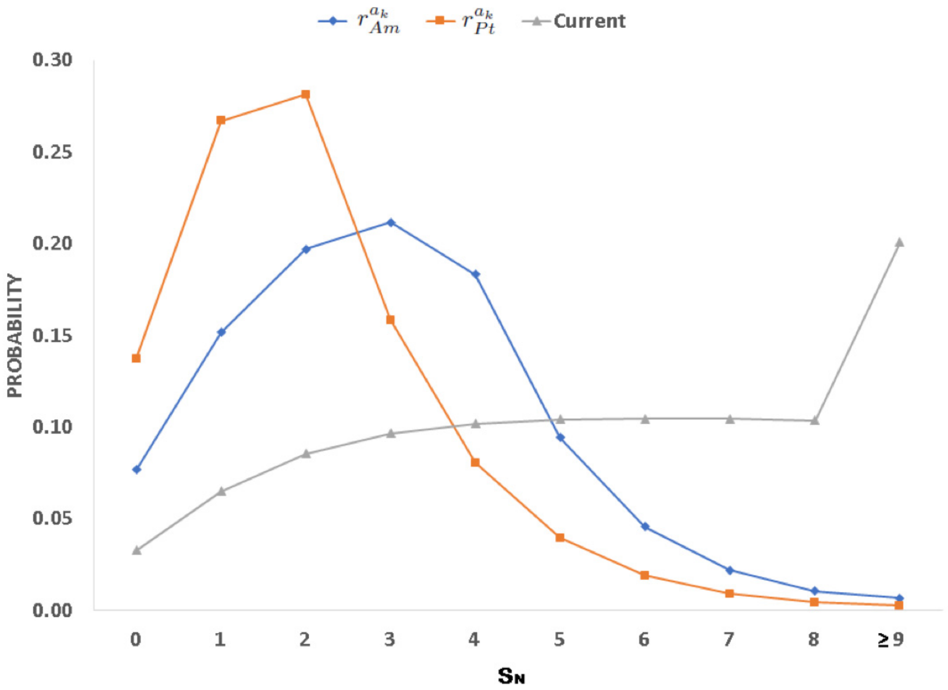


Figure 5.7: The probabilities of S_N states at the urban EDs when following policies: $r^{a_k}_{Am}$, $r^{a_k}_{Pt}$, and the current policy $Current$.

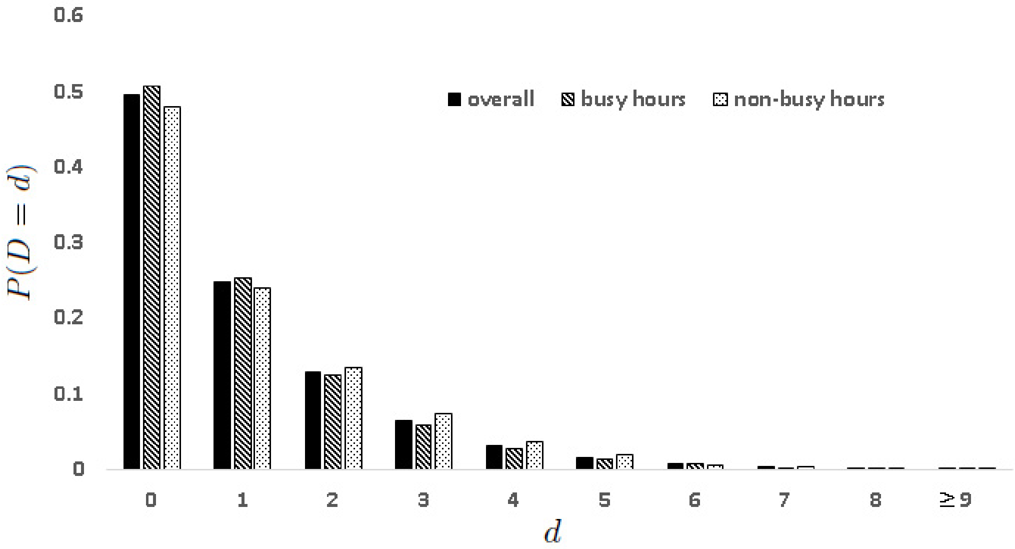


Figure 5.8: The empirical distribution of the probability of the numbers of ambulances being released from the urban EDs between decision epochs in non-busy hours and busy hours, in comparison to the overall distribution.

Chapter 6

CONCLUSION

Healthcare is an area of growing importance and cost around the world, thus a popular area for operations research. As a key element of the healthcare network, EMS systems require constant improvements to ensure capacity to adequately and efficiently respond to the emergency care needs of the public. Ambulance offload delay (AOD), as an EMS operational problem, has become common in many health care systems. However, research examining system performances associated with EMS is still limited.

In this thesis, we model different perspectives of the AOD problem using various operation research approaches to establish a better understanding of its impact within an EMS system. This includes designing models to help a provincial EMS provider mitigate AOD. To achieve these objectives, we develop three distinct research stages each with a different modeling approach.

In the first stage, we conduct a systematic literature review on AOD. To our best knowledge, this is the first published review of AOD related studies and models (*Chapter 2*). This chapter describes the causes and consequences of this growing problem, key measures that are used to assess system performance, and potential solutions investigated using various methods. Furthermore, we provide a comprehensive depiction of the AOD problem experienced by the provincial EMS provider in Nova Scotia, Canada (*Chapter 3*). We show how this problem has a substantial impact on ambulance performance, leading to prolonged ambulance turnaround times, total call times, and response times, as well as reduced ambulance availability. The descriptive analytics and statistical models are presented in a way that can be generalized to other EMS systems for measuring ambulance performance with respect to AOD.

The next stage of this research (*Chapter 4*) provides the EMS provider with a decision-support model that can predict AOD status based on the current system status. As information technology advances, such prediction models can be developed using the shared information between the EMS providers and the hospital EDs. We design various practical prediction settings for this application and utilize a hybrid decision tree algorithm to improve the performance of these models. This way, proactive interventions can be initiated by the decision makers based on different states of the system to mitigate the problem. Our predictive analytics suggest that the AOD status of the EMS system is robust and resistant to any sudden changes in a short period of time. The prediction models perform relatively well with accuracy rates of 60%-75%, 69%-83%, and 91%-95%, with respect to different prediction settings discussed in the chapter. As expected, the presence of high degrees of variability negatively impacts the performance of the prediction model. The variability is likely due to the complexity of EMS systems while modeling with realistic details incorporated to reflect the real-world situation. We also compare the hybrid decision tree algorithm with a basic decision tree algorithm (classification and regression tree). The prediction models generated using the hybrid decision tree algorithm outperform the ones generated by the traditional algorithm by an average accuracy improvement of 2.44%.

Our final stage of research began by gathering feasible intervention ideas from the key frontline personnel of the local EMS providers. Among these interventions, we model optimal ambulance destination policies for an EMS system when considering AOD, and evaluate their effects on the system performance (*Chapter 5*). These policies determine when it is advantageous for ambulance patient to be transported to an out-of-region ED (that is not affected by AOD) to achieve a shorter ambulance turnaround time. In specific cases, ambulances can return to service quicker, and thus reduce the effects of AOD on the system performance. It is anticipated that best practices produced from this study will be directly transferable.

The AOD problem is a consequence of a much bigger problem, which is the lack of capacity in the healthcare system to treat hospital inpatients, leading to ED overcrowding and access block. AOD includes clinical, operational, and administrative perspectives and must be addressed in a system-wide manner. Research has shown that initiatives and efforts from one party (EMS or ED) alone may not be sufficient to solve this problem, a more collaborative approach is required. Establishing better collaboration between EMS and hospital EDs should be the first step towards the goal of building a system-wide solution to this problem. EMS providers and hospital EDs should initiate dialogues at high management levels and work together to take appropriate steps to mitigate AOD. Timely information sharing between these two parties could allow interventions built to achieve benefits to both.

Evidence suggests that the root causes of AOD lie outside the EMS system and to address it will likely take significant time and effort and require system-wide policy changes. Meanwhile, EMS operation is impaired by this problem. Therefore, research should continue to develop interventions, either through operation research models or operation trials, to help EMS operate in this difficult environment and mitigate the negative impacts of AOD. While the AOD problem presents itself as a challenging problem, it also represents an opportunity for public health, EMS, and hospitals, to come together to identify best practices and implement positive changes. Ultimately, all key components of the health care system should work together to ensure the ED crowding problem is eliminated or minimized, thereby alleviating much of the AOD problem.

While this thesis provides a number of insights on the different perspectives of the AOD problem, there are several directions that can be considered for further research:

- There is limited research focused on the AOD problem specifically in the operation research field. However, operation research methodologies should be recognized as powerful tools for this problem. Several models are developed in this research to measure the effects of AOD, predict the system status, and develop optimal ambulance destination policies. Yet, there is much more to explore with either improvements of the existing models, or new developments with other approaches. For example, our models can be extended to include more factors to better capture the complexity of the EMS system but at the expense of higher dimensionality and tractability. The interface of EMS and EDs can be modeled using queueing theory. A simulation model may reveal more insights of this problem when modeling a broader perspective of the healthcare system that includes patients from arrival to the ED to being discharged from the hospital. To help better assess and mitigate this problem, models need to be further developed to estimate the system performance in a more realistic and detailed environment.

- For the interventions to be effective and true to real-world situations, the measurement of related metrics needs to be improved. Further study is required to standardize the definition and the measurements of the ambulance offload process. It is important that future research on this topic are based on solid measurements of the main components of this process.

- Another aspect of AOD assessment, which operation research may help investigate, is its impact on the workload of paramedics and ED staff. As it becomes a new norm, are there human resource consequences in terms of increased rates of human error, scheduling conflicts, etc.?

- There is a lack of clinical study that investigates if there is a relationship between the offload delay and patients risk levels. For example, is AOD more common or prolonged for patients with certain clinical conditions? If so, what are the impacts on the safety and outcomes of these patients? What policy should we consider to alleviate the consequences?

- There are few studies describing the relationship between AOD and EMS performance. Most studies use anecdote evidence and rationalizations as supposed to empirical studies. It would be beneficial for future work to further quantify this relationship.

Bibliography

[1] Mason AJ (2013) Simulation and real-time optimised relocation for improving ambulance operations, Springer, New York, NY. Handbook of Healthcare Operations Management: Methods and Applications, URL `https://search.proquest.com/docview/1569729197`

[2] Creemers S, Lambrecht M, Vandaele N (2007) Queueing models in healthcare. Tijdschrift voor economie en management 52(3):471–497

[3] Aboueljinane L, Sahin E, Jemai Z (2013) A review on simulation models applied to Emergency Medical Service operations. Computers and Industrial Engineering 66(4):734–750

[4] Smith L (2013) Modelling emergency medical services. PhD thesis, URL `https://orca.cf.ac.uk/47743/1/2013smithlphd.pdf`

[5] Almehdawe E, Jewkes B, He QM (2013) A Markovian queueing model for ambulance offload delays. European Journal of Operational Research 226(3):602–614

[6] Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. European Journal of Operational Research 147(3):451–463, DOI 10.1016/S0377-2217(02)00364-8

[7] Carter AJ, Overton J, Terashima M, Cone DC (2014) Can Emergency Medical Services use turnaround time as a proxy for measuring ambulance offload time? J Emerg Med 47(1):30–35, DOI 10.1016/j.jemermed.2013.08.109

[8] Reuter-Oppermann M, Berg PLvd, Vile JL (2017) Logistics for emergency medical service systems. Health Systems 6(3):187–208, DOI 10.1057/s41306-017-0023-x

[9] Spaite DW, Valenzuela TD, Meislin HW, Criss EA, Hinsberg P (1993) Prospective validation of a new model for evaluating Emergency Medical Services systems by in-field observation of specific time intervals in prehospital care. Annals of Emergency Medicine 22(4):638–645, DOI 10.1016/S0196-0644(05)81840-2

[10] Cone DC, Davidson SJ, Nguyen Q (1998) A time-motion study of the Emergency Medical Services turnaround interval. Ann Emerg Med 31(2):241–246

[11] Takeda RA, Widmer JA, Morabito R (2007) Analysis of ambulance decentralization in an urban Emergency Medical Service using the hypercube queueing model. Computers & Operations Research 34(3):727–741, DOI 10.1016/j.cor.2005.03.022

[12] Singer M, Donoso P (2008) Assessing an ambulance service with queuing theory. Computers & Operations Research 35(8):2549–2560

[13] Restrepo M, Henderson S, Topaloglu H (2009) Erlang loss models for the static deployment of ambulances. Health Care Management Science 12(1):67–79, DOI 10.1007/s10729-008-9077-4

[14] Spaite D, Benoit R, Brown D, Cales R, Kaufmann C, Pollock D, Yano EM, Ryan S, Glass C, Dawson D (1995) Uniform pre-hospital data elements and definitions: a report from the uniform pre-hospital emergency medical services data conference. Annals of Emergency Medicine 25(4):525–531, DOI 10.1016/S0196-0644(95)70271-7

[15] Pinto LR, Silva PMS, Young TP (2015) A generic method to develop simulation models for ambulance systems. Simulation Modelling Practice and Theory 51:170–183, DOI 10.1016/j.simpat.2014.12.001

[16] Fomundam S, Herrmann JW (2007) A survey of queuing theory applications in healthcare. Tech. rep., URL `http://hdl.handle.net/1903/7222`

[17] Richardson LD, Asplin BR, Lowe RA (2002) Emergency department crowding as a health policy issue: past development, future directions. Ann Emerg Med 40:388–393, DOI 10.1067/mem.2002.128012

[18] Schafermeyer RW, Asplin BR (2003) Hospital and emergency department crowding in the United States. Emergency Medicine 15(1):22–27, DOI 10.1046/j.1442-2026.2003.00403.x

[19] Ay D (2010) Patient population and factors determining length of stay in adult ED of a Turkish university medical center. Am J Emerg Med 28(3):325–330, DOI 10.1016/j.ajem.2008.12.011

[20] Crilly J, Keijzers G, Tippett V, O'Dwyer J, Lind J, Bost N, O'Dwyer M, Shiels S, Wallis M (2015) Improved outcomes for emergency department patients whose ambulance off-stretcher time is not delayed. Emerg Med Australas 27(3):216–224, DOI 10.1111/1742-6723.12399

[21] Almehdawe E (2012) Queueing network models of ambulance offload delays. PhD thesis, University of Waterloo, URL `http://hdl.handle.net/10012/7046`

[22] Eckstein M, Isaacs SM, Slovis CM, Kaufman BJ, Loflin JR, O'Connor RE, Pepe PE (2005) Facilitating EMS turnaround intervals at hospitals in the face of receiving facility overcrowding. Prehosp Emerg Care 9(3):267–275, DOI 10.1080/10903120590962102

[23] Silvestri S, Ralls GA, Sun J, Shah KJ, Parrish GA (2006) Evaluation of patients in delayed Emergency Medical Services unit off-load status. Acad Emerg Med 13(5):S70

[24] Silvestri S, Ralls GA, Papa L, Barnes M (2006) Impact of emergency department bed capacity on Emergency Medical Services unit off-load time. Acad Emerg Med 13(5):S70

[25] Lee YJ, Shin SD, Lee EJ, Cho JS, Cha WC (2015) Emergency department overcrowding and ambulance turnaround time. PLoS One 10(6):e0130,758, DOI 10.1371/journal.pone.0130758

[26] Eckstein M, Chan LS (2004) The effect of emergency department crowding on paramedic ambulance availability. Ann Emerg Med 43(1):100–105, DOI 10.1016/s0196064403007479

[27] Hitchcock M, Crilly J, Gillespie B, Chaboyer W, Tippett V, Lind J (2010) The effects of ambulance ramping on emergency department length of stay and in-patient mortality. Australasian Emergency Nursing Journal 13(1-2):17–24, DOI 10.1016/j.aenj.2010.02.004

[28] Cooney DR, Millin MG, Carter A, Lawner BJ, Nable JV, Wallus HJ (2011) Ambulance diversion and emergency department offload delay: resource document for the National Association of EMS Physicians position statement. Prehosp Emerg Care 15(4):555–561, DOI 10.3109/10903127.2011.608871

[29] Hamilton TE (2006) "Parking" of emergency medical service patients in hospitals. US Department of Health and Human Services

[30] Carter AJ, Gould JB, Vanberkel P, Jensen JL, Cook J, Carrigan S, Wheatley MR, Travers AH (2015) Offload zones to mitigate Emergency Medical Services (EMS) offload delay in the emergency department: a process map and hazard analysis. CJEM 17(6):670–678, DOI 10.1017/cem.2015.15

[31] Li M, Vanberkel P, Carter AJE (2018) A review on ambulance offload delay literature. Health Care Management Science pp 1–18, DOI 10.1007/s10729-018-9450-x

[32] Majedi M (2008) A queueing model to study ambulance offload delays. Master's thesis, University of Waterloo

[33] Almehdawe E, Jewkes B, He QM (2016) Analysis and optimization of an ambulance offload delay and allocation problem. Omega-International Journal of Management Science 65:148–158

[34] (2012) Patient handover delays being addressed by ambulance service. Nurs Stand 26(48):10, DOI 10.7748/ns.26.48.10.s15

[35] Segal E, Verter V, Colacone A, Afilalo M (2006) The in-hospital interval: a description of EMT time spent in the emergency department. Prehosp Emerg Care 10(3):378–382, DOI 10.1080/10903120600725884

[36] Cone DC, Middleton PM, Pour SM (2012) Analysis and impact of delays in ambulance to emergency department handovers. Emerg Med Australas 24(5):525–533, DOI 10.1111/j.1742-6723.2012.01589.x

[37] Cooney DR, Wojcik S, Seth N, Vasisko C, Stimson K (2013) Evaluation of ambulance offload delay at a university hospital emergency department. Int J Emerg Med 6(1):15, DOI 10.1186/1865-1380-6-15

[38] Hammond E, Holzhauser K, Shaban R, Melton N (2009) An exploratory study to examine the phenomenon and practice of 'ambulance ramping' at hospitals within the Southern Health Service Districts of Queensland and Queensland Ambulance Service. Australasian Emergency Nursing Journal 12(4):170

[39] Taylor C, Williamson D, Sanghvi A (2006) When is a door not a door? The difference between documented and actual arrival times in the emergency department. Emergency medicine journal 23(6):442–443, DOI 10.1136/emj.2005.029868

[40] Cooney DR, Wojcik S, Seth N (2011) Can NEDOCS score be used to predict ambulance offload delay? Ann Emerg Med 58(4):S217

[41] Steer S, Bhalla MC, Zalewski J, Frey J, Nguyen V, Mencl F (2016) Use of radio frequency identification to establish emergency medical service offload times. Prehosp Emerg Care 20(2):254–259, DOI 10.3109/10903127.2015.1076093

[42] Richardson DB, Mountain D (2009) Myths versus facts in emergency department overcrowding and hospital access block. Medical Journal of Australia 190(7):369

[43] Derlet RW, Richards JR (2000) Overcrowding in the nation's emergency departments: complex causes and disturbing effects. Annals of Emergency Medicine 35(1):63

[44] Fatovich DM, Hirsch RL (2003) Entry overload, emergency department overcrowding, and ambulance bypass. Emergency Medicine Journal 20(5):406–409, DOI 10.1136/emj.20.5.406

[45] Sayed ME, Mitchell PM, White LF, Rubin-Smith J, Maciejko TM, Obendorfer DT, Ulrich AS, Dyer S, Olshaker JS (2012) Impact of an emergency department closure on the local emergency medical services system. Prehosp Emerg Care 16(2):198–203, DOI 10.3109/10903127.2011.640418

[46] Schwartz B, Department HE, Group AEW (2005) Improving access to emergency services: a system commitment. Tech. rep., Ontario Ministry of Health and Long-Term Care

[47] Andrulis DP, Kellermann A, Hintz EA, Hackman BB, Weslowski VB (1991) Emergency departments and crowding in United States teaching hospitals. Ann Emerg Med 20(9):980–986

[48] Derlet RW, Richards JR, Kravitz RL (2001) Frequent overcrowding in U.S. emergency departments. Academic Emergency Medicine 8(2):151–155, DOI 10.1111/j.1553-2712.2001.tb01280.x

[49] Schneider S, Zwemer F, Doniger A, Dick R, Czapranski T, Davis E (2001) Rochester, New York: a decade of emergency department overcrowding. Acad Emerg Med 8(11):1044–1050

[50] Derlet RW, Richards JR (2002) Emergency department overcrowding in Florida, New York, and Texas. Southern Medical Journal 95(8):846

[51] Olshaker JS, Rathlev NK (2006) Emergency department overcrowding and ambulance diversion: the impact and potential solutions of extended boarding of admitted patients in the emergency department. J Emerg Med 30(3):351–356, DOI 10.1016/j.jemermed.2005.05.023

[52] Hoot NR, LeBlanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, Aronsky D (2008) Forecasting emergency department crowding: a discrete event simulation. Ann Emerg Med 52(2):116–125, DOI 10.1016/j.annemergmed.2007.12.011

[53] Schull MJ, Lazier K, Vermeulen M, Mawhinney S, Morrison LJ (2003) Emergency department contributors to ambulance diversion: a quantitative analysis. Ann Emerg Med 41(4):467–476, DOI 10.1067/mem.2003.23

[54] Moskop JC, David PS, Gelderman JM, Schears RM, Bookman KJ (2009) Emergency department crowding, part 1–concept, causes, and moral consequences. Ann Emerg Med 53(5):605–611, DOI 10.1016/j.annemergmed.2008.09.019

[55] Yancer DA, Foshee D, Cole H, Beauchamp R, la Pena de, Keefe T, Smith W, Zimmerman K, Lavine M, Toops B (2006) Managing capacity to reduce emergency department overcrowding and ambulance diversions. The Joint Commission Journal on Quality and Patient Safety 32(5):239–245

[56] Cameron PA, Joseph AP, McCarthy SM (2009) Access block can be managed. Med J Aust 190(7):364–368

[57] Olshaker JS (2009) Managing emergency department overcrowding. Emerg Med Clin North Am 27(4):603, viii, DOI 10.1016/j.emc.2009.07.004

[58] Ting JY (2008) The potential adverse patient effects of ambulance ramping, a relatively new problem at the interface between prehospital and ED care. J Emerg Trauma Shock 1(2):129, DOI 10.4103/0974-2700.43201

[59] Kingswell C, Shaban RZ, Crilly J (2015) The lived experiences of patients and ambulance ramping in a regional Australian emergency department: An interpretive phenomenology study. Australas Emerg Nurs J 18(4):182–189, DOI 10.1016/j.aenj.2015.08.003

[60] Esensoy AV (2008) Evaluation of the demonstration project to direct low acuity ambulance patients to urgent care centres to improve ambulance availability. Master thesis, University of Toronto.

[61] Perry M, Carter D (2017) The ethics of ambulance ramping. Emerg Med Australas 29(1):116–118, DOI 10.1111/1742-6723.12625

[62] Schwartz B (2015) Transfer of care and offload delay: continued resistance or integrative thinking? CJEM 17(6):679–684, DOI 10.1017/cem.2014.62

[63] NSHA (2015) Quarterly performance report emergency departments and system flow. URL `http://www.cdha.nshealth.ca/about-us-39`

[64] Laan CM, Vanberkel PT, Boucherie RJ, Carter AJE (2016) Offload zone patient selection criteria to reduce ambulance offload delay. Operations Research for Health Care 11:13–19

[65] Newell K (2013) Offload delay - returning paramedic unit hours to the street: the Ottawa approach. Canadian Paramedicine 36(6):20–22

[66] Clarey A, Allen M, Brace-McDonnell S, Cooke MW (2014) Ambulance handovers: can a dedicated ED nurse solve the delay in ambulance turnaround times? Emerg Med J 31(5):419–420, DOI 10.1136/emermed-2012-202258

[67] Greaves T, Mitchell M, Zhang P, Crilly J (2017) The impact of an emergency department ambulance offload nurse role: A retrospective comparative study. Int Emerg Nurs 32:39–44, DOI 10.1016/j.ienj.2016.12.005

[68] Han JH, Zhou C, France DJ, Zhong S, Jones I, Storrow AB, Aronsky D (2007) The effect of emergency department expansion on emergency department overcrowding. Acad Emerg Med 14(4):338–343, DOI 10.1197/j.aem.2006.12.005

[69] Crilly J, O'Dwyer J, Lind J, Tippett V, Thalib L, O'Dwyer M, Keijzers G, Wallis M, Bost N, Shiels S (2013) Impact of opening a new emergency department on healthcare service and patient outcomes: analyses based on linking ambulance, emergency and hospital databases. Intern Med J 43(12):1293–1303, DOI 10.1111/imj.12202

[70] Crilly JL, Keijzers GB, Tippett VC, O'Dwyer JA, Wallis MC, Lind JF, Bost NF, O'Dwyer MA, Shiels S (2014) Expanding emergency department capacity: a multisite study. Aust Health Rev 38(3):278–287, DOI 10.1071/ah13085

[71] Lee IH, Chen CT, Lee YT, Hsu YS, Lu CL, Huang HH, Hsu TF, How CK, Yen DH, Yang UC (2017) A new strategy for emergency department crowding: High-turnover utility bed intervention. J Chin Med Assoc 80(5):297–302, DOI 10.1016/j.jcma.2016.11.002

[72] Services AH (2010) AHS launches overcapacity protocols. URL `http://www.albertahealthservices.ca/news/releases/2010/Page3376.aspx`

[73] McRae A, Wang D, Blanchard IE, Almansoori W, Lang E, Innes G, Anton A (2012) Benefits on EMS offload delay of a provincial ED overcapacity protocol aimed at reducing ED boarding. Tech. rep.

[74] Lagoe RJ, Jastremski MS (1990) Relieving overcrowded emergency departments through ambulance diversion. Hosp Top 68(3):23–27

[75] Deo S, Gurvich I (2011) Centralized vs. decentralized ambulance diversion: a network perspective. Management Science 57(7):1300–1319

[76] Lagoe RJ, Hunt RC, Nadle PA, Kohlbrenner JC (2002) Utilization and impact of ambulance diversion at the community level. Prehosp Emerg Care 6(2):191–198

[77] Burt CW, McCaig LF, Valverde RH (2006) Analysis of ambulance transports and diversions among US emergency departments. Ann Emerg Med 47(4):317–326, DOI 10.1016/j.annemergmed.2005.12.001

[78] Warden CR, Bangs C, Norton R, Huie J (2003) Temporal trends in ambulance diversion in a mid-sized metropolitan area. Prehosp Emerg Care 7(1):109–113

[79] Leegon J, Hoot N, Aronsky D, Storkey A (2007) Predicting ambulance diversion in an adult emergency department using a Gaussian process. AMIA Annu Symp Proc p 1026

[80] Hagtvedt R, Ferguson M, Griffin P, Jones GT, Keskinocak P (2009) Cooperative strategies to reduce ambulance diversion. In: Simulation Conference (WSC), Proceedings of the 2009 Winter, Winter Simulation Conference Proceedings, pp 1861–1874

[81] Ramirez-Nafarrate A, Fowler JW, Wu T (2011) Design of centralized ambulance diversion policies using simulation-optimization. In: Simulation Conference (WSC), Proceedings of the 2011 Winter, Winter Simulation Conference Proceedings, pp 1251–1262, DOI 10.1109/WSC.2011.6147846

[82] Ramirez-Nafarrate A, Hafizoglu AB, Gel ES, Fowler JW (2014) Optimal control policies for ambulance diversion. European Journal of Operational Research 236(1):298–312

[83] Lin CH, Kao CY, Huang CY (2015) Managing emergency department overcrowding via ambulance diversion: a discrete event simulation model. J Formos Med Assoc 114(1):64–71, DOI 10.1016/j.jfma.2012.09.007

[84] Kao CY, Yang JC, Lin CH (2015) The impact of ambulance and patient diversion on crowdedness of multiple emergency departments in a region. PLoS One 10(12):e0144,227, DOI 10.1371/journal.pone.0144227

[85] Scheulen JJ, Li G, Kelen GD (2001) Impact of ambulance diversion policies in urban, suburban, and rural areas of Central Maryland. Acad Emerg Med 8(1):36–40

[86] Carter AJ, Grierson R (2007) The impact of ambulance diversion on EMS resource availability. Prehosp Emerg Care 11(4):421–426, DOI 10.1080/10903120701536909

[87] Asplin BR (2003) Does ambulance diversion matter? Ann Emerg Med 41(4):477–480, DOI 10.1067/mem.2003.112

[88] Redd JM, Bair AE, Jayaraman S (2003) Implications of ambulance diversion. Ann Emerg Med 42(4):S93

[89] Nakajima Y, Vilke GM (2015) Editorial: ambulance diversion: the con perspective. Am J Emerg Med 33(6):818–819, DOI 10.1016/j.ajem.2015.03.005

[90] Redelmeier DA, Blair PJ, Collins WE (1994) No place to unload: a preliminary analysis of the prevalence, risk factors, and consequences of ambulance diversion. Ann Emerg Med 23(1):43–47

[91] Mund E (2011) Ending ambulance diversion. Eighteen hospitals in King County, Wash., work toward a perpetual zero-divert status. EMS World 40(4):31–38

[92] Vilke GM, Brown L, Skogland P, Simmons C, Guss DA (2004) Approach to decreasing emergency department ambulance diversion hours. J Emerg Med 26(2):189–192, DOI 10.1016/j.jemermed.2003.07.003

[93] Shealy RM, Sorrell JF, French DM (2014) Ambulance diversion by cooperation: a positive experience with a physician-directed ambulance diversion policy in Charleston County, South Carolina. Ann Emerg Med 64(1):97–98, DOI 10.1016/j.annemergmed.2014.03.021

[94] Glushak C, Delbridge TR, Garrison HG (1997) Ambulance diversion. Prehosp Emerg Care 1(2):100–103

[95] McConnell KJ, Richards CF, Daya M, Weathers CC, Lowe RA (2006) Ambulance diversion and lost hospital revenues. Ann Emerg Med 48(6):702–710, DOI 10.1016/j.annemergmed.2006.05.001

[96] Williams RM (2006) Ambulance diversion: economic and policy considerations. Ann Emerg Med 48(6):711–712, DOI 10.1016/j.annemergmed.2006.06.009

[97] Weaver J (2007) Ed overcrowding and ambulance diversion cause potential liabilities. ED Legal Letter 18(9):97–101

[98] Upfold J (2002) Emergency department overcrowding: ambulance diversion and the legal duty to care. Canadian Medical Association Journal 166(4):445–446

[99] Litzenburg TA, Dorsey NB (2011) Ambulance diversion: solution or problem? ED Legal Letter 22(3):30–33

[100] Geiderman JM, Marco CA, Moskop JC, Adams J, Derse AR (2015) Ethics of ambulance diversion. Am J Emerg Med 33(6):822–827, DOI 10.1016/j.ajem.2014.12.002

[101] Adkins EJ, Werman HA (2015) Ambulance diversion: ethical dilemma and necessary evil. Am J Emerg Med 33(6):820–821, DOI 10.1016/j.ajem.2015.03.007

[102] Brennan JA, Allin DM, Calkins AM, Enguidanos ER, Heimbach LJ, S JNP, Stilley DG (2000) Guidelines for ambulance diversion. American College of Emergency Physicians. Ann Emerg Med 36(4):376–377

[103] Khaleghi M, Loh A, Vroman D, Chan TC, Vilke GM (2007) The effects of minimizing ambulance diversion hours on emergency departments. J Emerg Med 33(2):155–159, DOI 10.1016/j.jemermed.2007.02.014

[104] Lindstrom A (2009) Always open: study finds no adverse effects from stopping ambulance diversion. JEMS: Journal of Emergency Medical Services 34(10):16

[105] Patel PB, Derlet RW, Vinson DR, Williams M, Wills J (2006) Ambulance diversion reduction: the Sacramento solution. Am J Emerg Med 24(2):206–213, DOI 10.1016/j.ajem.2005.09.005

[106] Patel PB, Vinson DR (2012) Ambulance diversion reduction and elimination: the 3-2-1 plan. J Emerg Med 43(5):363, DOI 10.1016/j.jemermed.2012.01.031

[107] Friedman FD, Rathlev NK, White L, Epstein SK, Sayah A, Pearlmutter M, Biddinger P, Zane R, Moyer P (2011) Trial to end ambulance diversion in Boston: report from the conference of the Boston teaching hospitals consortium. Prehosp Disaster Med 26(2):122–126, DOI 10.1017/s1049023x11000070

[108] Lagoe RJ, Kohlbrenner JC, Hall LD, Roizen M, Nadle PA, Hunt RC (2003) Reducing ambulance diversion: a multihospital approach. Prehosp Emerg Care 7(1):99–108

[109] Barthell EN, Foldy SL, Pemble KR, Felton CW, Greischar PJ, Pirrallo RG, Bazan WJ (2003) Assuring community emergency care capacity with collaborative Internet tools: the Milwaukee experience. J Public Health Manag Pract 9(1):35–42

[110] Castillo EM, Vilke GM, Williams M, Turner P, Boyle J, Chan TC (2009) Collaborative to decrease ambulance diversion: the California ED diversion project. Ann Emerg Med 54(3):S78

[111] Poliakoff R, Vilke GM (2005) New ambulance policy slashes diversion hours: average number of patients drops 70%. ED Management 17(6):67–68

[112] Vilke GM, Castillo EM, Stepanski BM, Murrin PA, Upledger-Ray L, Metz MA, Chan TC (2006) San Diego county patient destination trial to decrease ambulance diversion hours: three year follow-up. Ann Emerg Med 48(4):S90

[113] Darrab AA, Fernandes CM, Worster A, Woolfrey K, Moneta S (2005) A city wide approach to reduce ambulance diversion: the Hamilton model. Ann Emerg Med 46(3):S41

[114] Burke L (2010) Ending ambulance diversion in Massachusetts. Virtual Mentor 12(6):483–486, DOI 10.1001/virtualmentor.2010.12.6.pfor2-1006

[115] Rathlev NK, Blank F, Osborne B, Kellogg A, Li H, Blanchet J, Conway RF, Durkin L, Gerstein R, Strzempko S, Vig M, Santoro JP, Visintainer P (2013) No diversion in Western Massachusetts. J Emerg Med 44(2):313–320, DOI 10.1016/j.jemermed.2012.06.017

[116] Burke LG, Joyce N, Baker WE, Biddinger PD, Dyer KS, Friedman FD, Imperato J, King A, Maciejko TM, Pearlmutter MD, Sayah A, Zane RD, Epstein SK (2013) The effect of an ambulance diversion ban on emergency department length of stay and ambulance turnaround time. Ann Emerg Med 61(3):311.e301, DOI 10.1016/j.annemergmed.2012.09.009

[117] O'Keefe SD, Bibi S, Rubin-Smith J, Feldman J (2014) "no diversion": a qualitative study of emergency medicine leaders in Boston, MA, and the effects of a statewide diversion ban policy. Ann Emerg Med 63(5):597.e587, DOI 10.1016/j.annemergmed.2013.09.007

[118] Holley J (2003) Memphis adopts no-ambulance-diversion policy: simple strategy returns rich rewards. EMS Insider 30(12):1–3

[119] Strear C, Vissers R, Yoder E, Barnett H, Shanks T, Jones L (2010) Applying the theory of constraints to emergency department workflow: reducing ambulance diversion through basic business practice. Ann Emerg Med 56(3):S11

[120] McLeod B, Zaver F, Avery C, Martin DP, Wang D, Jessen K, Lang ES (2010) Matching capacity to demand: a regional dashboard reduces ambulance avoidance and improves accessibility of receiving hospitals. Acad Emerg Med 17(12):1383–1389, DOI 10.1111/j.1553-2712.2010.00928.x

[121] El-Masri S, Saddik B (2012) An emergency system to improve ambulance dispatching, ambulance diversion and clinical handover communication - a proposed model. J Med Syst 36(6):3917–3923, DOI 10.1007/s10916-012-9863-x

[122] Beechner PM (2013) A fuzzy inference system for preventing ambulance diversion in emergency departments. Master thesis, State University of New York.

[123] Asamoah OK, Weiss SJ, Ernst AA, Richards M, Sklar DP (2008) A novel diversion protocol dramatically reduces diversion hours. Am J Emerg Med 26(6):670–675, DOI 10.1016/j.ajem.2007.10.020

[124] Pham JC, Patel R, Millin MG, Kirsch TD, Chanmugam A (2006) The effects of ambulance diversion: a comprehensive review. Acad Emerg Med 13(11):1220–1227, DOI 10.1197/j.aem.2006.05.024

[125] Delgado MK, Meng LJ, Mercer MP, Pines JM, Owens DK, Zaric GS (2013) Reducing ambulance diversion at hospital and regional levels: systemic review of insights from simulation models. West J Emerg Med 14(5):489–498, DOI 10.5811/westjem.2013.3.12788

[126] Shah MN, Fairbanks RJ, Maddow CL, Lerner EB, Syrett JI, Davis EA, Schneider SM (2006) Description and evaluation of a pilot physician-directed Emergency Medical Services diversion control program. Acad Emerg Med 13(1):54–60, DOI 10.1197/j.aem.2005.07.026

[127] Larson G (2008) Ambulance destination determination system for ambulance distribution as an alternative to ambulance diversion. J Emerg Nurs 34(4):357–358, DOI 10.1016/j.jen.2008.04.004

[128] Halliday MH, Bouland AJ, Lawner BJ, Comer AC, Ramos DC, Fletcher M (2016) The medical duty officer: an attempt to mitigate the ambulance at-hospital interval. West J Emerg Med 17(5):662–668, DOI 10.5811/westjem.2016.7.30266

[129] Snooks H, Foster T, Nicholl J (2004) Results of an evaluation of the effectiveness of triage and direct transportation to minor injuries units by ambulance crews. Emerg Med J 21(1):105–111

[130] Clawson JJ (1988) Principles of emergency medical dispatch. Englewood Cliffs, N.J.: Prentice-Hall, Englewood Cliffs, N.J.

[131] Shah MN, Bishop P, Lerner EB, Czapranski T, Davis EA (2003) Derivation of emergency medical services dispatch codes associated with low - acuity patients. Prehospital Emergency Care 7(4):434–439, DOI 10.1080/312703002132

[132] Shah MN, Bishop P, Lerner EB, Fairbanks RJ, Davis EA (2005) Validation of using EMS dispatch codes to identify low-acuity patients. Prehospital Emergency Care 9(1):24–31, DOI 10.1080/10903120590891651

[133] Woollard M (2003) Emergency calls not requiring an urgent ambulance response: expert consensus. Prehospital Emergency Care 7(3):384–391, DOI 10.1080/10903120390936626

[134] Villarreal M, Leach J, Kandala NB, Dale J (2017) Can a partnership between general practitioners and ambulance services reduce conveyance to emergency care? Emerg Med J 34:459–465, DOI 10.1136/emermed-2015-204924

[135] Millin MG, Brown LH, Schwartz B (2011) EMS provider determinations of necessity for transport and reimbursement for EMS response, medical care, and transport: combined resource document for the National Association of EMS Physicians position statements. Prehosp Emerg Care 15(4):562–569, DOI 10.3109/10903127.2011.598625

[136] Snooks HA, Dale J, Hartley-Sharpe C, Halter M (2004) On-scene alternatives for emergency ambulance crews attending patients who do not need to travel to the accident and emergency department: a review of the literature. Emergency medicine journal 21(2):212–215, DOI 10.1136/emj.2003.005199

[137] Schaefer RA, Rea TD, Plorde M, Peiguss K, Goldberg P, Murray JA (2002) An emergency medical services program of alternate destination of patient care. Prehospital Emergency Care 6(3):309–314, DOI 10.1080/10903120290938355

[138] Barishansky RM, O'Connor KE, Eckstein M, Chan LS (2004) The effect of emergency department crowding on ambulance availability. Ann Emerg Med 44(3):280–281

[139] Stewart D, Lang E, Wang D, Innes G (2019) Are emergency medical services offload delay patients at increased risk of adverse outcomes? Canadian Journal of Emergency Medicine pp 1–8

[140] Cooney DR, Vasisko C, Stimson K, Wojcik S (2013) Analysis of ambulance offload delay at an academic level 1 trauma center with adult and pediatric emergency departments. Ann Emerg Med 62(4):S2

[141] McNamara L (2017) Measuring emergency care network population coverage using location-allocation models and geographic information systems. Master thesis, Dalhousie University.

[142] (2015) Canadian Triage and Acuity Scale (CTAS) / Prehospital CTAS (Pre-CTAS). URL `http://caep.ca/resources/ctas`

[143] (2011) Ambulance diversion and emergency department offload delay. Prehosp Emerg Care 15(4):543, DOI 10.3109/10903127.2011.598620

[144] Kotsiantis S (2013) Decision trees: a recent overview. Artificial Intelligence Review 39(4):261–283, DOI 10.1007/s10462-011-9272-4

[145] Alpaydin E (2010) Introduction to machine learning, 2nd edn. Cambridge, Mass.: MIT Press, Cambridge, Mass.

[146] Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. Systems, Man and Cybernetics, IEEE Transactions on 21(3):660–674, DOI 10.1109/21.97458

[147] Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R (2014) Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Systems with Applications 41(4):1937–1946

[148] Loh WY (2011) Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1):14–23, DOI 10.1002/widm.8

[149] Tan PN (2006) Introduction to data mining. Boston: Pearson Addison Wesley, Boston

[150] Breiman L (1984) Classification and regression trees. New York, N.Y.: Chapman & Hall, New York, N.Y.

[151] Quinlan JR (1986) Induction of decision trees. Machine Learning 1(1):81–106, DOI 1022643204877

[152] Quinlan JR (1993) C4.5: programs for machine learning. San Mateo, Calif.: Morgan Kaufmann Publishers, San Mateo, Calif.

[153] Lim TS, Loh WY, Shih YS (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning 40(3):203–228, DOI 1007608224229

[154] Murthy S (1998) Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining and Knowledge Discovery 2(4):345–389, DOI 1009744630224

[155] Sankari ES, Manimegalai D (2017) Predicting membrane protein types using various decision tree classifiers based on various modes of general pseaac for imbalanced datasets. Journal of Theoretical Biology 435:208–217, DOI 10.1016/j.jtbi.2017.09.018

[156] Han L, Wang Y, Bryant SH (2008) Developing and validating predictive decision tree models from mining chemical structural fingerprints and high–throughput screening data in PubChem. BMC bioinformatics 9(1):401

[157] Chang CL, Chen CH (2009) Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Systems with Applications 36(2):4035–4041, DOI 10.1016/j.eswa.2008.03.007

[158] Handley TE, Hiles SA, Inder KJ, Kay-Lambkin F, Kelly BJ, Lewin TJ, Mcevoy M, Peel R, Attia JR (2014) Predictors of suicidal ideation in older people: a decision tree analysis. The American Journal of Geriatric Psychiatry 22(11):1325–1335, DOI 10.1016/j.jagp.2013.05.009

[159] Chen KH, Wang KJ, Tsai ML, Wang KM, Adrian AM, Cheng WC, Yang TS, Teng NC, Tan KP, Chang KS (2014) Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinformatics 15:49, DOI 10.1186/1471-2105-15-49

[160] Snousy MBA, El-Deeb H, Badran K, Khlil IAA (2011) Suite of decision tree-based classification algorithms on cancer gene expression data. Egyptian Informatics Journal 12(2):73–82, DOI 10.1016/j.eij.2011.04.003

[161] Rita BP, Maya BM, Gould M, Timothy MU, Bhattacharya J, Xiao Y, Khazeni N (2014) Demographic and clinical predictors of mortality from highly pathogenic avian influenza a H5N1 virus infection: CART analysis of international cases. PLoS ONE 9(3):e91,630, DOI 10.1371/journal.pone.0091630

[162] Luk JM, Lam BY, Lee NPY, Ho DW, Sham PC, Chen L, Peng J, Leng X, Day PJ, Fan ST (2007) Artificial neural networks and decision tree model analysis of liver cancer proteomes. Biochemical and Biophysical Research Communications 361(1):68–73, DOI 10.1016/j.bbrc.2007.06.172

[163] Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7(Jan):1–30

[164] Alpaydm E (1999) Combined 5× 2 cv F test for comparing supervised classification learning algorithms. Neural Computation 11(8):1885–1892

[165] Brazdil PB, Soares C (2000) A comparison of ranking methods for classification algorithm selection. In: European Conference on Machine Learning, Springer, pp 63–75

[166] Elomaa T, Kaariainen M (2001) An analysis of reduced error pruning. Journal of Artificial Intelligence Research 15:163–187

[167] Esposito F, Malerba D, Semeraro G, Kay J (1997) A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(5):476–491

[168] Quinlan JR (1999) Simplifying decision trees. International Journal of Human-Computer Studies 51(2):497–510

[169] Bradford JP, Kunz C, Kohavi R, Brunk C, Brodley CE (1998) Pruning decision trees with misclassification costs. In: European Conference on Machine Learning, Springer, pp 131–136

[170] Appavu alias Balamurugan S, Rajaram R (2009) Effective solution for unhandled exception in decision tree induction algorithms. Expert Systems with Applications 36(10):12,113–12,119

[171] Garofalakis M, Hyun D, Rastogi R, Shim K (2000) Efficient algorithms for constructing decision trees with constraints. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp 335–339

[172] Polat K, Güneş S (2009) A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Systems with Applications 36(2):1587–1592

[173] Chandra B, Varghese PP (2009) Fuzzifying Gini Index based decision trees. Expert Systems with Applications 36(4):8549–8559

[174] Aviad B, Roy G (2011) Classification by clustering decision tree-like classifier based on adjusted clusters. Expert Systems with Applications 38(7):8220–8228

[175] Li J, Weng J, Shao C, Guo H (2016) Cluster-based logistic regression model for holiday travel mode choice. Procedia Engineering 137:729–737

[176] De Caigny A, Coussement K, De Bock KW (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 269(2):760–772

[177] Aitkenhead MJ (2008) A co-evolving decision tree classification method. Expert Systems with Applications 34(1):18–25

[178] Llora X, Garrell JM (2001) Evolution of decision trees. In: Forth Catalan Conference on Artificial Intelligence (CCIA'2001), pp 115–122

[179] NEDOCS (2018) The standard for hospital - NEDOCS. URL https://www.nedocs.org/

[180] Hu Q, Yue W (2007) Markov decision processes with their applications, vol 14. Springer Science & Business Media

[181] White DJ (1993) A survey of applications of Markov decision processes. Journal of the Operational Research Society 44(11):1073–1096

[182] Alagoz O, Ayvaci MU, Linderoth JT (2015) Optimally solving Markov decision processes with total expected discounted reward function: linear programming revisited. Computers & Industrial Engineering 87:311–316

[183] Bandara D, Mayorga ME, McLay LA (2012) Optimal dispatching strategies for emergency vehicles to increase patient survivability. International Journal of Operational Research 15(2):195–214

[184] McLay LA, Mayorga ME (2013) A dispatching model for server-to-customer systems that balances efficiency and equity. Manufacturing & Service Operations Management 15(2):205–220

[185] McLay LA, Mayorga ME (2013) A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. IIE Transactions 45(1):1–24

[186] Jarvis JP (1981) Optimal assignments in a Markovian queueing system. Computers & Operations Research 8(1):17–23

[187] Keneally SK, Robbins MJ, Lunday BJ (2016) A Markov decision process model for the optimal dispatch of military medical evacuation assets. Health Care Management Science 19(2):111–129

[188] Alanis R, Ingolfsson A, Kolfal B (2013) A Markov chain model for an EMS system with repositioning. Production and Operations Management 22(1):216–231

[189] Berman O (1981) Dynamic repositioning of indistinguishable service units on transportation networks. Transportation Science 15(2):115–136

[190] Berman O (1981) Repositioning of distinguishable urban service units on networks. Computers & Operations Research 8(2):105–118

[191] Zhang L, Mason A, Philpott A (2010) The optimisation of a single ambulance moveup. Tech. rep., Faculty of Engineering, University of Auckland, New Zealand.

[192] Maxwell MS, Henderson SG, Topaloglu H (2009) Ambulance redeployment: an approximate dynamic programming approach. In: Winter Simulation Conference, WSC, pp 1850–1860

[193] Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. INFORMS Journal on Computing 22(2):266–281

[194] Puterman ML (2005) Markov decision processes: discrete stochastic dynamic programming. Hoboken, N.J.: J. Wiley, Hoboken, N.J.

[195] Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. Journal of Artificial Intelligence Research 4:237–285

[196] Ross J (2010) The patient journey through emergency care in Nova Scotia: a prescription for new medicine. Nova Scotia Department of Health

[197] Budge S, Ingolfsson A, Zerom D (2010) Empirical analysis of ambulance travel times: the case of Calgary emergency medical services. Management Science 56(4):716–723

[198] Kolesar P, Walker W, Hausner J (1975) Determining the relation between fire engine travel times and travel distances in New York city. Operations Research 23(4):614–627

Appendix A

Copyright Permission Letter

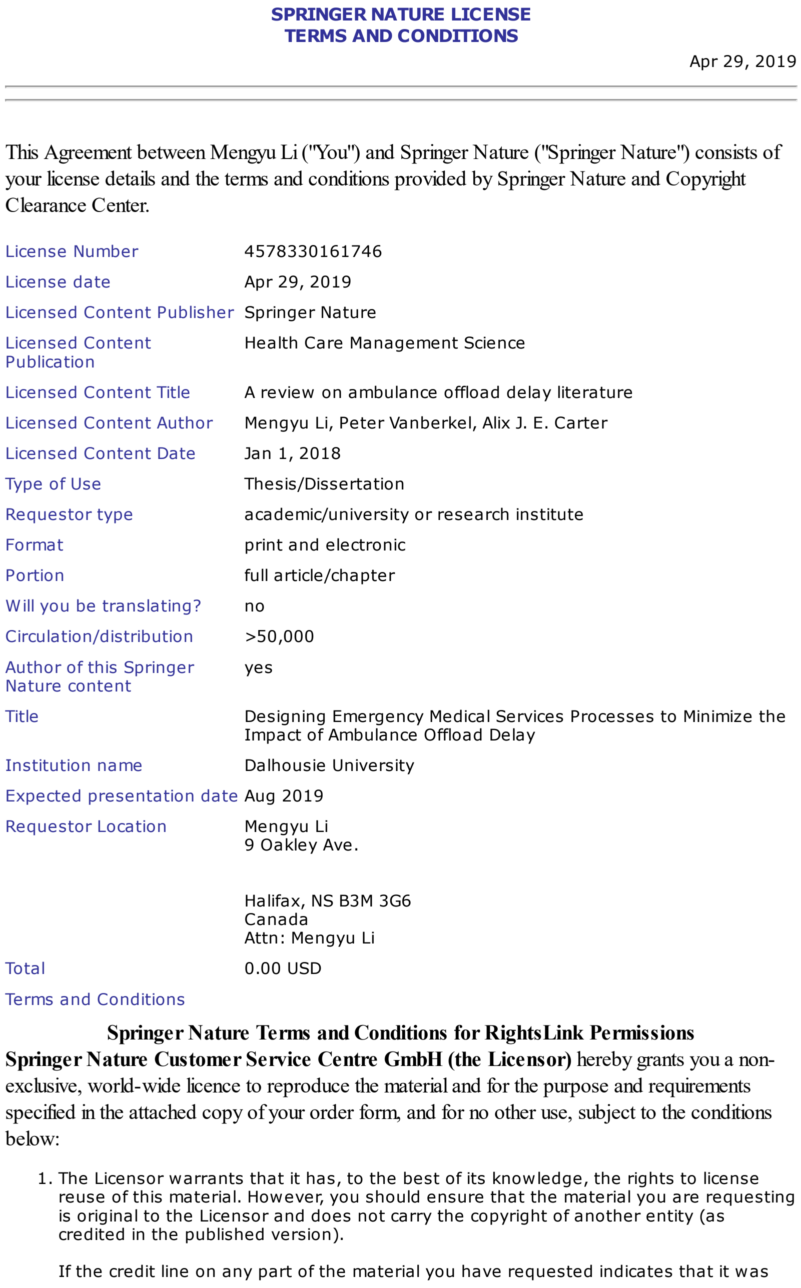
The thesis includes a manuscript version of the following paper as a chapter:

Mengyu Li, Peter Vanberkel, and Alix J. E. Carter (2018). A Review on Ambulance Offload Delay Literature. *Health Care Management Science.* https://doi.org/10.1007/s10729-018-9450-x.

Permission is granted by the licensed content publisher, Springer Nature, for:

- the inclusion of the material described above in my thesis.
- for the material described above to be included in the copy of my thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Full publication details and a copy of the permission letter has been included in Appendix A of this thesis.



SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Apr 29, 2019

This Agreement between Mengyu Li ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

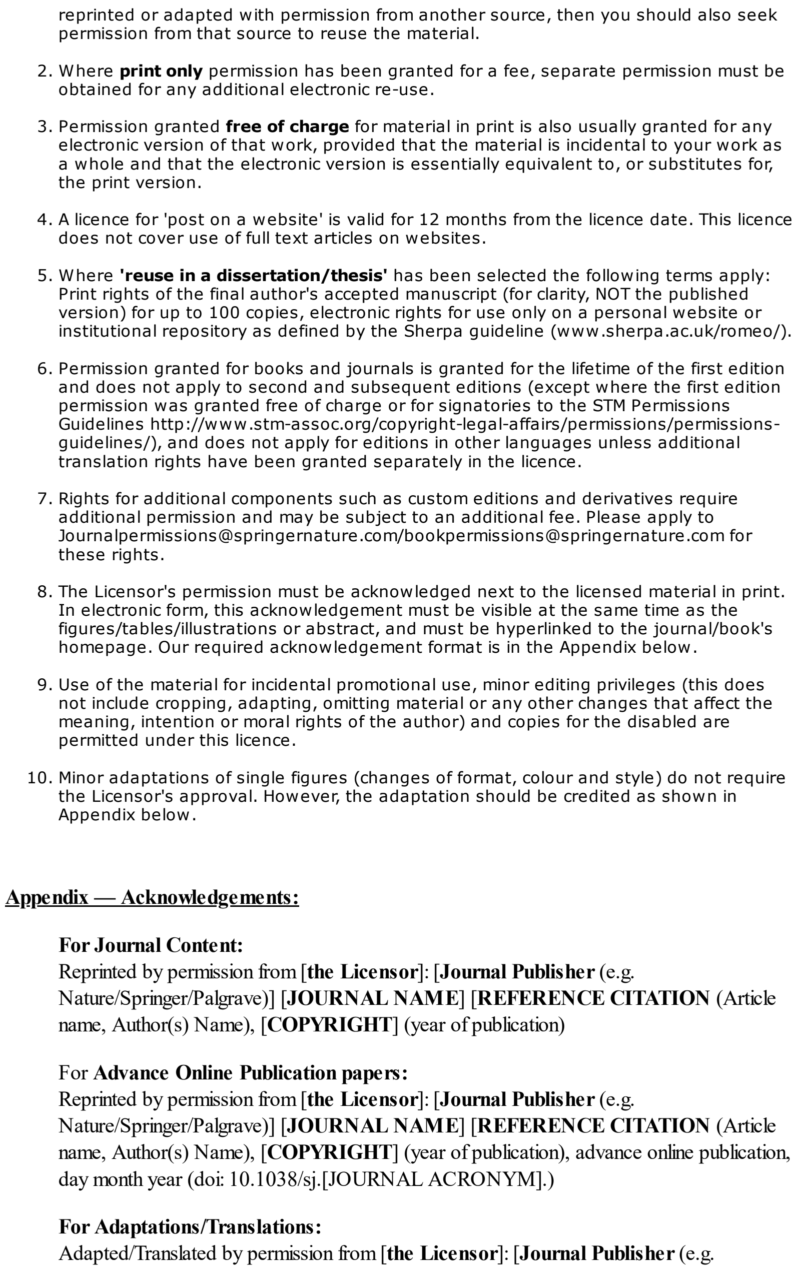
License Number	4578330161746
License date	Apr 29, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Health Care Management Science
Licensed Content Title	A review on ambulance offload delay literature
Licensed Content Author	Mengyu Li, Peter Vanberkel, Alix J. E. Carter
Licensed Content Date	Jan 1, 2018
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	>50,000
Author of this Springer Nature content	yes
Title	Designing Emergency Medical Services Processes to Minimize the Impact of Ambulance Offload Delay
Institution name	Dalhousie University
Expected presentation date	Aug 2019
Requestor Location	Mengyu Li 9 Oakley Ave. Halifax, NS B3M 3G6 Canada Attn: Mengyu Li
Total	0.00 USD
Terms and Conditions	

Springer Nature Terms and Conditions for RightsLink Permissions

Springer Nature Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was



reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

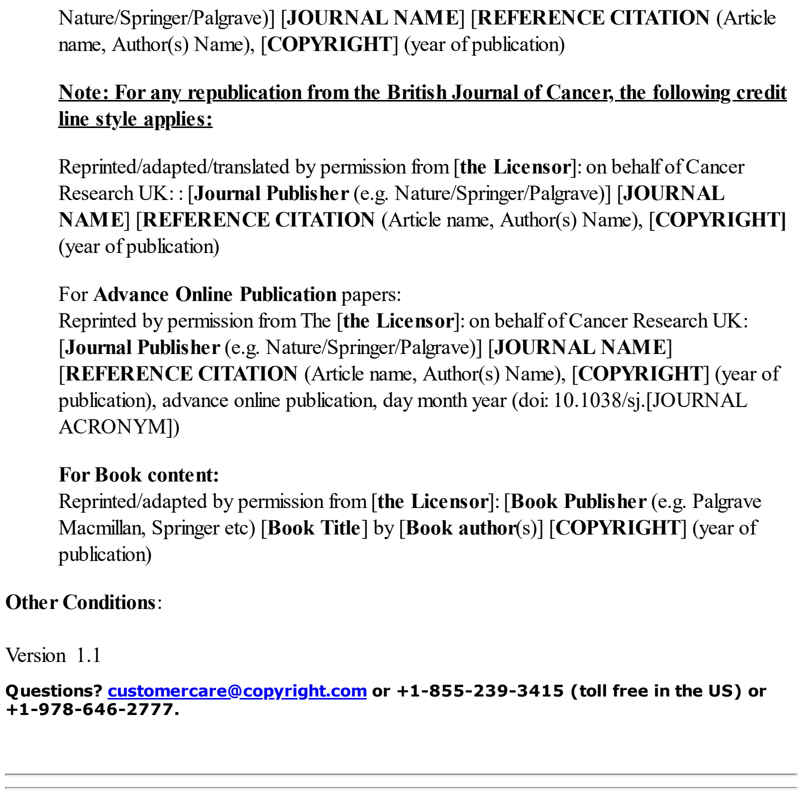
10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g.



Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT]** (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

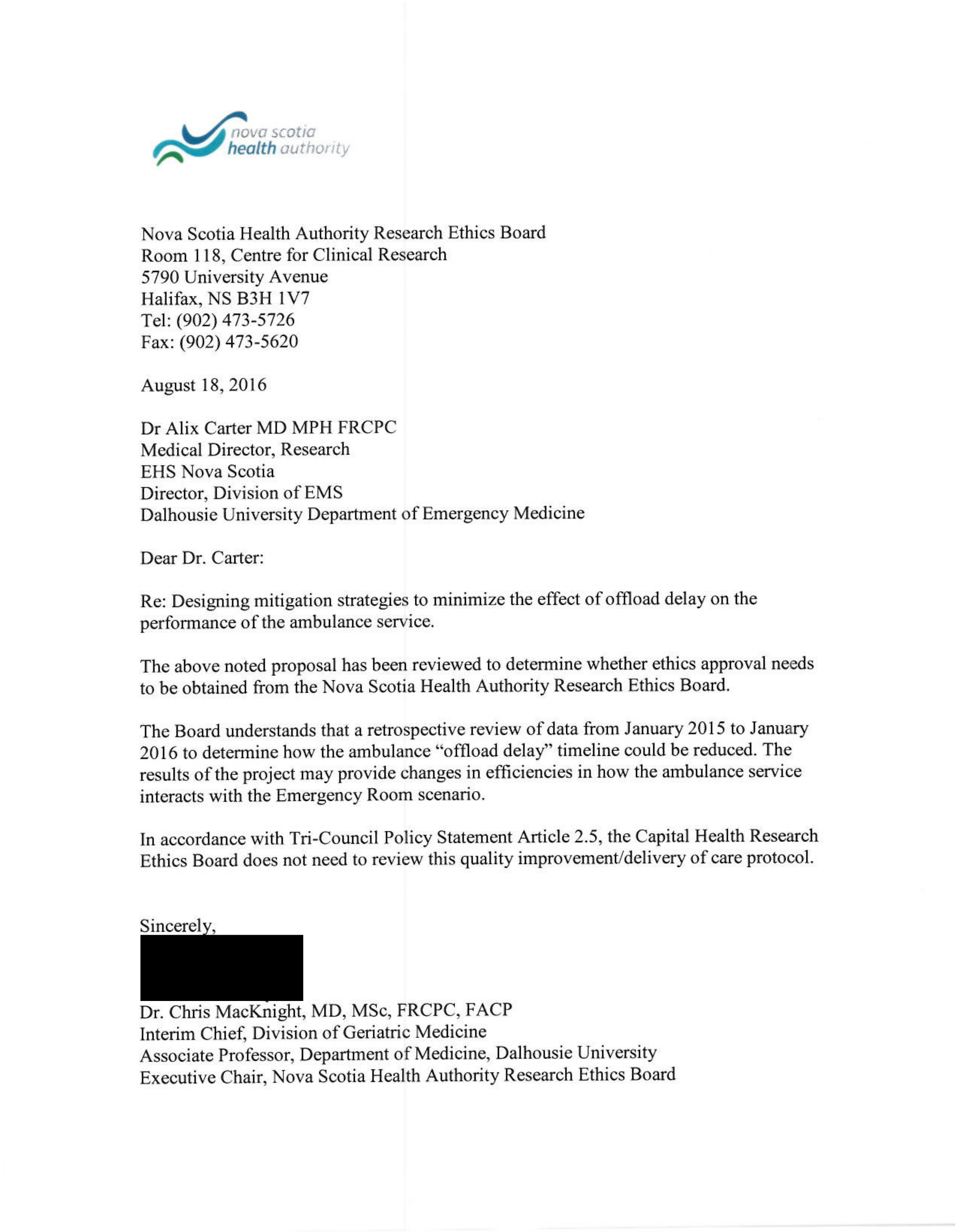
Version 1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix B

Ethics Statement

The data obtained for the research projects reported in this dissertation have been reviewed by the Nova Scotia Health Authority Research Ethics Board. The board deemed this research as a quality improvement / delivery of call protocol, therefore provided an ethical approval waiver. The full statement from the Nova Scotia Health Authority Research Ethics Board can be found in Appendix B.



Nova Scotia Health Authority Research Ethics Board
Room 118, Centre for Clinical Research
5790 University Avenue
Halifax, NS B3H 1V7
Tel: (902) 473-5726
Fax: (902) 473-5620

August 18, 2016

Dr Alix Carter MD MPH FRCPC
Medical Director, Research
EHS Nova Scotia
Director, Division of EMS
Dalhousie University Department of Emergency Medicine

Dear Dr. Carter:

Re: Designing mitigation strategies to minimize the effect of offload delay on the performance of the ambulance service.

The above noted proposal has been reviewed to determine whether ethics approval needs to be obtained from the Nova Scotia Health Authority Research Ethics Board.

The Board understands that a retrospective review of data from January 2015 to January 2016 to determine how the ambulance "offload delay" timeline could be reduced. The results of the project may provide changes in efficiencies in how the ambulance service interacts with the Emergency Room scenario.

In accordance with Tri-Council Policy Statement Article 2.5, the Capital Health Research Ethics Board does not need to review this quality improvement/delivery of care protocol.

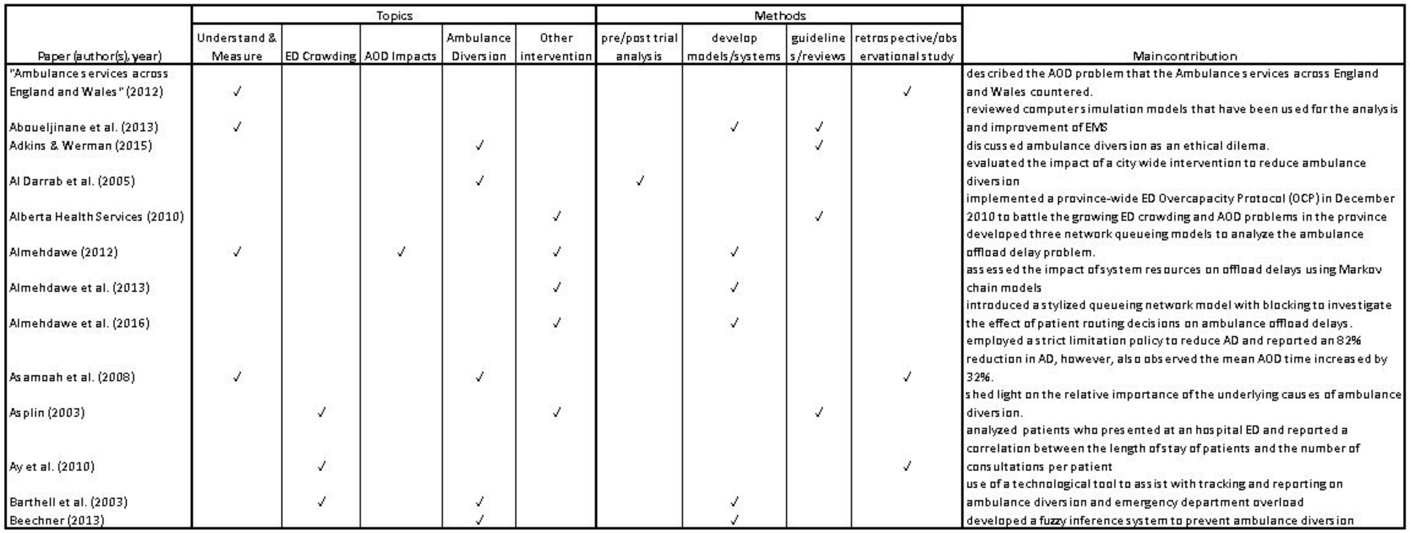
Sincerely,

[redacted]

Dr. Chris MacKnight, MD, MSc, FRCPC, FACP
Interim Chief, Division of Geriatric Medicine
Associate Professor, Department of Medicine, Dalhousie University
Executive Chair, Nova Scotia Health Authority Research Ethics Board

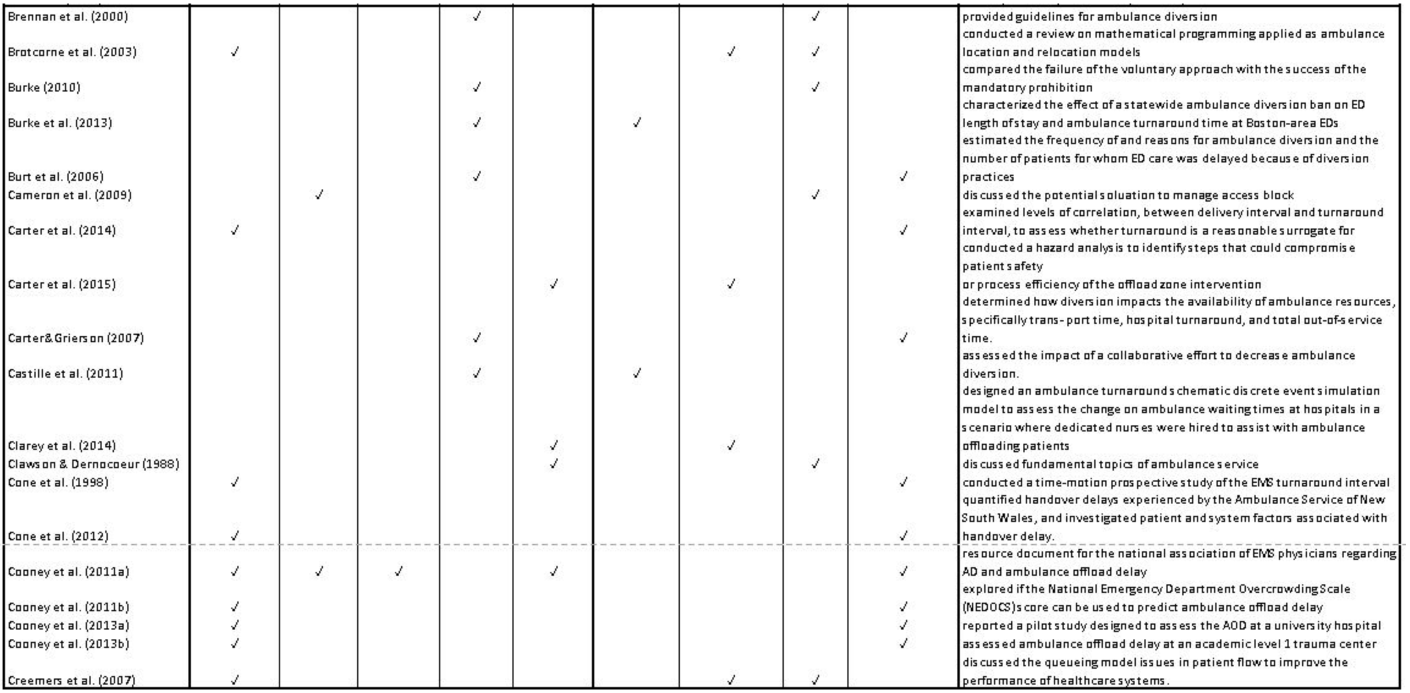
Appendix C

Literature Summary (Chapter 2)



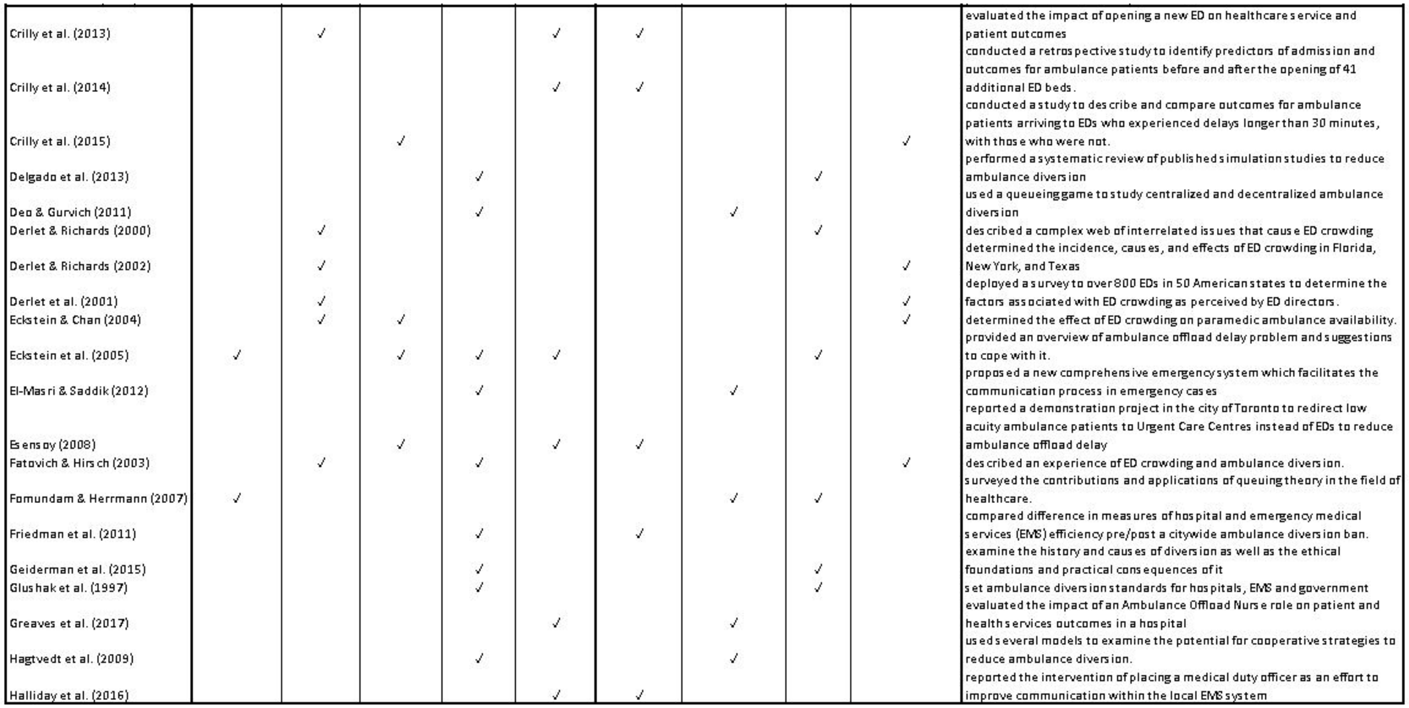
	Topics					Methods				
Paper (author(s), year)	Understand & Measure	ED Crowding	AOD Impacts	Ambulance Diversion	Other intervention	pre/post trial analysis	develop models/systems	guidelines/reviews	retrospective/observational study	Main contribution
"Ambulance services across England and Wales" (2012)	✓								✓	described the AOD problem that the Ambulance services across England and Wales countered.
Aboueljinane et al. (2013)	✓						✓	✓		reviewed computer simulation models that have been used for the analysis and improvement of EMS
Adkins & Werman (2015)				✓				✓		discussed ambulance diversion as an ethical dilema.
Al Darrab et al. (2005)				✓		✓				evaluated the impact of a city wide intervention to reduce ambulance diversion
Alberta Health Services (2010)					✓			✓		implemented a province-wide ED Overcapacity Protocol (OCP) in December 2010 to battle the growing ED crowding and AOD problems in the province
Almehdawe (2012)	✓		✓		✓		✓			developed three network queueing models to analyze the ambulance offload delay problem.
Almehdawe et al. (2013)					✓		✓			assessed the impact of system resources on offload delays using Markov chain models
Almehdawe et al. (2016)					✓		✓			introduced a stylized queueing network model with blocking to investigate the effect of patient routing decisions on ambulance offload delays.
Asamoah et al. (2008)	✓			✓					✓	employed a strict limitation policy to reduce AD and reported an 82% reduction in AD, however, also observed the mean AOD time increased by 32%.
Asplin (2003)		✓			✓			✓		shed light on the relative importance of the underlying causes of ambulance diversion.
Ay et al. (2010)		✓							✓	analyzed patients who presented at an hospital ED and reported a correlation between the length of stay of patients and the number of consultations per patient
Barthell et al. (2003)		✓		✓			✓			use of a technological tool to assist with tracking and reporting on ambulance diversion and emergency department overload
Beechner (2013)				✓			✓			developed a fuzzy inference system to prevent ambulance diversion

Figure C.1: Literature Summary Table



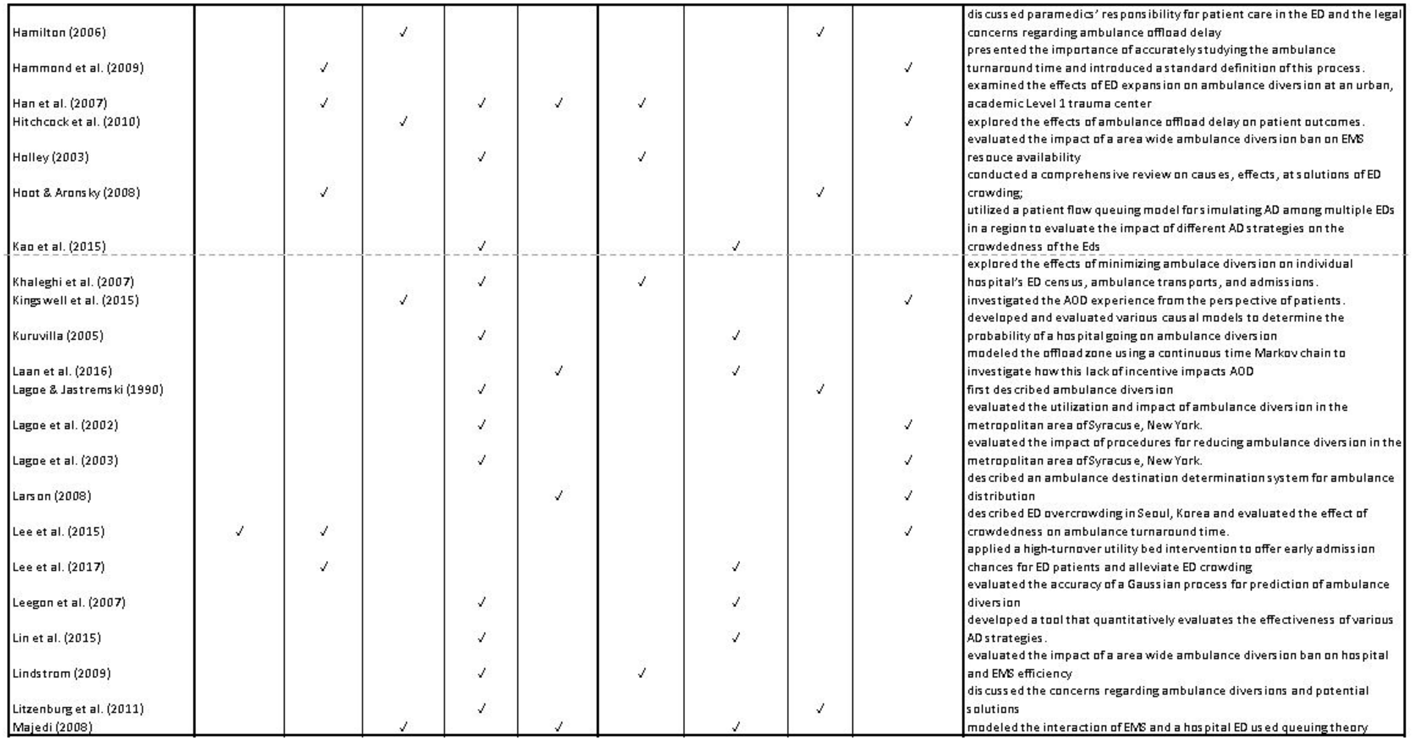
Brennan et al. (2000)				✓				✓		provided guidelines for ambulance diversion
Bratcorne et al. (2003)	✓						✓	✓		conducted a review on mathematical programming applied as ambulance location and relocation models
Burke (2010)				✓				✓		compared the failure of the voluntary approach with the success of the mandatory prohibition
Burke et al. (2013)				✓		✓				characterized the effect of a statewide ambulance diversion ban on ED length of stay and ambulance turnaround time at Boston-area EDs
Burt et al. (2006)				✓					✓	estimated the frequency of and reasons for ambulance diversion and the number of patients for whom ED care was delayed because of diversion practices
Cameron et al. (2009)		✓						✓		discussed the potential soluation to manage access block
Carter et al. (2014)	✓								✓	examined levels of correlation, between delivery interval and turnaround interval, to assess whether turnaround is a reasonable surrogate for
Carter et al. (2015)					✓		✓			conducted a hazard analysis to identify steps that could compromise patient safety or process efficiency of the offload zone intervention
Carter&Grierson (2007)				✓					✓	determined how diversion impacts the availability of ambulance resources, specifically trans- port time, hospital turnaround, and total out-of-service time.
Castille et al. (2011)				✓		✓				assessed the impact of a collaborative effort to decrease ambulance diversion.
Clarey et al. (2014)					✓		✓			designed an ambulance turnaround schematic discrete event simulation model to assess the change on ambulance waiting times at hospitals in a scenario where dedicated nurses were hired to assist with ambulance offloading patients
Clawson & Dernocoeur (1988)					✓			✓		discussed fundamental topics of ambulance service
Cone et al. (1998)	✓								✓	conducted a time-motion prospective study of the EMS turnaround interval
Cone et al. (2012)	✓								✓	quantified handover delays experienced by the Ambulance Service of New South Wales, and investigated patient and system factors associated with handover delay.
Cooney et al. (2011a)	✓	✓	✓		✓				✓	resource document for the national association of EMS physicians regarding AD and ambulance offload delay
Cooney et al. (2011b)	✓								✓	explored if the National Emergency Department Overcrowding Scale (NEDOCS) score can be used to predict ambulance offload delay
Cooney et al. (2013a)	✓								✓	reported a pilot study designed to assess the AOD at a university hospital
Cooney et al. (2013b)	✓								✓	assessed ambulance offload delay at an academic level 1 trauma center
Creemers et al. (2007)	✓						✓	✓		discussed the queueing model issues in patient flow to improve the performance of healthcare systems.

Figure C.1: Literature Summary Table (cont.)



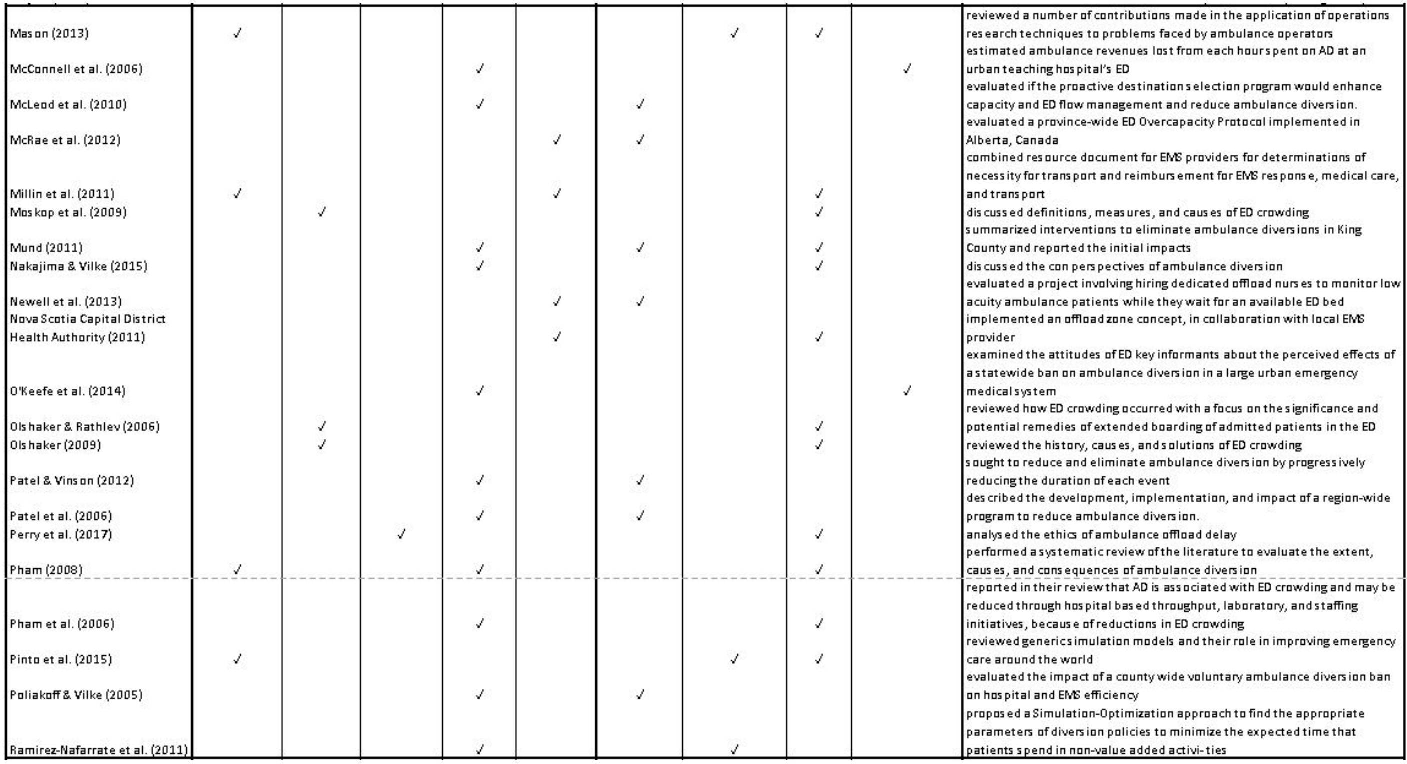
Crilly et al. (2013)		✓			✓	✓				evaluated the impact of opening a new ED on healthcare service and patient outcomes
Crilly et al. (2014)					✓	✓				conducted a retrospective study to identify predictors of admission and outcomes for ambulance patients before and after the opening of 41 additional ED beds.
Crilly et al. (2015)			✓						✓	conducted a study to describe and compare outcomes for ambulance patients arriving to EDs who experienced delays longer than 30 minutes, with those who were not.
Delgado et al. (2013)				✓				✓		performed a systematic review of published simulation studies to reduce ambulance diversion
Deo & Gurvich (2011)				✓			✓			used a queueing game to study centralized and decentralized ambulance diversion
Derlet & Richards (2000)		✓						✓		described a complex web of interrelated issues that cause ED crowding
Derlet & Richards (2002)		✓							✓	determined the incidence, causes, and effects of ED crowding in Florida, New York, and Texas
Derlet et al. (2001)		✓							✓	deployed a survey to over 800 EDs in 50 American states to determine the factors associated with ED crowding as perceived by ED directors.
Eckstein & Chan (2004)		✓	✓						✓	determined the effect of ED crowding on paramedic ambulance availability.
Eckstein et al. (2005)	✓		✓	✓	✓			✓		provided an overview of ambulance offload delay problem and suggestions to cope with it.
El-Masri & Saddik (2012)				✓			✓			proposed a new comprehensive emergency system which facilitates the communication process in emergency cases
Esensoy (2008)			✓		✓	✓				reported a demonstration project in the city of Toronto to redirect low acuity ambulance patients to Urgent Care Centres instead of EDs to reduce ambulance offload delay
Fatovich & Hirsch (2003)		✓		✓					✓	described an experience of ED crowding and ambulance diversion.
Fomundam & Herrmann (2007)	✓						✓	✓		surveyed the contributions and applications of queuing theory in the field of healthcare.
Friedman et al. (2011)				✓		✓				compared difference in measures of hospital and emergency medical services (EMS) efficiency pre/post a citywide ambulance diversion ban.
Geiderman et al. (2015)				✓				✓		examine the history and causes of diversion as well as the ethical foundations and practical consequences of it
Glushak et al. (1997)				✓				✓		set ambulance diversion standards for hospitals, EMS and government
Greaves et al. (2017)					✓		✓			evaluated the impact of an Ambulance Offload Nurse role on patient and health services outcomes in a hospital
Hagtvedt et al. (2009)				✓			✓			used several models to examine the potential for cooperative strategies to reduce ambulance diversion.
Halliday et al. (2016)					✓	✓				reported the intervention of placing a medical duty officer as an effort to improve communication within the local EMS system

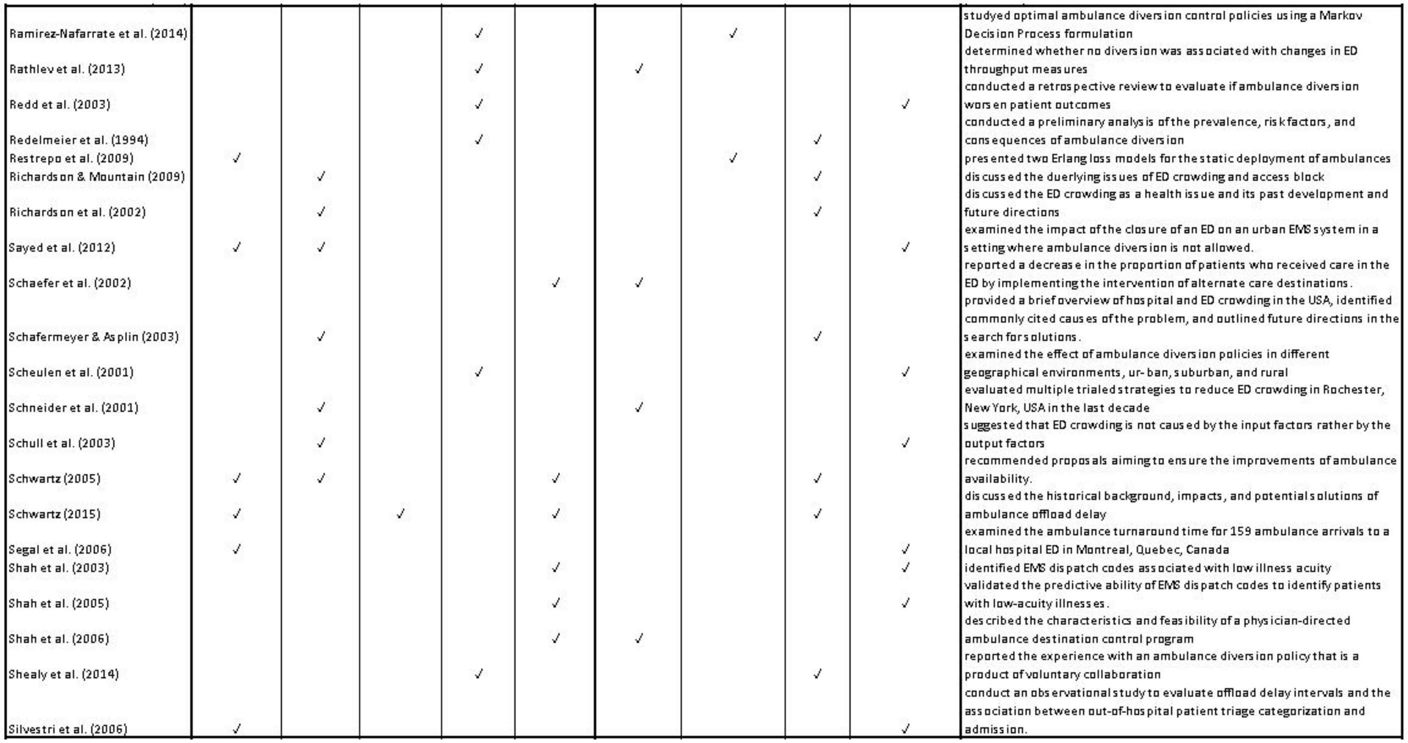
Figure C.1: Literature Summary Table (cont.)



Hamilton (2006)			✓					✓		discussed paramedics' responsibility for patient care in the ED and the legal concerns regarding ambulance offload delay
Hammond et al. (2009)		✓							✓	presented the importance of accurately studying the ambulance turnaround time and introduced a standard definition of this process.
Han et al. (2007)		✓		✓	✓	✓				examined the effects of ED expansion on ambulance diversion at an urban, academic Level 1 trauma center
Hitchcock et al. (2010)			✓						✓	explored the effects of ambulance offload delay on patient outcomes.
Holley (2003)				✓		✓				evaluated the impact of a area wide ambulance diversion ban on EMS resouce availability
Hoot & Aronsky (2008)		✓						✓		conducted a comprehensive review on causes, effects, at solutions of ED crowding;
Kao et al. (2015)				✓			✓			utilized a patient flow queuing model for simulating AD among multiple EDs in a region to evaluate the impact of different AD strategies on the crowdedness of the Eds
Khaleghi et al. (2007)				✓		✓				explored the effects of minimizing ambulace diversion on individual hospital's ED census, ambulance transports, and admissions.
Kingswell et al. (2015)			✓						✓	investigated the AOD experience from the perspective of patients.
Kuruvilla (2005)				✓			✓			developed and evaluated various causal models to determine the probability of a hospital going on ambulance diversion
Laan et al. (2016)					✓		✓			modeled the offload zone using a continuous time Markov chain to investigate how this lack of incentive impacts AOD
Lagoe & Jastremski (1990)				✓				✓		first described ambulance diversion
Lagoe et al. (2002)				✓					✓	evaluated the utilization and impact of ambulance diversion in the metropolitan area of Syracuse, New York.
Lagoe et al. (2003)				✓					✓	evaluated the impact of procedures for reducing ambulance diversion in the metropolitan area of Syracuse, New York.
Larson (2008)					✓				✓	described an ambulance destination determination system for ambulance distribution
Lee et al. (2015)	✓	✓							✓	described ED overcrowding in Seoul, Korea and evaluated the effect of crowdedness on ambulance turnaround time.
Lee et al. (2017)		✓					✓			applied a high-turnover utility bed intervention to offer early admission chances for ED patients and alleviate ED crowding
Leegon et al. (2007)				✓			✓			evaluated the accuracy of a Gaussian process for prediction of ambulance diversion
Lin et al. (2015)				✓			✓			developed a tool that quantitatively evaluates the effectiveness of various AD strategies.
Lindstrom (2009)				✓		✓				evaluated the impact of a area wide ambulance diversion ban on hospital and EMS efficiency
Litzenburg et al. (2011)				✓				✓		discussed the concerns regarding ambulance diversions and potential solutions
Majedi (2008)			✓		✓		✓			modeled the interaction of EMS and a hospital ED used queuing theory

Figure C.1: Literature Summary Table (cont.)





Ramirez-Nafarrate et al. (2014)				✓			✓			studyed optimal ambulance diversion control policies using a Markov Decision Process formulation
Rathlev et al. (2013)				✓		✓				determined whether no diversion was associated with changes in ED throughput measures
Redd et al. (2003)				✓					✓	conducted a retrospective review to evaluate if ambulance diversion worsen patient outcomes
Redelmeier et al. (1994)				✓				✓		conducted a preliminary analysis of the prevalence, risk factors, and consequences of ambulance diversion
Restrepo et al. (2009)	✓						✓			presented two Erlang loss models for the static deployment of ambulances
Richardson & Mountain (2009)		✓						✓		discussed the duerlying issues of ED crowding and access block
Richardson et al. (2002)		✓						✓		discussed the ED crowding as a health issue and its past development and future directions
Sayed et al. (2012)	✓	✓							✓	examined the impact of the closure of an ED on an urban EMS system in a setting where ambulance diversion is not allowed.
Schaefer et al. (2002)					✓	✓				reported a decrease in the proportion of patients who received care in the ED by implementing the intervention of alternate care destinations.
Schafermeyer & Asplin (2003)		✓						✓		provided a brief overview of hospital and ED crowding in the USA, identified commonly cited causes of the problem, and outlined future directions in the search for solutions.
Scheulen et al. (2001)				✓					✓	examined the effect of ambulance diversion policies in different geographical environments, ur- ban, suburban, and rural
Schneider et al. (2001)		✓				✓				evaluated multiple trialed strategies to reduce ED crowding in Rochester, New York, USA in the last decade
Schull et al. (2003)		✓							✓	suggested that ED crowding is not caused by the input factors rather by the output factors
Schwartz (2005)	✓	✓			✓			✓		recommended proposals aiming to ensure the improvements of ambulance availability.
Schwartz (2015)	✓		✓		✓			✓		discussed the historical background, impacts, and potential solutions of ambulance offload delay
Segal et al. (2006)	✓								✓	examined the ambulance turnaround time for 159 ambulance arrivals to a local hospital ED in Montreal, Quebec, Canada
Shah et al. (2003)					✓				✓	identified EMS dispatch codes associated with low illness acuity
Shah et al. (2005)					✓				✓	validated the predictive ability of EMS dispatch codes to identify patients with low-acuity illnesses.
Shah et al. (2006)					✓	✓				described the characteristics and feasibility of a physician-directed ambulance destination control program
Shealy et al. (2014)				✓				✓		reported the experience with an ambulance diversion policy that is a product of voluntary collaboration
Silvestri et al. (2006)	✓								✓	conduct an observational study to evaluate offload delay intervals and the association between out-of-hospital patient triage categorization and admission.

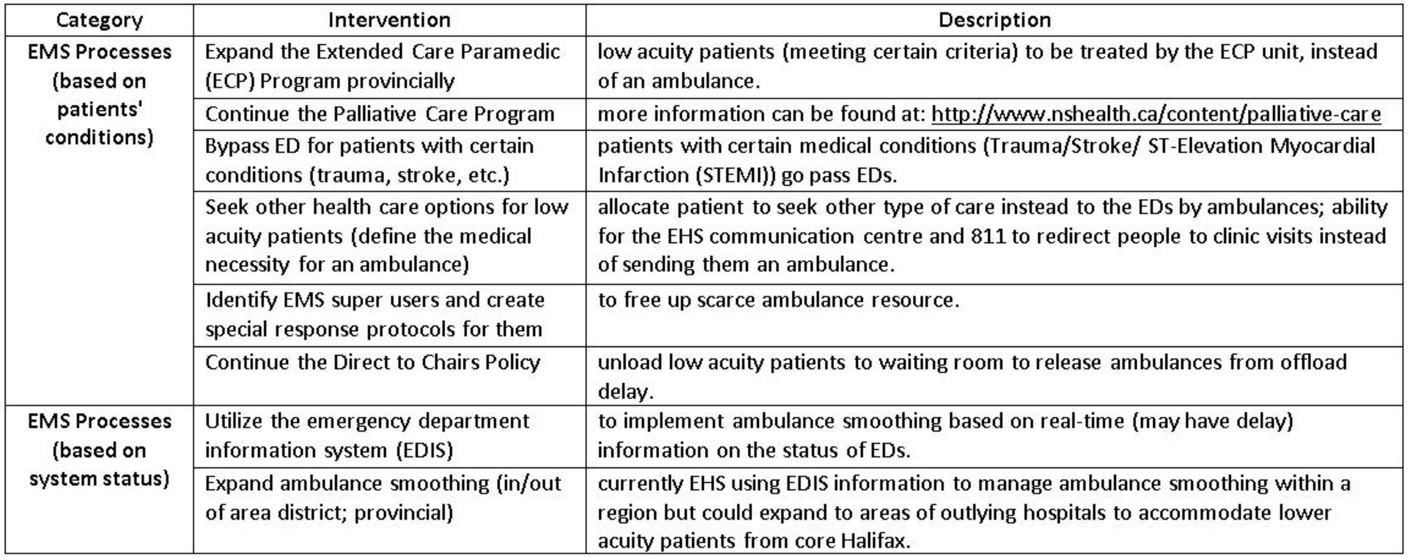
Figure C.1: Literature Summary Table (cont.)

Silvestri et al. (2006)					✓	✓				examined the impact of ED bed availability on the off-load time of EMS units.
Singer & Donoso (2008)	✓						✓			described an ambulance service in terms of its main operation parameters and strategic decision variables
Smith (2013)	✓		✓				✓			reported ambulance offload delay as a costly issue for England ambulance service providers.
Snooks et al. (2004)					✓				✓	evaluated triage and transportation to a minor injury unit instead of ED by ambulance
Snooks et al. (2004)					✓			✓		reviewed the literature concerning on-scene alternatives to conveyance to an ED
Spaite et al. (1995)	✓							✓		defined the “delivery interval” component of the ambulance turnaround time
Steer et al. (2016)	✓								✓	introduced radio frequency identification as a novel method for monitoring offload times and identifying variance
Strear et al. (2010)				✓			✓			applied the Theory of Constraints to ED workflow to reducing ambulance diversion
Takeda et al. (2007)	✓						✓			studied the application of the hypercube queueing model to the urban EMS of Campinas in Brazil
Taylor et al. (2006)	✓								✓	determined the difference between the recorded arrival of an ambulance outside an ED and the actual delivery of the patient to the clinical area of the ED.
Ting (2008)			✓					✓		discussed the potential adverse patient effects of ambulance offload delay
Upfold (2002)		✓		✓				✓		discussed the legal concern of ED crowding and ambulance diversion
Vilke et al. (2004)				✓		✓				concluded that reciprocating effects can be decreased with one institution’s commitment to avoid diversion, thus decreasing the need for diversion at a neigh- boring facility
Vilke et al. (2004)				✓		✓				evaluated a community intervention to reduce ambulance diversion.
Vilke et al. (2006)				✓		✓				followed up with a voluntary ambulance diversion ban in San Diego County after three years
Villarreal et al. (2017)					✓		✓			investigated a new model of patient screening where a partnership between general practitioners and ambulance services was formed to reduce conveyance rates to the Hospital EDs.
Warden et al. (2003)				✓					✓	evaluated the amount of ambulance diversion in an EMS system and investigated potential predictive factors
Weaver (2007)				✓				✓		discussed the potential liabilities caused by ED overcrowding and ambulance diversion
Williams (2006)				✓				✓		examined the ambulance diversion issue from the financial perspective of an individual institution
Woollard (2003)					✓				✓	sought expert consensus about which ambulance dispatch codes could be appropriate for a nonemergency response
Yancer et al. (2006)		✓		✓		✓				summarized interventions to decrease ambulance diversions and ED crowding in a hospital

Figure C.1: Literature Summary Table (cont.)

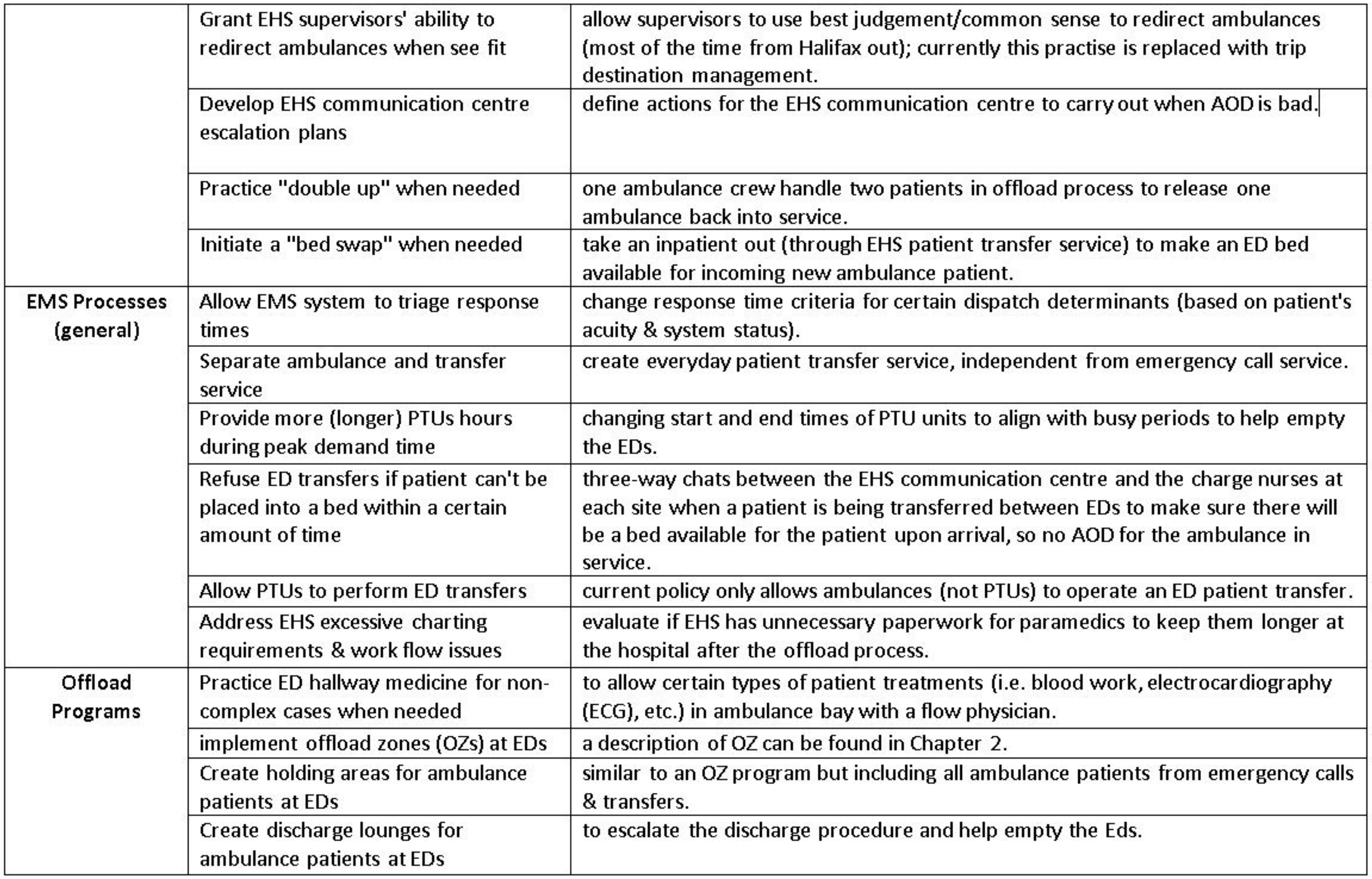
Appendix D

Summary of interventions from the focus group discussions (Chapter 5)



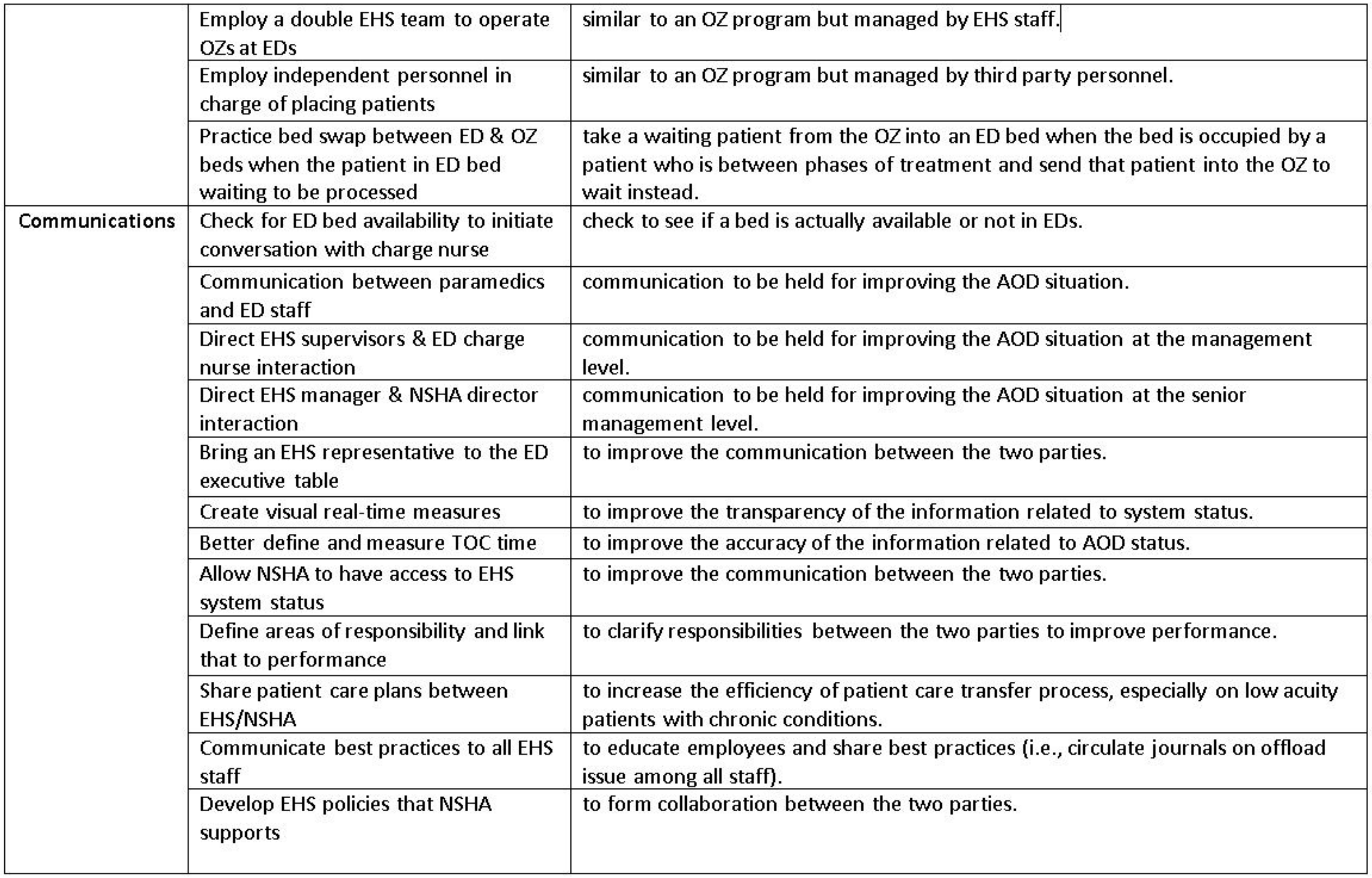
Category	Intervention	Description
EMS Processes (based on patients' conditions)	Expand the Extended Care Paramedic (ECP) Program provincially	low acuity patients (meeting certain criteria) to be treated by the ECP unit, instead of an ambulance.
	Continue the Palliative Care Program	more information can be found at: http://www.nshealth.ca/content/palliative-care
	Bypass ED for patients with certain conditions (trauma, stroke, etc.)	patients with certain medical conditions (Trauma/Stroke/ ST-Elevation Myocardial Infarction (STEMI)) go pass EDs.
	Seek other health care options for low acuity patients (define the medical necessity for an ambulance)	allocate patient to seek other type of care instead to the EDs by ambulances; ability for the EHS communication centre and 811 to redirect people to clinic visits instead of sending them an ambulance.
	Identify EMS super users and create special response protocols for them	to free up scarce ambulance resource.
	Continue the Direct to Chairs Policy	unload low acuity patients to waiting room to release ambulances from offload delay.
EMS Processes (based on system status)	Utilize the emergency department information system (EDIS)	to implement ambulance smoothing based on real-time (may have delay) information on the status of EDs.
	Expand ambulance smoothing (in/out of area district; provincial)	currently EHS using EDIS information to manage ambulance smoothing within a region but could expand to areas of outlying hospitals to accommodate lower acuity patients from core Halifax.

Figure D.1: The summary of interventions obtained from the EHS focus group discussions



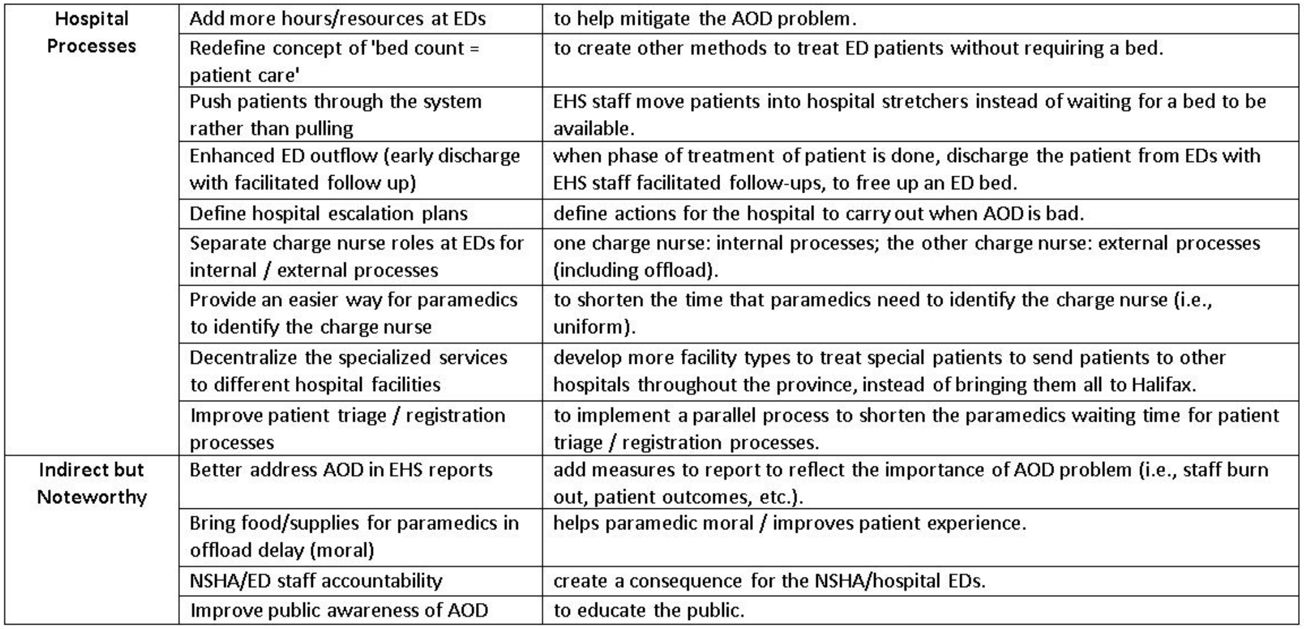
	Grant EHS supervisors' ability to redirect ambulances when see fit	allow supervisors to use best judgement/common sense to redirect ambulances (most of the time from Halifax out); currently this practise is replaced with trip destination management.
	Develop EHS communication centre escalation plans	define actions for the EHS communication centre to carry out when AOD is bad.
	Practice "double up" when needed	one ambulance crew handle two patients in offload process to release one ambulance back into service.
	Initiate a "bed swap" when needed	take an inpatient out (through EHS patient transfer service) to make an ED bed available for incoming new ambulance patient.
EMS Processes (general)	Allow EMS system to triage response times	change response time criteria for certain dispatch determinants (based on patient's acuity & system status).
	Separate ambulance and transfer service	create everyday patient transfer service, independent from emergency call service.
	Provide more (longer) PTUs hours during peak demand time	changing start and end times of PTU units to align with busy periods to help empty the EDs.
	Refuse ED transfers if patient can't be placed into a bed within a certain amount of time	three-way chats between the EHS communication centre and the charge nurses at each site when a patient is being transferred between EDs to make sure there will be a bed available for the patient upon arrival, so no AOD for the ambulance in service.
	Allow PTUs to perform ED transfers	current policy only allows ambulances (not PTUs) to operate an ED patient transfer.
	Address EHS excessive charting requirements & work flow issues	evaluate if EHS has unnecessary paperwork for paramedics to keep them longer at the hospital after the offload process.
Offload Programs	Practice ED hallway medicine for non-complex cases when needed	to allow certain types of patient treatments (i.e. blood work, electrocardiography (ECG), etc.) in ambulance bay with a flow physician.
	implement offload zones (OZs) at EDs	a description of OZ can be found in Chapter 2.
	Create holding areas for ambulance patients at EDs	similar to an OZ program but including all ambulance patients from emergency calls & transfers.
	Create discharge lounges for ambulance patients at EDs	to escalate the discharge procedure and help empty the Eds.

Figure D.1: The summary of interventions obtained from the EHS focus group discussions (cont.)



	Employ a double EHS team to operate OZs at EDs	similar to an OZ program but managed by EHS staff.
	Employ independent personnel in charge of placing patients	similar to an OZ program but managed by third party personnel.
	Practice bed swap between ED & OZ beds when the patient in ED bed waiting to be processed	take a waiting patient from the OZ into an ED bed when the bed is occupied by a patient who is between phases of treatment and send that patient into the OZ to wait instead.
Communications	Check for ED bed availability to initiate conversation with charge nurse	check to see if a bed is actually available or not in EDs.
	Communication between paramedics and ED staff	communication to be held for improving the AOD situation.
	Direct EHS supervisors & ED charge nurse interaction	communication to be held for improving the AOD situation at the management level.
	Direct EHS manager & NSHA director interaction	communication to be held for improving the AOD situation at the senior management level.
	Bring an EHS representative to the ED executive table	to improve the communication between the two parties.
	Create visual real-time measures	to improve the transparency of the information related to system status.
	Better define and measure TOC time	to improve the accuracy of the information related to AOD status.
	Allow NSHA to have access to EHS system status	to improve the communication between the two parties.
	Define areas of responsibility and link that to performance	to clarify responsibilities between the two parties to improve performance.
	Share patient care plans between EHS/NSHA	to increase the efficiency of patient care transfer process, especially on low acuity patients with chronic conditions.
	Communicate best practices to all EHS staff	to educate employees and share best practices (i.e., circulate journals on offload issue among all staff).
	Develop EHS policies that NSHA supports	to form collaboration between the two parties.

Figure D.1: The summary of interventions obtained from the EHS focus group discussions (cont.)



Hospital Processes	Add more hours/resources at EDs	to help mitigate the AOD problem.
	Redefine concept of 'bed count = patient care'	to create other methods to treat ED patients without requiring a bed.
	Push patients through the system rather than pulling	EHS staff move patients into hospital stretchers instead of waiting for a bed to be available.
	Enhanced ED outflow (early discharge with facilitated follow up)	when phase of treatment of patient is done, discharge the patient from EDs with EHS staff facilitated follow-ups, to free up an ED bed.
	Define hospital escalation plans	define actions for the hospital to carry out when AOD is bad.
	Separate charge nurse roles at EDs for internal / external processes	one charge nurse: internal processes; the other charge nurse: external processes (including offload).
	Provide an easier way for paramedics to identify the charge nurse	to shorten the time that paramedics need to identify the charge nurse (i.e., uniform).
	Decentralize the specialized services to different hospital facilities	develop more facility types to treat special patients to send patients to other hospitals throughout the province, instead of bringing them all to Halifax.
	Improve patient triage / registration processes	to implement a parallel process to shorten the paramedics waiting time for patient triage / registration processes.
Indirect but Noteworthy	Better address AOD in EHS reports	add measures to report to reflect the importance of AOD problem (i.e., staff burn out, patient outcomes, etc.).
	Bring food/supplies for paramedics in offload delay (moral)	helps paramedic moral / improves patient experience.
	NSHA/ED staff accountability	create a consequence for the NSHA/hospital EDs.
	Improve public awareness of AOD	to educate the public.

Figure D.1: The summary of interventions obtained from the EHS focus group discussions (cont.)