

UNIVERSITY OF CALGARY

Improving Emergency Department Efficiency: A Study of Physician Scheduling Strategies to
Reduce Patient Wait Times

by

Negar Ganjoughaghi

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MANAGEMENT

CALGARY, ALBERTA

JUNE, 2024

© Negar Ganjoughaghi 2024

Abstract

Prolonged wait times and overcrowding in emergency departments (EDs) represent significant national challenges in Canada. Within this thesis, I examine three distinct approaches aimed at assisting managers, decision-makers, and schedulers within EDs in addressing these pressing issues. Emergency departments serve as the initial point of contact for patients within the healthcare system, constituting a crucial yet interconnected component of healthcare provision. This study narrows its focus to the aspect of physician scheduling within EDs, recognizing its pivotal role in mitigating wait times and improving efficiency. The initial focus of my investigation lies in optimizing physician schedules within EDs to align with the fluctuating supply and demand dynamics. Extensive literature review and our own dataset reveal the variable productivity levels among physicians in these settings. In this thesis, productivity primarily refers to the number of new patients seen (or treated) by a physician per hour of their shift—a crucial metric in our pursuit of reducing patient wait times. Acknowledging this variability, I first delve into incorporating physician productivity (measured in Patients Per Hour, or PPH rate) into a staffing and shift scheduling problem for physicians. Numerical results show that significant improvements can be obtained in terms of average wait times for patients if we consider the variable productivity of physicians in the staffing and shift scheduling problem. Despite the optimization of schedules, the inherent stochastic nature of ED operations implies occasions where patient volumes exceed expectations, which lead to increased wait times. In response to these fluctuations, in the second and third studies, I propose and assess two distinct strategies aimed at managing ED crowding levels. Particularly, I derive optimal policies for EDs on when and how to extend physicians' shifts or call in physicians in response to a surge in demand. Using the simulation model to evaluate these strategies, I show the effectiveness of having flexibility in physicians' schedules in reducing the average wait time of patients.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr. Marco Bijvank and Dr. Alireza Sabouri, for their invaluable guidance, support, and encouragement throughout the course of my research. Their expertise, patience, and dedication have been instrumental in shaping this work and guiding me through its challenges.

I am also profoundly grateful to my husband, Majid, and my daughter, Elin, for their unwavering love, understanding, and patience during this journey. Their constant support, encouragement, and belief in me have been a constant source of strength and motivation.

Furthermore, I extend my heartfelt appreciation to my parents for their endless love, encouragement, and sacrifices. Their unwavering belief in my abilities and their continuous support have been the cornerstone of my academic endeavors.

I want to thank my Ph.D. committee for their time, encouragement and direction over these past years; FGS, GSA, Haskayne's Ph.D. office, Dr. Osman Alp, Dr. Giovanni J. C. da Silveira, Dr. Nadia Lahrichi, and Dr. Eddy S. Lang.

Contents

Abstract	i
Acknowledgments	ii
Table of Contents	iii
List of Figures	vi
List of Tables	viii
1 Introduction	1
2 Physician Staffing and Shift Scheduling at EDs with Time Varying Productivity	6
2.1 Introduction	6
2.2 Literature Review	9
2.2.1 Staffing and Scheduling of ED Physicians	10
2.2.2 Time Varying Productivity of Employees	14
2.2.3 Workforce Planning with Productivity Considerations	15
2.3 Problem Formulation	16
2.4 Solution methodology	19

2.4.1	Benders Decomposition	20
2.4.2	Enhancements	22
2.4.3	Proposed heuristic for improving upper and lower bounds	22
2.4.4	Symmetry Breaking Constraints	24
2.5	Numerical Experiments	25
2.5.1	Simulation model	29
2.5.2	Simulation Results	31
2.5.3	Sensitivity Analysis	35
2.6	Conclusion	35
2.7	Appendix	44
3	When to extend a shift in Emergency Departments	46
3.1	Introduction	46
3.2	Literature Review	50
3.3	Problem Formulation	53
3.3.1	Decision Epochs	54
3.3.2	State Space	55
3.3.3	Action Sets	57
3.3.4	Transition Probabilities	57
3.3.5	Costs	60
3.3.6	Optimality Equations and Solution Approaches	60
3.3.7	Structural Properties	61
3.4	Numerical Results	64

3.4.1	Simulation	65
3.5	Conclusion	74
4	Enhancing Efficiency and Applicability of Surge Calls in EDs	78
4.1	Introduction	78
4.2	Literature Review	80
4.3	Problem Formulation	82
4.3.1	Decision Epochs	83
4.3.2	State Space	84
4.3.3	Action Sets	85
4.3.4	Transition Probabilities	85
4.3.5	Costs	87
4.3.6	Optimality Equations and Solution Approaches	88
4.4	Numerical Results	88
4.5	Simulation	92
4.6	Conclusion	95
5	Conclusion	100

List of Figures

2.1	Average number of arrivals to FMC ED in a day	26
2.2	FMC vs Stochastic schedules	29
2.3	Comparison of Schedules with 16 shifts in a day in terms of workload distribution	33
2.4	Hourly average waiting time of the Stochastic Schedule vs NV-Stochastic, NV-Deterministic, Deterministic, and FMC Schedules compared to the hourly average arrivals to the ED	34
3.1	The minimum number of patients waiting to be seen for which the optimal policy is to extend a shift for different hours of the day and different values of the previous actions.	66
3.2	Simulation Results: Average waiting time and average number of extensions per day for different values of α for the model PI-2h-15	68
3.3	The average and standard deviation of critical parameters	70
3.4	The minimum number of patients waiting to be seen for which the optimal policy is to extend a shift for different hours of the day and different values of the previous actions for the policies derived from the Policy Iteration algorithm and 4-step, 2-step, and 1-step lookahead models. All policies result in the average wait time of 0.7 hours	73
4.1	PPH distribution for different hours of the shift	90

4.2	Optimal Policy for issuing a surge call based on the number of patients awaiting assessment and the total Patient Per Hour capacity of current physicians. The “Never” region indicates instances where a surge call should not be issued, while the “Always” region indicates when it should be. In the “Maybe” region, the decision to issue a surge call depends on the number of successful surge calls made within the past 8 hours.	91
4.3	90 th percentile Wait time comparison for different surge call policies	96

List of Tables

2.1	FMC Schedule	25
2.2	The scale parameter of the Weibull distribution for the IIAT of patients in the shifts with different total number of patients seen in a shift	27
2.3	Stochastic Schedule	28
2.4	Simulation Results	31
2.5	Sensitivity Analysis	35
2.6	Estimation Results for the effect of various factors on PPH rates. Model 1 includes all variables as discussed in Section 2.5. Models 2-7 gradually remove variables from model 1.	45
3.1	Example Schedule, all shifts are initially 6 hours and might be extended by two hours	54
3.2	The Schedule resulting from the optimization model	64
3.3	The average PPH rate of shifts with lengths of 6 and hours and the average PPH used for the shift that is extended at the beginning of its 5 th hour	69
4.1	Example Schedule	83
4.2	The optimal fixed schedule with 14 shifts, each 7 hours length	89

Chapter 1

Introduction

Everything should be made as simple as possible, but not simpler

Albert Einstein

Emergency Departments (EDs) serve as the initial point of contact for patients within the health-care system, constituting a crucial yet interconnected component of healthcare provision. However, prolonged wait times and overcrowding in EDs represent significant national challenges in Canada. Long wait times can result in increased morbidity, more patients leaving without being seen (LWBS), and increased readmission rates, among other consequences. Due to an increasing population and shortcomings in other parts of the healthcare system, EDs often operate above capacity, leading to burnout among physicians. Fatigue and heavy workloads in physicians can lead to lower quality of care and increased medical errors. Therefore, long wait times in EDs is a medical issue as well as an operations management challenge. This challenging problem has been addressed from different perspectives and using different methods in the literature of operations management and management science.

Some studies attempt to address fluctuations in demand by improving forecasts, redirecting arrivals to other healthcare providers, or delaying arrivals by announcing ED wait times. Other studies focus on resources such as the number of available beds or nurses. Some research aims to predict discharge or hospital admission times to expedite processes. Efficient physician scheduling is crucial for reducing patient wait times, as these times are typically measured from patient triage to initial

physician assessment. Moreover, physicians are among the most expensive resources in the healthcare system, and limited budgets further emphasize the need for optimal physician schedules. Therefore, there are extensive amount of research on physicians availability in EDs. Even in studies focused on ED physicians, various perspectives are employed. Some explore behavioral factors affecting physician productivity; others address shift rostering among physicians; and a group of papers focuses on physician staffing and shift scheduling. This thesis positions itself in the latter group, filling important gaps in the literature.

First, many studies on physician and shift scheduling problems assume deterministic arrivals and productivity. In this thesis, arrivals and physician productivity are assumed to be stochastic, allowing for any distribution. Importantly, based on our data and ED literature, physician productivity, measured by the number of new patients seen per hour, has a decreasing average during a shift. We account for this decreasing PPH rate while addressing the challenging problem of physician staffing and shift scheduling. Additionally, there is no clear policy in the literature for response strategies during ED overcrowding. This thesis fills this gap by providing an easy-to-implement procedure for managing situations when the number of patients waiting to be seen by a physician exceeds a certain threshold. In summary, the focus of our investigation lies in optimizing and improving physician schedules within EDs to align with the fluctuating supply and demand dynamics. This thesis examines three distinct approaches to assist managers, decision-makers, and schedulers in EDs in addressing these critical issues.

The initial approach to addressing prolonged wait times in EDs involves finding an optimal schedule for physicians to match supply with demand. However, the variability in the Patient Per Hour (PPH) rates of physicians, coupled with the stochastic nature of patient arrivals to the ED (i.e., the demand for emergency care), leads to a mismatch between demand and the number of patients served by scheduled physicians (i.e., the available capacity to serve patients) at different times. Ignoring the variable productivity of physicians in the staffing and scheduling problem can result in overstaffing during the initial hours of a shift and understaffing during the latter hours of a shift, hence longer wait times. This discrepancy arises because having a physician in the last hour of their shift does not equate to having one in the first hour due to fatigue and other factors affecting productivity. To mitigate this, it is crucial to account for the variable productivity when staffing decisions are made and shift are scheduled.

In chapter 2, We propose a two-stage stochastic programming formulation for the staffing and shift scheduling problem with the objective of minimizing this demand-supply mismatch. To efficiently solve this problem, we employ Benders decomposition and integrate various enhancements and heuristics. Subsequently, we conduct numerical experiments using data from a Canadian emergency department and evaluate the schedules through simulation modeling. Our findings demonstrate that incorporating the variable PPH yields schedules with lower average wait times compared to schedules that do not account for it.

Despite efforts to optimize schedules, the unpredictable nature of operations in EDs can lead to instances where patient volumes exceed expectations, resulting in longer wait times. EDs must have protocols and strategies in place to effectively manage surges in demand. Current strategies and policies face two main challenges in responding to these surges: determining which strategies to employ and how to implement them effectively. Given that EDs are designed to provide urgent care, they must have the ability to scale resources and adapt as needed during periods of increased demand, ensuring timely service for patients. Therefore, in the third and fourth chapters of this thesis, I study the strategies to deal with the challenging problem of overcrowding in EDs.

To address overcrowding effectively, we must incorporate flexibility into the schedules of physicians such that we can accommodate fluctuations in demand. One straightforward and efficient method is to utilize overtime, i.e., extending workers' shifts beyond their regular hours. In the third chapter of this thesis, the goal is to find an optimal policy on when to extend a shift, depending on the hour of the day and the total PPH capacity from current available physicians. Using a simulation model to evaluate the policy, I show that by finding an optimal policy for extending shifts dynamically versus just adding extra shifts or increasing the length of some shifts statically, it is possible to decrease the average wait time of patients. We formulate the problem as an infinite horizon Markov Decision Process and solve the problem using a policy iteration algorithm. The shift extension policy that we propose incorporates the system status (or ED census) as well as the projected patient arrivals and physician productivity in real time. For our numerical analysis, we use data from an ED in Calgary, Canada. We also solve the problem using look-ahead policies where instead of an infinite horizon we only consider the next few hours. Numerical results show that even though we will have some improvements in the average wait time by using the policies resulting from the look ahead models, the policy from the policy iteration algorithm, significantly outperforms the look-ahead models.

Other strategies are also employed across EDs to address patient crowding levels and to make efficient use of physicians. The most common approaches include on-call physician rotations and the implementation of surge calls. The ED at the Foothills Medical Center in Calgary employs the latter strategy. However, its success has been hampered by specific challenges. The main challenge is the absence of a clear and optimal policy specifying when a surge call should be issued. Additionally, the response rate to these calls is low despite financial incentives designed to encourage physicians' involvement. This results in sub optimal physician participation.

Chapter 4 addresses identified shortcomings by focusing on two primary issues. First, I tackle the absence of a well-defined policy for issuing surge calls by formulating the problem as a MDP. The objective of the model is to minimize the weighted average of the number of patients waiting and the cost incurred by issuing surge calls. The model incorporates the likelihood that physicians may not respond to a surge call, a parameter intricately linked to the incentives and consequences associated with their response or non-response to surge calls. Second, by manipulating this probability within our MDP framework, my research offers decision-makers a tool to discern the most effective policy given the available incentives and constraints. We evaluate the effectiveness of our policy using a simulation model. Results show that the flexibility that comes from issuing surge calls, rather than solely adding fixed shifts to the schedule, can improve the average and 90th percentile of the wait time by almost 13 and 40 minutes, respectively. The improvement in wait time is even more pronounced when resources become more scarce and fewer fixed shifts are scheduled.

To conclude, the main goal of this thesis is to address the challenging and common problem of long wait times in EDs. We focus on the availability of physicians and find optimal schedules and policies that result in a better match between the demand for timely health care services and the capacity of the ED to provide it. The promising numerical results of this thesis show that with the proposed schedule for physicians, it is possible to decrease the average wait time in EDs without adding any extra shifts for physicians. Moreover, this thesis provides valuable and easy-to-implement policies for how and when to respond to surges in demand. To the best of our knowledge, this is the first study to find such policies and formulate the staffing and shift scheduling problem by considering all sources of variability and stochasticity in demand and the productivity of physicians.

This thesis is organized as follows: Chapter 2 explains and presents the physician staffing and

shift scheduling problem. Chapter 3 introduces the shift extension idea. Chapter 4 discusses strategies to improve the implementation of surge calls in EDs. Finally, Chapter 5 reviews and compares the findings from all chapters, discussing the limitations of the study and suggesting future extensions and areas for further research.

Chapter 2

Physician Staffing and Shift Scheduling at EDs with Time Varying Productivity

2.1 Introduction

EDs are among the most important locations for providing health care services to the general public. However, due to an aging population, among other reasons, the demand for emergency care has been increasing over the last decades. This results in elevated patient crowding levels and high workloads for health professionals. From the patient perspective, ED crowding results in extended patient wait times. Based on the Canadian Institute for Health Information report in 2023 ¹, one out of ten Canadian patients reported waiting five or more hours the last time they went to an ED to obtain care. Prolonged waiting times are associated with increased morbidity, more patients leaving without being seen (LWBS), and increased readmission rates among other consequences (Sun et al., 2013; Batt and Terwiesch, 2015; Richardson and Bryant, 2004).

Since most patients in an ED are walk-in patients, and there is a high uncertainty regarding the arrival of new patients as well as the care that these patients require, it is very challenging to schedule a sufficient number of physicians. When a large number of physicians are scheduled to work, it can balance any high demand for emergency care. Even though this would be ideal for patients, it is very

¹Canadian Institute for Health Information. Emergency Department Wait Time for Physician Initial Assessment (90% Spent Less, in Hours). Accessed April 26, 2024.

costly when physicians are idle. Alternatively, the shifts of physicians can be scheduled more effectively to balance the (uncertain) patient arrivals. A physician's shift schedule comprises of the number of shifts to be scheduled in a day as well as their individual start times and lengths. It determines the number of physicians working at the ED for each hour of the day. ED managers face a variety of trade-offs between system costs and patient needs. For instance, while an ED with fewer shifts in a day can be desirable from a cost perspective, this can result in poor quality of care for patients and have unsatisfactory patient outcomes (such as longer wait times and the consequences discussed earlier). Such trade offs between service quality and operational efficiency need to be carefully considered when determining staffing levels and creating physician schedules. Usually it is not feasible to increase ED treatment capacity due to budget constraints, and therefore, it is crucial to make the best use of the existing resources.

This is exactly one of the most challenging issues for ED managers: when are physicians needed the most: Should they be scheduled when demand is (expected to be) highest? Or should they be scheduled before the demand peaks to prevent a buildup of patients waiting to receive care? This could provide helpful information to arrange their shift schedules. Especially since demand is uncertain, it is not straightforward to develop an efficient schedule. To make things worse, other phenomena exist in ED environments that make the physician staffing and shift scheduling problem at EDs ever more challenging.

Another source of uncertainty in healthcare systems is the patients' characteristics and their needs. In EDs, patients are heterogeneous in terms of their main complaints and their required treatment plans. As a result, the time it requires for a physician to spend on each individual patient (i.e., the treatment or service time) will vary. Furthermore, the number of new patients a physician will see in an hour is another source of complexity. This measure is known as patient-per-hour (PPH) rate in the literature, and it has been shown empirically that physicians accept fewer new patients near the end of their shift (Chan, 2018). In explaining this decreasing pattern in PPH rate, it should be mentioned that this physician productivity metric only accounts for the number of new patients the physician signs up for and is not a representative of how much work the physician performs in each hour (since patients typically stay in the ED for multiple hours). This decrease can be attributed to two main reasons: First, in the first hours of the shift, there are less patients under the care of the physician such that the physician has more available time to sign up for more (new) patients. In

contrast, during the later hours of the shift, the physician would be busy with follow-up visits and treatment procedures on the patients, who have been initially assessed in an earlier hour of the shift and are still under the care of the physician. Consequently, the physician would have less time to start the treatment of new patients. Second, if patients are still under the care of the physician at the end of their shift, their care needs to be transferred to another physician. This passing of work from one individual to another individual on the following shift is called hand-offs. It has been shown in the literature that these hand-offs lower clinical quality and patients who have been handed off are more likely to revisit the ED within three days (Batt et al., 2019). In order to minimize hand-offs of patients to the next physician, physicians are advised not to leave the ED after their shift ends unless the medical path for each of their patients is finalized. Since the treatment time of patients is usually a lengthy process, physicians tend to sign up for fewer new patients in the last hours of their shift.

The goal in this chapter is to include these characteristics in the staffing and shift scheduling decision for ED physicians. It is important to not only include the stochastic nature of the patient arrivals and their treatment time, but also to consider this time-varying PPH rate during the physician's shift when matching the available capacity at the ED in the schedule with the patient needs (which is a multiplication of their arrivals and their treatment times). Ignoring this variability would result in overstaffing during the earlier hours of the shift and understaffing in the last hours of the shift. The main contribution of this paper to the existing literature of workforce staffing and shift scheduling is to develop a general framework that allows for stochastic and time-varying productivities of servers where the need for these servers also follows a time-dependent arrival pattern.

Considering the mentioned sources of stochasticity, we formulate the staffing and scheduling problem as a two-stage stochastic problem where the objective is to minimize the number of patients waiting to be seen during a planning horizon. The output of the model is a detailed schedule plan that determines the number of shifts to schedule in a day as well as their start and end times (i.e., the shift lengths). We solve the problem using the Benders Decomposition algorithm. In the first stage, we find a schedule before knowing the number of arrivals and productivity of physicians in each hour of the planning horizon. Then, in the second stage, based on the output schedule from the first stage problem, we calculate the number of patients waiting to be seen at the beginning of each hour of the planning horizon for multiple scenarios. The values of the dual variables that we get from solving the second-stage problem are used to add optimality cuts to the first-stage problem and then we go back

to solve this problem again with the additional constraint. This iterative procedure between the two stages is repeated until optimality conditions are satisfied. In order to speed up the process, we add enhancements and a heuristic to the algorithm, which are explained in detail in Section 2.4.

Our study is motivated by a real-world setting at the Foothills Medical Center (FMC) in Calgary, Alberta (Canada). Using data from this ED, we solve the problem and evaluate and compare the resulting schedules with the existing schedule at FMC by employing a discrete event simulation model. Based on the numerical experiments and simulation results, the stochastic algorithm can find a schedule that effectively reduces the average patient waiting times compared to the schedule generated by the deterministic problem as well as to the schedule used in the ED of FMC. In the deterministic problem, we use the hourly average values for the number of arrivals and PPH of physicians in each hour of the shift.

The rest of the paper is organized as follows. Section 2.2 briefly reviews the related literature. Section 2.3 presents the formulation of the problem, followed by the solution approach in Section 2.4. Section 2.5 describes the numerical results and Section 2.6 presents the concluding remarks.

2.2 Literature Review

There is a large body of literature addressing the staffing and shift scheduling problem, which is reviewed by Defraeye and Van Nieuwenhuysse (2016) and Van den Bergh et al. (2013). According to these literature reviews, in most studies, the productivity of servers or PPH rate of physicians is considered to be deterministic and constant during a shift. By including time-varying and stochastic productivity levels for employees in the staffing and shift scheduling problem, there are three streams of literature relevant to our study. In Section 2.2.1, we present the staffing and shift scheduling problems that have been studied for physicians in EDs. The notion of employee productivity varying over the duration of a shift is addressed in Section 2.2.2. Finally, incorporating time-varying employee productivity in workforce planning is discussed in Section 2.2.3.

2.2.1 Staffing and Scheduling of ED Physicians

To avoid ambiguity with regards to terminology, we begin with defining the steps of workforce planning. The process of personnel capacity planning has four main steps (Atlason et al., 2008): First, in the forecasting step, the demand or the customer arrival rate is estimated; Second, in the staffing step, the forecasted demand is used to determine the optimal number of servers required to meet this demand; Third, in the scheduling step, a detailed scheduling plan, including the number of shifts as well as their start time and length, is created based on the required staffing levels; Finally, in the rostering step, shifts are assigned to individual servers. In the literature, the terms staffing, scheduling and rostering have been used inconsistently. What is called scheduling in many papers, is actually the rostering step, and what we call the scheduling step, is called the staffing step in some other papers. To be clear, we follow the definition of Atlason et al. (2008) in this paper and the focus of our work is on the staffing and scheduling steps and not the rostering step.

Erhard et al. (2018) present an overview of the literature on shift scheduling problems for physicians in the last 30 years (not restricted to an ED context). When patients have medical appointments to see a physician (e.g., in outpatient clinics), the scheduling of physicians is considered a *task-coverage problem* (or skill-coverage problem). In such problems, the actual tasks (or skills) that need to be executed are already known before the employee starts the shift (sometimes even before the employee is scheduled). In contrast, patient arrivals and their associated required care needs are less predictable in EDs. Consequently, the scheduling of emergency physicians is considered a *workload-coverage problem*, where the demand for employees needs to be forecasted based on expected workloads and employees are scheduled to cover these predicted personnel demands.

There are two main research methodologies to tackle the staffing and scheduling problem of physicians in EDs. One is simulation and the other is stochastic models based on queueing theory. The reason that simulation is a common research method is because EDs are highly complex systems and it is difficult to use analytical models that accurately describe the dynamics in EDs.

Badri and Hollingsworth (1993) are one of the first authors to use a *simulation model* to analyze the effect that various changes in the number of resources (physicians, nurses, pharmacists and beds) have on the average wait time, LOS and LWBS percentage. In Rossetti et al. (1999), the impact of only four different staffing levels on pre-existing 8-hour shifts are simulated. Al-Najjar and Ali (2011)

study the existing staffing levels of physicians and nurses at two EDs. In both locations there were three consecutive, non-overlapping shifts of 8 hours each. The number of physicians and nurses for two consecutive, non-overlapping shifts is determined by Ghanes et al. (2015), such that the LOS is minimized when there is a budget constraint. Ahmed and Alkhamis (2009) developed a simulation-based heuristic to determine the optimal number of ED staff members (i.e., doctors, lab technicians and nurses) required to maximize the patient throughput and to reduce the patient waiting time subject to budget and average wait time constraints. Staggered shifts are investigated by Kuo (2014) to better match physicians with patient demand. The author proposes the integration of simulation in a simulated annealing algorithm to create a physician schedule with 8-hour shifts.

Except for Kuo (2014), simulation techniques are mostly used to assess the impact of different staffing levels for predetermined shift schedules on wait times. A number of papers address the ED staffing problem by a two-step approach. In the first step, a model is constructed to determine minimal staffing levels for each period. In a second step, a model is developed to create actual shifts that cover these staffing requirements of the first step. Centeno et al. (2003) use a simulation model to first set the required nurse levels such that a patient demand service level is satisfied in each period, and these staffing levels are then used as input to an integer linear programming (ILP) model that assigns nurses to five different 12-hour shifts. A similar approach is proposed by Sinreich and Jabali (2007), who combine the two steps of a simulation model coupled with an ILP model in an iterative manner to produce shift schedules for doctors, nurses, and imaging technicians. In contrast to Centeno et al. (2003), different shift lengths are considered.

Besides simulation, *queuing theory* is also extensively used to determine staffing levels because it allows the evaluation of quality-of-service criteria such as the average waiting time for patients or the probability that the wait time is within a certain threshold value (e.g, no more than 20% of the patients wait more than one hour before being seen by a healthcare provider). However, no analytical formulas exist to determine optimal staffing levels in complex systems where the demand process is stochastic and varying over time (which is the case in EDs). Often the steady-state expressions for an $M(t)/M/s(t)$ queueing model are used, one for each planning period, which is referred to as the “stationary independent period-by-period” (SIPP) approach (Green et al., 2001). Each of these period-specific models is independently solved for the minimum number of servers needed to meet the service target in that period. Once staffing levels are determined, a set of permissible shifts is selected

based on an integer programming (IP) model formulation that minimizes labor cost and satisfies the minimum number of employees in each time period. This is often referred to as the SIPP-IP approach (Ingolfsson et al., 2002).

The advantage of a stationary approximation is its simplicity. The disadvantages is that it can only be applied if the stationary behavior is obtained within each time period. However, EDs are commonly overloaded and many patients have to wait beyond one period (frequently even beyond two or three periods). A lagged SIPP queuing model is proposed by Green et al. (2006), who use it to analyze different physician staffing patterns in EDs that reduce the number of LWBS patients. Alternative to the lagged SIPP approach, Stolletz (2008) approximates the queuing system with an Erlang-loss system where blocked demand is carried over to the next period (referred to as the stationary backlog-carryover approximation). Jennings et al. (1996) and Green et al. (2007) present a normal approximation for the $M(t)/M/s(t)$ model by using an infinite-server model to determine minimal staffing levels with the square-root staffing formula for different application settings including EDs. A numerical comparison of different approximation approaches that account for time-varying queueing effects is presented by Ingolfsson et al. (2007). More recent overviews on other approximation methodologies are provided by Schwarz et al. (2016) and Defraeye and Van Nieuwenhuysse (2016). In an ED setting, Zeltyn et al. (2011) calculate time-varying staffing levels with the classical square-root safety-staffing rule from queueing theory, where the offered-load is estimated with simulation.

In all papers mentioned so far, only staffing levels are determined and no shift schedules are considered. In contrast, Ingolfsson et al. (2002) extend the SIPP approach by including shift structures, and a genetic algorithm is used to search for schedules that satisfy all service level constraints. Izady and Worthington (2012) use staffing and scheduling routines in two consecutive steps. First, they extend the square-root staffing method of Jennings et al. (1996) for a network of ED resources. Next, the IP model of Sinreich and Jabali (2007) is modified to produce shifts for each resource type (i.e., shift start times, durations, and number of physicians assigned) that comply with the hourly staffing levels of the first stage. More traditional queueing theory is used by Liu and Xie (2018), who use $M(t)/M/s(t)$ approximations for the average total waiting time of patients in an ED by separating the patient arrivals to those who obtain medical service within the period of arrival and those who carry over to another period. This results in a non-linear objective function that the authors want to minimize. After linearization, the resulting mixed-integer programming (MIP) formulation can only

be solved for problems of a small size. Therefore, the authors propose a variable neighborhood search (VNS) to solve the staffing problem.

Another issue when applying queueing models is that assumptions have to be imposed on the distributions for the patient arrival and service processes, such as Poisson arrivals and exponentially distributed service times, in order to use closed-form expressions for some service quality measures of the ED system (e.g., patient waiting time or fraction of patients who become LWBS). However, these assumptions do not always comply with reality.

Not many papers in the literature directly construct shift schedules without the two-step approach of determining staffing levels first. Savage et al. (2015) propose a simple MIP model to generate an ED physician schedule for an acute care area (patients with CTAS level 1, 2 or 3) and a fast-track clinic (patients with CTAS level 4 or 5) based on historical patient arrival data. Such a formulation doesn't incorporate the stochastic nature of the patient arrivals or physician productivities observed in EDs. A chance-constrained programming model to schedule medical personnel with different skill levels in an emergency department is used by Ganguly et al. (2014). The probability that the scheduled workload capacity of qualified physicians is sufficient to satisfy the work content requested by arriving patients has to exceed a target service level in each period of the planning horizon. The objective is to minimize staffing cost. The authors argue that staffing and scheduling decisions have to be made jointly. Interestingly, both papers do not incorporate how to handle unmet demand for emergency care in any given time period: either it is prevented as much as possible in the objective function (Savage et al., 2015) or it is restricted to rarely occur in service level constraints (Ganguly et al., 2014). An alternative approach to jointly solve staffing and scheduling problems is with stochastic programming. This methodology has only been applied to call center settings. Robbins and Harrison (2010), Gans et al. (2015) and Bodur and Luedtke (2017) make staffing and shift decisions in the first stage, and the actual service level is calculated in the second stage after the call volume is realized. Gans et al. (2015) also provide a stochastic programming model where workers can be added and removed from an initial schedule as recourse variables in the second stage, similar to Kim and Mehrotra (2015).

2.2.2 Time Varying Productivity of Employees

Shift work is very common in hospital care. There have been a number of studies that analyze the impact of the duration of a shift on the productivity of emergency physicians, where productivity is defined as average PPH rate during the shift. Hart and Drall (2007) and Jeanmonod et al. (2008a) conclude that shifts with a duration of 8 to 9 hours result in a higher productivity compared to 12-hour shifts. In contrast, the study of Yang and Lee (2012) suggests no difference in average productivity. Chan (2018) concludes that physicians accept fewer patients in the last two to four hours prior to the end of a 9-hour shift, which is referred to as the end-of-shift phenomenon. Physicians do this to avoid handoffs. In that same study, the author finds that the physician's productivity decreases even earlier when there is more shift overlap with other physicians at the end of their shift.

Other than the end-of-shift effect, a worker's service rate can also be impacted by environmental parameters such as workload (KC and Terwiesch, 2009), interactions with other workers (Mass and Moretti, 2009; Saghafian et al., 2021) and system design (Shunko et al., 2017). Especially the impact of workload on the service time has received quite some attention in the literature. See the review by Delasay et al. (2019). The workload is usually defined as the extent to which a system is busy or congested at a given time as reflected by the number of customers in the system. We only discuss the existing studies that examine the effect of workload on service times in hospital settings, where service time is measured as the time from when a patient is placed in a treatment bed to when treatment is completed as either the patient being discharged to go home or an inpatient bed is requested to admit the patient in the hospital. KC and Terwiesch (2009) and Batt and Terwiesch (2012) show that service times for a given set of patient care tasks decrease when the load on the system increases. In addition, Kuntz et al. (2014) and KC and Terwiesch (2012) illustrate that patients are discharged earlier with an increase in workload. Other than these workload-dependent speedup effects, there is also evidence of slowdown when there is excess load over a longer period of time (KC and Terwiesch, 2009; Armony et al., 2015). This is referred to as overwork (Delasay et al., 2016) or the saturation effect (Berry Jaeker and Tucker, 2017).

More recent papers observe that patient service times can be expressed as an inverted U-shaped function of the hospital workload (Berry Jaeker and Tucker, 2017; Batt and Terwiesch, 2017). This means that emergency physicians initially slow down when the ED becomes more congested, but at

some point their service time speeds up again. Armony et al. (2015) find opposite speeding-up and slowing-down effects. Delasay et al. (2016) provide a theoretical framework to model the effect of workload on service times.

Even though the impact of ED workload on service times is studied in the literature, this doesn't imply that the same (complex) relationship exists between ED workload and the PPH rate of physicians. For instance, physicians can sign up more patients when the ED gets crowded (since it might take longer before diagnostic tests are performed). This is called multitasking, and the effect of this on ED physician performance (measured as throughput rate and outcome quality) is studied by Kc (2014). The author concludes that multitasking is only beneficial up to a certain threshold level. The relationship between ED workload and multitasking is not taken into consideration. In contrast, Batt and Terwiesch (2017) observe that the number of diagnostic tests ordered by physicians is not impacted by an increased ED workload. The authors also study the impact of early task initiation and conclude that this reduces service times, but this reduction does not increase with ED workload. The only two studies that investigate the impact of ED workload on productivity are by Jeanmonod et al. (2008b) and Zaerpour et al. (2022). In the former, the authors observe no strong relationship between patient volume or shift time of day and the productivity (or PPH rate) of residents in the ED. In the more recent study by Zaerpour et al. (2022), the time varying PPH rate is considered in the ED physician rostering problem. Using an empirical model, the authors show that the shift hour and the individual physician are the main factors affecting the PPH rate of the emergency physician and not the workload.

2.2.3 Workforce Planning with Productivity Considerations

One of the earliest works where productivity is considered in personnel scheduling is by Li et al. (1991), who study a shift scheduling problem where a relative productivity is included to distinguish between full-time and part-time employees. A similar approach is used in multi-skill workforce planning, where employees can be cross-trained. Brusco and Johns (1998) and Brusco and Jacobs (1998) study a two-skill and multi-skill staffing problem, respectively, where employees have a lower productivity in secondary skills.

A productivity parameter for individual employees is hardly included in rostering problems

where workers are assigned to shifts. Goodale and Thompson (2004) take individual productivity levels from a normal distribution and the labor cost of each employee is approximately proportional to their productivity levels. Employees are grouped by their productivity level in Thompson and Goodale (2006). The productivity level in Akbari et al. (2013) is exogenous and dependent on the shift they are assigned. Furthermore, when employees are scheduled to work consecutive shifts on the same day, their productivity decreases by a constant factor. In these papers, the authors formulate a deterministic optimization model and use heuristic procedures to assign workers to shifts. As mentioned before, Campbell et al. (2011) and Gnanlet and Gilland (2014) are the exception as they use stochastic programming.

Worker fatigue creates a reduction in productivity that is increased again after a break period. Such non-constant service times are included in staffing and shift scheduling problems by Bechtold and Brusco (1994). This work is extended by Goodale and Tunc (1998), who also take learning effects into account to explain changes in productivity over time. To the best of our knowledge, the time in-homogeneous productivity of servers, or PPH of physicians, has never been considered in a staffing and shift scheduling problem, which is a gap in the literature that this paper tries to fill.

2.3 Problem Formulation

In this study, we propose an Integrated Staffing and Scheduling (ISS) formulation to determine a daily shift schedule that minimizes the sum of the number of patients waiting to be seen by a physician in the ED at the beginning of each hour of the planning horizon. The number of hourly arrivals to the ED and the number of new patients that a physician can visit in an hour, or PPH, are considered to be stochastic. Let $\Lambda_{h,d}$ denote the number of arrivals to the ED at hour h of day d of the planning horizon and $\Psi_{i,j,k,d}$ denotes the number of new patients seen by the physician working at hour i of shift j with length k on day d of the planning horizon. We assume that the average number of patients a physician can see in an hour (Patient Per Hour) is variable during a shift.

To simplify the problem formulation, we assume without loss of generality that the shift schedule will be the same for each day in the planning horizon. However, our formulation and solution approach can be easily extended to include different shift schedules per day to accommodate different

patient arrival patterns between days of the week. Based on our assumption, the output of our problem formulation is a one-day shift schedule that will be the same on all days of the planning horizon T . In the problem formulation, we include parameter J to present the maximum number of shifts in a day and we also have a limit on the number of hours to be scheduled in a day, which is denoted by F . This accommodates the scheduling of shifts with different lengths. For example, when $J = 10$ and $F = 70$, this allows for 10 shifts of 7 hours each, or 5 shifts with a length of 6 hours and another 5 shifts but with a length of 8 hours. By limiting the total number of hours that are scheduled in a day (i.e., by using parameter F) in our problem formulation, we do not have to consider the cost of scheduling physicians. Since we will formulate the objective of our shift scheduling problem such that the total number of patients waiting to be seen at the beginning of each hour in the planning horizon is minimized, the shift schedule will always use all F hours in a day. In other words, we keep the resources fixed and try to optimally allocate them throughout the day. We also allow shifts to have different lengths, ranging from K_{min} to K_{max} .

We formulate the ISS problem as a two-stage stochastic integer problem. In the first stage, before observing the uncertain patient arrivals and physician productivities, we decide on the start time and length of each shift. The binary decision variables are denoted by $x_{j,h,k}$ for $j = 1, \dots, J, h = 0, \dots, 23$, and $k = K_{min}, \dots, K_{max}$, and they represent the schedule of the system, vector X . More specifically, $x_{j,h,k} = 1$ if shift j with length k starts at hour h of the day. The ISS problem formulation follows as:

$$\min_X \quad \theta = E_{\Lambda, \Psi}[Q(\Lambda, \Psi, X)] \quad (2.1)$$

$$s.t. \quad \sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} x_{j,h,k} \leq 1 \quad j \in \{1, \dots, J\} \quad (2.2)$$

$$\sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} \sum_{j=1}^J kx_{j,h,k} \leq F \quad (2.3)$$

$$x_{j,h,k} \in \{0, 1\} \quad j \in \{1, \dots, J\}, h \in \{0, \dots, 23\}, k \in \{K_{min}, \dots, K_{max}\} \quad (2.4)$$

The objective function in Eq. (2.1) minimizes the expected number of patients waiting to be seen at the beginning of each hour over the planning horizon T . Constraints 2.2 enforce that each shift is scheduled at most once in a day while Constraint 2.3 limits the total number of hours that are covered by the schedule to at most F . Finally, Constraints 2.4 require that the variables x are binary variables.

The function $Q(A, P, X)$ calculates the total number of patients waiting to be seen at the

beginning of each hour over the entire planning horizon for a given schedule X , where the matrices A and P present realizations of the random variables $\Lambda_{h,d}$ and $\Psi_{i,j,k,d}$ (i.e., representing the actual arrivals and physician productivities, respectively). Function $Q(A, P, X)$ is defined as:

$$Q(A, P, X) = \sum_{t \in T} M_t \quad (2.5)$$

$$\text{where} \quad M_t = \max[0, M_{t-1} - A_{t-1} + \sum_{k=K_{min}}^{K_{max}} \sum_{j=1}^J \sum_{i=1}^k x_{jh'k} P_{ijkd}] \quad (2.6)$$

In the second stage model, the function $Q(A, P, X)$ finds the total number of patients who are waiting to be seen at the beginning of hour t , denoted by M_t , during the planning horizon T . In this formulation, any unmet demand at the end of the hour t will be transferred to the next hour. The variable M_t is defined as the maximum of zero and the total number of patients waiting to be seen at the beginning of hour $t - 1$ (i.e., M_{t-1}) plus the number of arrivals during hour $t - 1$ (i.e., A_{t-1}) minus the total number of patients who were actually seen by a physician during hour $t - 1$. This function is not linear and is linearized in the extended form by using additional constraints that are discussed later.

The output of the first stage problem is a daily schedule while the planning horizon of the second stage problem is more than a day. This is where we use our assumption that the shift schedule is the same for all days of the week (i.e., that the arrival pattern is almost the same on all days of the week in our study ED). Note that in the second stage problem, we use $x_{jh'k}$ in which hour h' will be the result of $t - (i - 1)$ modulo 24. For example, to subtract the productivity of a physician who is working the fourth hour in its seven-hour shift at hour $t = 35$ of the planning horizon, then we need to know that the shift has started at hour $h' = 8$ of the day (since $i = 4$, $k = 7$ and $t = 35$ corresponds to the 11th hour the day). Moreover, to find the day d in P_{ijkd} that corresponds to hour t of the planning horizon, we simply calculate the floor of t divided by 24, since the planning horizon starts with day 0 at $t = 0$.

Furthermore, note that the second stage problem is a trivial problem and we only calculate the number of patients waiting at the beginning of each hour during the planning horizon. It is just an iterative procedure to calculate M_t when the arrivals A_t and productivities P_t are given for a schedule X . In other words, there are no decisions to make in the second stage problem, and it is not an actual optimization problem. However, since we need the dual variables associated with the constraints in

the second stage problem, and the second stage problem is an easy problem to solve, we formulate it as an optimization problem.

The ISS problem is a 2-stage stochastic integer program with binary variables in the first stage and continuous variables in the second stage. Since in the second stage we only calculate the number of patients waiting at the beginning of each hour of the planning horizon, it is always feasible, bounded and has complete recourse. Moreover, since the second stage problem is a linear problem, we can conclude that the function $Q(\cdot)$ is convex and piece-wise linear (Birge, 1997).

2.4 Solution methodology

The objective function of the ISS problem is the expected value of a function over the random matrices $\Lambda_{h,d}$ and $\Psi_{i,j,k,d}$. Real-life situations involve a large number of scenarios, which makes it difficult to solve the problem. To overcome this challenge, the Sample Average Approximation (SAA) algorithm is used. The main advantage of using the SAA algorithm is its ability to find near-optimal solutions while considering samples comprising a smaller number of scenarios. The idea consists of generating a random sample of the parameters that are uncertain and integrate them into the model. In particular, assuming N scenarios of ξ^1, \dots, ξ^N , the SAA approach approximates the following model which is considered a stochastic programming problem with the finite set $\{\xi^1, \dots, \xi^N\}$ of scenarios each with equal probability of N^{-1} (Shapiro, 2003). The expected value is replaced by the sample average over these scenarios: $E_{\Lambda, \Psi}[Q(\Lambda, \Psi, X)] \approx \frac{1}{N} \sum_{i=1}^N Q(A(\xi^i), P(\xi^i), x)$.

With fixed scenarios, we obtain the extensive form or the deterministic equivalent of the SAA problem, presented below:

$$\begin{aligned}
 \text{Min} \quad & \frac{1}{N} \sum_{i=1}^N \sum_{t \in T} M_t^{\xi^i} \\
 \text{s.t.} \quad & \sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} x_{jhk} \leq 1 \quad j \in \{1, \dots, J\} \\
 & \sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} \sum_{j=1}^J kx_{jhk} \leq F \\
 & M_t^{\xi^i} \geq M_{t-1}^{\xi^i} + A_{t-1}^{\xi^i} - \sum_{k=K_{min}}^{K_{max}} \sum_{j=1}^J \sum_{i=1}^k x_{jh'k} P_{ijkd}^{\xi^i} \quad t \in \{1, \dots, T-1\}, i \in \{1, \dots, N\}
 \end{aligned}$$

$$\begin{aligned}
 M_t^{\xi^i} &\geq 0 & t \in \{1, \dots, T-1\}, i \in \{1, \dots, N\} \\
 M_0^{\xi^i} &= 0 & i \in \{1, \dots, N\} \\
 x_{jhk} &\in \{0, 1\} & j \in \{1, \dots, J\}, h \in \{0, \dots, 23\}, k \in \{K_{min}, \dots, K_{max}\}
 \end{aligned}$$

However, as the number of scenarios increases, this formulation (extensive form) becomes too large and therefore we use Benders Decomposition (BD) strategies to solve it.

2.4.1 Benders Decomposition

The classic Benders Decomposition algorithm (or BD in short) was initially introduced in early 1960's for mixed integer linear programming problems (Benders, 1962). The idea is to divide the problem into a Master Problem (MP) which includes the integer variables and a Sub Problem (SP) where the values of the integer variables are given as input. This results in a continuous linear problem for which standard duality theory can be used to develop optimality cuts that improve the approximation in the objective of the MP. BD iterates between solving the MP, passing the solution to SP, solving the SP, deriving optimality cuts, and passing these cuts to the MP. The MP involves a lower approximation of the function Q through the variable θ and the optimality cuts. The SP has a linear programming formulation, and since it always finds a solution that is feasible for the MP, it yields an upper bound for the MP. In any iteration, if the difference between the best upper bound and lower bound (i.e., the optimality gap) is less than a specified ϵ , then the procedure terminates. Otherwise, the dual information from the SP is used to add an optimality cut to the MP and we continue with the next iteration.

We decompose the ISS problem into an MP where the scheduling decisions are made and a set of SPs, one for each scenario. Each SP is formulated as below and we solve each independent SP for one fixed vector of values for A and P .

$$Q(A, P, X) = \min \sum_{t \in T} M_t \quad (2.7)$$

$$s.t. \quad M_t \geq A_{t-1} + M_{t-1} - \sum_{k=K_{min}}^{K_{max}} \sum_{j=1}^J \sum_{i=1}^k x_{jh'k} P_{ijkd} \quad t \in \{1, \dots, T-1\} \quad (2.8)$$

$$M_t \geq 0 \quad t \in \{0, \dots, T-1\} \quad (2.9)$$

$$M_0 = 0 \quad (2.10)$$

The Benders MP is:

$$\min_X \quad \theta \quad (2.11)$$

$$s.t. \quad \sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} x_{jhk} \leq 1 \quad j \in \{1, \dots, J\} \quad (2.12)$$

$$\sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} \sum_{j=1}^J kx_{jhk} \leq F \quad (2.13)$$

$$E^l X + \theta \geq e^l \quad l = 1, \dots, v - 1 \quad (2.14)$$

$$x_{jhk} \in \{0, 1\} \quad j \in \{1, \dots, J\}, h \in \{0, \dots, 23\}, k \in \{K_{min}, \dots, K_{max}\} \quad (2.15)$$

After solving the MP in iteration v , the optimal solution of X^v will be transmitted to the SPs and θ^v will be saved as the new lower bound. Since we add a new optimality cut to the MP in each iteration, this lower bound will be non-decreasing in the minimization problem. Constraints 2.14 in the MP are the optimality cuts included from the previous iterations. To derive Benders optimality cuts in each iteration, we solve the SP and use the dual variables $\pi_{\xi_i}^t$ associated with Constraints 2.8 for each scenario. We then can define the optimality cut for each iteration l , $l = 1, \dots, v$, by E^l and e^l as follows: $e^l = \frac{1}{N} \sum_{i=1}^N (\sum_{t \in T} \pi_{\xi_i}^t A_{t-1}^{\xi_i})$ and $E^l X = \frac{1}{N} \sum_{i=1}^N (\sum_{t \in T} \sum_{j=1}^J \sum_{k=K_{min}}^{K_{max}} \sum_{i=0}^k \pi_{\xi_i}^t P_{i,j,k,d}^{\xi_i} x_{jh'k})$

After solving the SP in iteration v , we calculate $w^v = e^l - E^l X^v$ for each $l = 1, \dots, v$. Then we find the maximum w^v and that will be the objective of the subproblem for that iteration and provides us an upper bound for MP objective. If w^v is lower than the previous iterations, we update the upper bound and compare it to the latest lower bound or θ^v . The reason is that, the SP is in fact the original problem except that the value of a subset of variables, binary decision variables in our case, are fixed. Therefore, the SP gives a solution with an objective value that is not better than the optimal value of the original problem, or in other words, it gives upper bounds on the original problem. However, there is no reason for these upper bounds to change strictly and we can only find the best upper bound found so far and update it at each iteration. Then these new upper bounds change decreasingly and that determines the convergence of the algorithm, not the particular objective value of the SP at each iteration.

2.4.2 Enhancements

To solve the problem using a BD strategy, we need to solve the MP (which is an integer problem) in each iteration, which will become more difficult to solve after each time that a new cut is added. Besides, while the BD algorithm is a finite scheme, the number of iterations that is required may be too large in practice (Santoso et al., 2005). Therefore, in order to improve the convergence behaviour of the generic BD algorithm outlined above, we use a number of enhancements to be able to solve the problem in a reasonable time to optimality. First, we use a multicut BD algorithm, in which we add one cut per scenario in each iteration, which results in faster convergence of the upper bound and the lower bound. Second, we add knapsack inequalities to the master problem. The knapsack inequalities at iteration l are: $\lfloor E^l \rfloor X \geq \lfloor e^l - UB \rfloor$, where UB is the best upper bound found so far and $\lfloor a \rfloor$ is the component-wise floor of a (Santoso et al., 2005). It has been shown that adding the above knapsack inequalities in addition to the optimality cuts can have a significant impact on improving the lower bound of the BD algorithm and generating a better MP solution (Santoso et al., 2005). Third, as proposed by Bodur and Luedtke (2017), we include a set of constraints in the master problem based on the average value of the sampled values of the stochastic parameters, which will result in the improvement of the convergence of the BD algorithm. Particularly, this technique is motivated by Jensen's inequality: $E_{\Lambda, \Psi}[Q(\Lambda, \Psi, X)] \geq Q(E_{\Lambda}[\Lambda], E_{\Psi}[\Psi], X)$. For more detailed information about this initialization of the MP technique, we refer to Bodur and Luedtke (2017). Lastly, we apply a new heuristic strategy to the BD algorithm which is described comprehensively in the next section. We also add symmetry breaking constraints to the MP to reduce the size of the feasible region. As a result, we are able to solve the MP faster. These symmetry breaking constraints are presented in Section 2.4.4.

2.4.3 Proposed heuristic for improving upper and lower bounds

The solution of the SP and the UB achieved in each iteration of the BD algorithm corresponds to the current solution of the MP. However, since this solution may be quite far from the optimal solution, the resulting UB might not improve as fast as desired. If a good solution can be found through some heuristic, then the current solution of the MP can be replaced by this better solution, resulting in a lower upper bound (Santoso et al., 2005). The aim of using a heuristic is to find better upper bounds as well as multiple optimality cuts at each iteration of the BD algorithm. By improving the lower

bound (from the MP objective function) and upper bound (from the best solution to the SP found so far), we can reduce the number of iterations and consequently the number of MPs that need to be solved to reach optimality. While hybridizing the BD algorithm with heuristics or meta heuristics can simultaneously improve the lower and upper bounds (Rei et al., 2009), effective strategies to generate cuts in the neighbourhood of the MP solution has been seldomly considered in the literature (Moreno et al., 2020). The upper bound can be enhanced when new better solutions are found during the evaluation of the neighbourhood of the current MP solution, while the lower bound can be boosted by the optimality cuts generated from the solutions found by the heuristics.

Our proposed solution is based on the heuristics used in Rei et al. (2009) and Santoso et al. (2005) with some modifications. In Rei et al. (2009), the idea of local branching is used and a step is added to the BD after solving the MP. In each iteration, after solving the MP, they solve the original problem with the addition of extra constraints, which use a Hamming distance function to divide the feasible region of the problem into smaller subregions based on the distance from the current solution of the MP. Then these smaller problems are solved to derive better solutions. They show that with this strategy the BD algorithm converges faster. Similarly, Santoso et al. (2005) apply a heuristic after solving the MP in each iteration, which they call the upper bounding heuristic. In their proposed heuristic, they fix some of the decision variables and solve the deterministic equivalent of the problem for a subset of the sampled scenarios for the remaining variables that are not fixed to derive better solutions (if possible).

In this study, we use similar ideas. In particular, we add a heuristic strategy to the BD algorithm after solving the MP and before solving the SP. After solving the MP in iteration v , we have the solution X^v that represents the schedule of the ED. In this solution, the value of $x_{j,h,k}$ is 1 if shift j starts at hour h with length of k . Therefore, a total of J decision variables of X^v will be equal to 1. We then solve the extensive form of the SAA problem as presented before, for all N scenarios with the addition of J new constraints. Based on this heuristic, when we solve the extensive form problem, we add a set of constraints which limit the start time of each shift to one hour before or one hour after the start time of the same shift in the current MP solution. Moreover, in order to shrink the feasible region even more, we keep the shift length the same as the shift length in the solution X^v . Therefore, if $x_{j,h,k}^v$ is equal to 1 for all $j = \{1, \dots, J\}$, $h = \{0, \dots, 23\}$ and $k = \{K_{min}, \dots, K_{max}\}$, we add the following constraint to the extensive form problem that we solve after solving the MP in each iteration:

$$x_{j(h-1)k} + x_{j(h)k} + x_{j(h+1)k} = 1$$

For instance, if based on the current solution of the MP, shift 1 starts at 5am (or hour $h = 5$) with a length of 7 hours (i.e., $x_{1,5,7}^v = 1$), then we add a constraint to the extensive form of the SAA that states that shift 1 should have a length of 7 hours and it can only start at hours $h = 4, 5$ or 6. For the cases where the start time of the shift in solution x^v is either hour 0 or hour 23, we only allow the shift to start one hour later or one hour earlier, respectively. By adding this set of constraints we can solve the problem to optimality in just a few minutes. We then pass the solution of the heuristic, which will be at least as good as the current solution of the MP to the SP and we continue from there. The advantage of this constraint compared to having a constraint based on Hamming distance is that the feasible region is narrower and the problem can be solved to optimality faster.

2.4.4 Symmetry Breaking Constraints

Symmetry is an important characteristic in many combinatorial problems, including our ISS problem. By adding additional constraints to the original problem formulation, we can reduce the size of the feasibility region and eliminate symmetric solutions. Below, we present the symmetry breaking constraints that we add to the MP.

$$\sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} x_{(j+1)hk} \leq \sum_{k=K_{min}}^{K_{max}} \sum_{h=0}^{23} x_{jhk} \quad j \in \{1, \dots, J-1\} \quad (2.16)$$

$$\sum_{k=K_{min}}^{K_{max}} x_{(j+1)hk} \leq \sum_{k=K_{min}}^{K_{max}} \sum_{h'=0}^h x_{jh'k} \quad j \in \{1, \dots, J-1\}; h \in \{0, \dots, 23\} \quad (2.17)$$

$$x_{jhk} \leq 1 - x_{(j+1)h(k-1)} \quad j \in \{1, \dots, J-1\}; h \in \{0, \dots, 23\}; k \in \{K_{min} + 1, \dots, K_{max}\} \quad (2.18)$$

Based on constraints 2.16, we should first start with scheduling shifts that have a lower index. More specifically, we can only schedule shift $j + 1$ in a day, if shift j has been scheduled already. Constraints 2.17 state that the start time of shifts with a lower index should be before the start times of shifts that have a higher index. In other words, the start time of shift j should be before the start time of shift $j + 1$. Lastly, when two shifts with different lengths are scheduled to start at the same time, constraints 2.18 require that the shift with the lower index should have a shorter length.

2.5 Numerical Experiments

This section presents the numerical results when the staffing and shift scheduling problem as formulated in Section 2.3 is applied to an ED. The case study that is used to generate the test instance in our study is the ED at the Foothills Medical Centre (FMC) in Calgary, Alberta (Canada). The ED consists of two separate areas. First, there is the fast track area, which is intended for patients who have been classified with a low severity score during triage and who are expected to be treated without having to be admitted to the hospital since the treatment plan is straightforward. Second, patients with a higher severity score are diverted to the main or intake area. Since the fast track area is assigned two dedicated physicians who are scheduled to work sequentially, we only consider the main and intake area in our case study.

Our data covers approximately 100,000 visits to the ED for the period of January 2017 until December 2018. During this two-years period, FMC has tried different schedules for their physicians and the latest schedule has 16 shifts of 7 hours each, which was introduced in July 2018. The actual schedule is presented in Table 2.1.

Shift	Hour of the day																							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	█	█	█	█	█	█	█																	
2								█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█
3																								
4																								
5																								
6																								
7																								
8																								
9																								
10																								
11																								
12																								
13	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█
14																								
15																								
16																								

Table 2.1. FMC Schedule

The data contains time stamps when the triage is performed as well as when an initial assessment and any re-assessments (if needed) are performed. This allows us to identify the arrival pattern of patients to the ED. We observed that the arrival process can be modelled as a Poisson process with varying averages during the day. The average number of patients arriving to the ED for each hour of the day is presented in Figure 2.1. It illustrates that the demand for emergency care increases quickly in the morning and slowly decreases in the afternoon and evening. According to this figure, the highest

number of arrivals is recorded around noon each day.

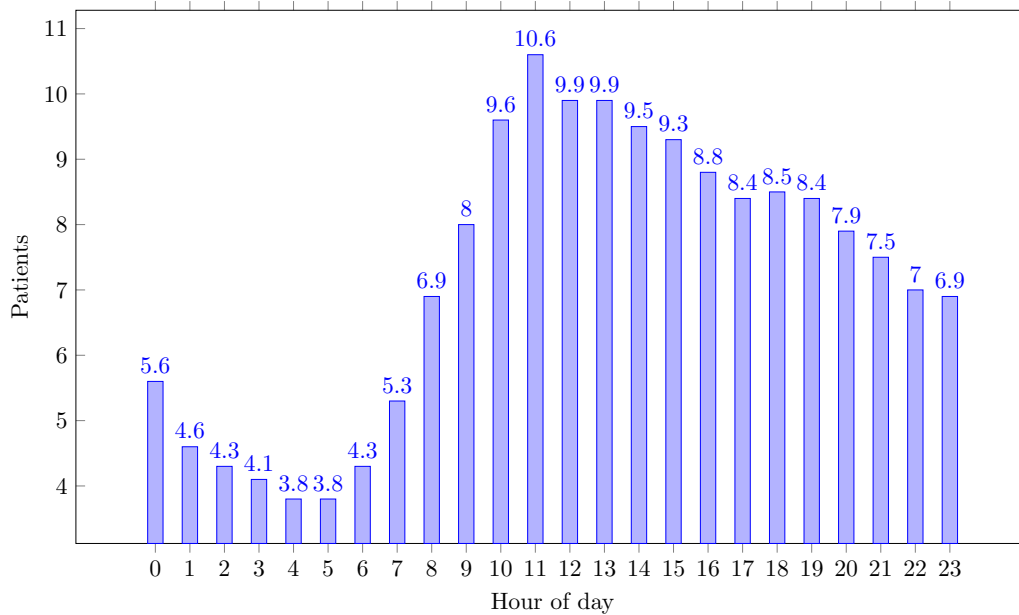


Figure 2.1. Average number of arrivals to FMC ED in a day

As discussed in previous sections, the number of new patients a physician can visit in each hour of the shift has a decreasing pattern. The total number of initial assessments performed in a physician's shift can be as low as 5 patients and as high as 19 patients. For more than half of the shifts, this number is between 9 and 12 patients. Rather than modeling the number of patients seen per hour, we model this number implicitly in two stages. First, we model the number of patients seen in a shift. Next, we model the time between initial assessments. We refer to this time interval as the inter-initial assessment time (or IIAT), which is conditional on the number of patients seen during a shift. Based on our data analysis, we find that a Weibull distribution best describes this conditional random variable. The scale parameter of the Weibull distribution for the IIAT depends on the patient number and the total number of patients seen in a shift, and is presented in Table 2.2. Based on this table, it becomes clear that the average IIAT increases toward the end of the shift, and it decreases as the total number of patients seen in a shift increases. We could not identify any theoretical distribution to fit to our data regarding the number of patients seen in a shift, and therefore, we use an empirical distribution to model this random variable. Based on this modeling approach, we can generate scenarios for IIAT, which are then translated to physician productivity (or PPH).

We also validate our variable PPH rate assumption using a multiple regression model, where

the average PPH serves as the dependent variable, while the hour of the shift, physician code, Waiting Room census, and several other variables act as independent variables. The results (Table 2.6) indicate that the hour of the shift (represented as a binary variable) is highly significant, demonstrating a notable decrease in PPH as the shift progresses.

IIAT	Total Number of patients seen in a shift														
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1-2	38	33	29	27	23	22	21	19	18	17	14	14	13	17	10
2-3	53	45	40	33	30	27	25	23	21	20	18	19	16	14	15
3-4	65	55	44	39	34	30	27	26	23	21	19	19	15	17	15
4-5	87	69	54	45	38	35	30	27	24	25	22	20	20	15	14
5-6		84	66	56	46	39	36	31	27	25	23	20	16	19	20
6-7			74	63	51	46	40	35	31	28	24	23	21	18	19
7-8				68	59	48	44	39	34	31	24	23	23	21	25
8-9					62	54	47	41	38	33	28	27	23	25	17
9-10						56	48	43	39	35	32	30	27	33	20
10-11							52	44	41	37	35	30	28	26	19
11-12								48	40	36	35	33	33	26	17
12-13									46	36	36	34	25	23	29
13-14										43	34	35	32	30	40
14-15											42	33	29	30	30
15-16												38	36	27	32
16-17													32	34	27
17-18														37	23
18-19															32

Table 2.2. The scale parameter of the Weibull distribution for the IIAT of patients in the shifts with different total number of patients seen in a shift

Using the generated scenarios, we solve the ISS problem for a planning horizon of 30 days and 100 scenarios. By adding the proposed heuristics to the enhanced Benders Decomposition algorithm, we can solve the problem to optimality (with an optimality gap of 1%) in 24 hours, which is about a 50% reduction in processing time compared to the enhanced BD algorithm without the proposed heuristics. The resulting schedule, which we call the Stochastic Schedule, is presented in Table 2.3. In order to have a fair comparison with the FMC Schedule, we assume that all shifts have a length of 7 hours and there are a total of 16 shifts in a day (i.e., $K_{min} = K_{max} = 7$ and $F = 7 \times 16 = 112$). When we compare the two physician schedules, there are more available physicians from 10 am to 3 pm in

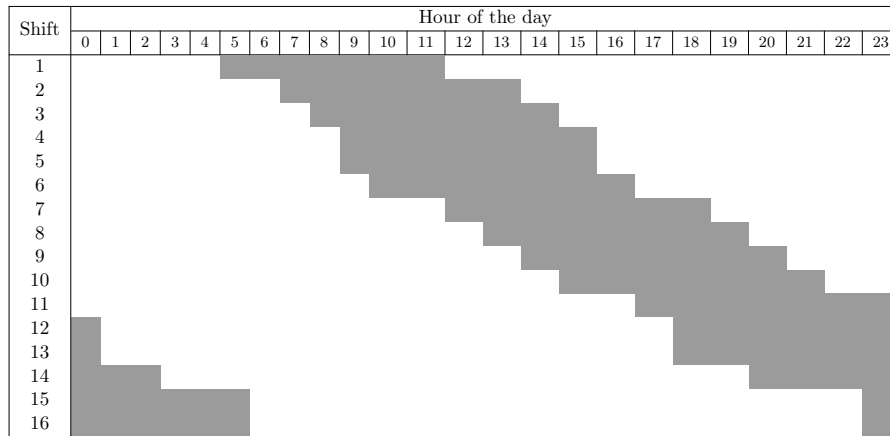


Table 2.3. Stochastic Schedule

our Stochastic Schedule. This exactly corresponds to when we expect to have more patient arrivals to the ED. In Figure 2.2 we compare the average demand for emergency care (measured as arrivals) to the average supply of emergency care available at the ED (measured as the total productivity over all physicians scheduled to work in a particular hour) for both schedules. For this figure, we can conclude the same result that on average the capacity of the ED is more around noon if we use the stochastic schedule. Moreover, the total PPH is more aligned with the average hourly arrivals in the stochastic schedule, which can result in a more efficient use of existing resources.

Other than the comparison against the shift schedule from FMC, we also study the impact of including the stochastic nature of both supply and demand for emergency care as well as including the time-dependent PPH rates when constructing the shift schedule. A schedule that is created without considering the randomness of the patient arrivals or the physicians’ productivity is referred to as Deterministic. Alternatively, when this randomness is taken into account, the schedule is referred to as Stochastic (similar to before). When it is assumed that the PPH rate is constant during the shift (when developing the shift schedule), this is denoted by NV (or Not Variable). Schedules that were created while considering variable (or time dependent) PPH rates is denoted by V. As an example, the V-Stochastic Schedule corresponds to the stochastic schedule in Figure 2.2. To evaluate and compare the schedules, we use a discrete event simulation model, which is presented in the following section.

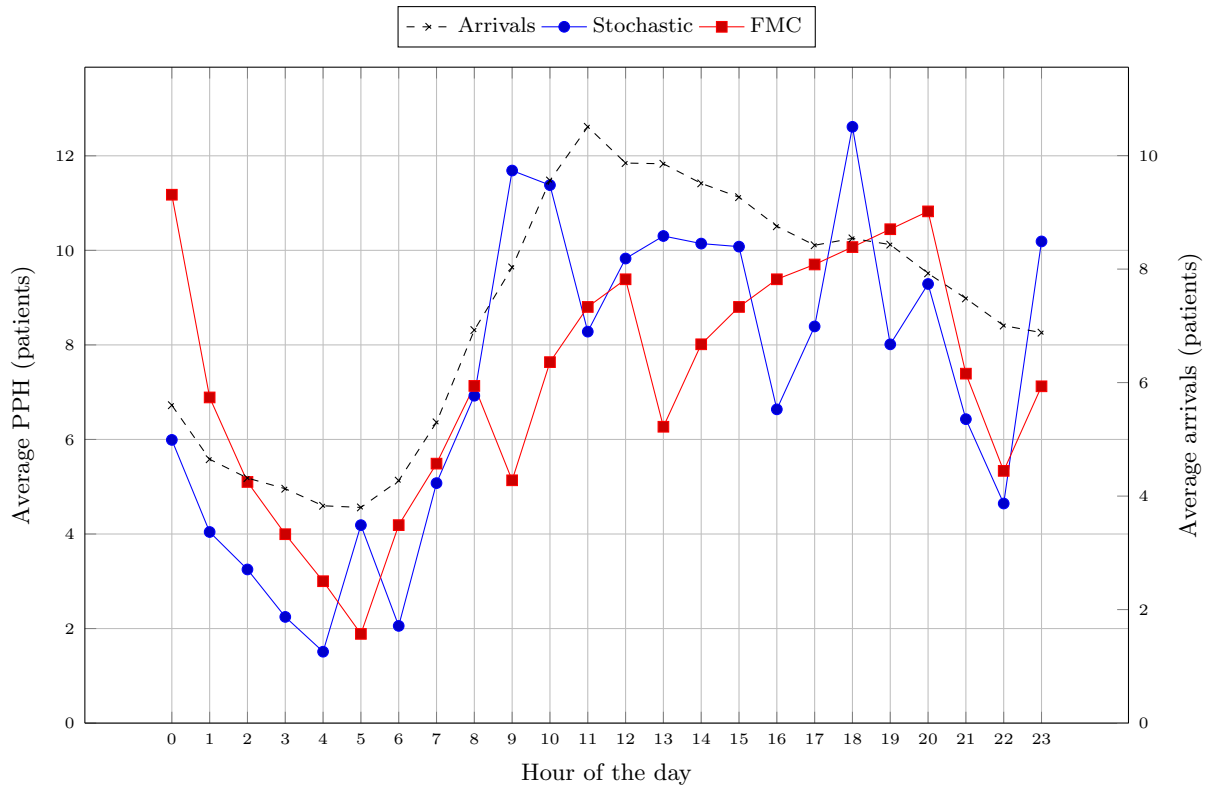


Figure 2.2. FMC vs Stochastic schedules

2.5.1 Simulation model

Our discrete event simulation model is developed based on the patient flow in an ED in the city of Calgary, Canada. Similar to most EDs in North America, a patient is seen by a triage nurse upon arrival to the ED. The nurse performs a quick assessment of the patient and creates an electronic record for the patient. After triage, all patients wait in a common waiting room, the patients' electronic records are placed in a virtual queue, and the waiting time of the patients begins. Then, whenever a physician becomes available, the physician chooses a new patient to visit from the virtual queue based on the patients' clinical information and their waiting times so far. The patient's initial assessment by the physician marks the end of the patient's waiting time.

Our simulation model, generates patients according to a non-homogeneous Poisson process with hourly rates estimated from the average number of patient arrivals in each hour of the day over the 2-year study period. Then, the patient joins a single station queue with a time dependent number of available physicians. At any time of the day, under a given schedule, the number of available physicians and the hour of their shifts are known. Once a physician completes the treatment of a

patient and becomes available, the physician is ready to sign up for a new patient. We assume patients are similar and do not consider patient prioritization in the simulation model as it does not impact our performance measure of the average waiting time of patients to see a physician in ED. Therefore, patients are seen by a physician and removed from the virtual queue based on a First Come First Serve basis. As mentioned before, we assume that the number of new patients a physician can see in an hour depends on the hour of the shift and the length of the shift that the physician is working at. In order to generate the inter-initial assessment times ($IIAT$) of patients for each physician in each shift, we first generate the total number of patients seen in a shift as a random number based on the empirical distribution from the data. Given this number (denoted by P), we use the Weibull distribution for the inter-initial assessment times between patients i and patient $i + 1$ visited in a shift ($IIAT_{i,P}$) conditional on P . We use this distribution as it yielded a best fit based on maximum likelihood estimation among other theoretical distributions.

The main outcome of interest in our simulation model is the average wait time of all patients in the 10 weeks period that we run the simulation for. We start the simulation with no patients waiting at the beginning of hour zero. Hence, we include a 2 days warm up period before each replications of the simulation. We simulate for 100 replications and use the average value of waiting time of all replications as the main output. The simulation model is developed using the Stochastic Simulation in Java (SSJ) library (L'Ecuyer, 2016).

Additionally, there are patients who will leave the ED without being seen by a physician if their wait time is longer than their patience time. In the data, there is no recorded time stamp when a patient actually left the ED before a disposition decision was made (either if the patient didn't receive an initial assessment or if the patient left before all re-assessments were completed). The only available time stamp for these patients is what is called the last contact time of the patient, which is the time stamp when a healthcare provider (either nurse or physician) noticed that the patient is not present anymore at the ED, which could be the time the patient is called to be seen by a physician (whereas the patient has left already). However, the percentage of patients who are classified as LWBS can be calculated by comparing the number of patients who got triaged by a nurse to the number of patients who got assessed by a physician. Furthermore, the patience time of a patient waiting in the waiting area for their initial assessment is usually modelled by an exponential distribution in the literature. We will use the same modeling approach in this paper. We choose an average patience time for patients

Schedule	Wait Time (hours)					Average LWBS
	Mean	Median	Std Dev	90% percentile	Max	
FMC	2.00	1.84	1.35	3.75	8.10	2.9%
NV-Deterministic	2.03	1.67	1.63	4.29	9.15	3.1%
V-Deterministic	1.85	1.56	1.46	3.85	8.12	2.9%
NV-Stochastic	1.87	1.49	1.52	3.98	8.00	2.6%
V-Stochastic	1.67	1.42	1.33	3.43	7.30	2.2%

Table 2.4. Simulation Results

such that the percentage of patients who leave the ED without being seen in our simulation model (with the same physician shift schedule as Figure 2.1) is similar to what we observe in the data for the ED at FMC. The same average patience time is used for evaluating the performance of all other shift schedules.

Moreover, in order to compare the schedules based on how shifts are similar to each other in terms of workload, we capture the total idle time of physicians during each shift. The idle time is calculated from the time a physician is ready and available to see a new patient until the time the physician actually starts assessing a new patient. There might be some times in the ED that there is sufficient capacity to assess a new patient, but there is no patient in the waiting room to be seen and the physician has to wait for a new patient arrival to the ED. In an ideal setting, all shifts would have similar average idle time, which shows a well balanced use of resources.

2.5.2 Simulation Results

We run the simulation model for all five schedule: FMC, NV-Deterministic, V-Deterministic, NV-Stochastic, and V-Stochastic. FMC schedule is the schedule that is being used by the FMC center during the period 2017-2018. NV-Deterministic schedule is the schedule resulting from the proposed formulation while instead of considering the stochasticity in arrivals and PPH we only use the hourly averages (hence deterministic) and do not consider the variable average for PPH during the shift (hence NV). V-Deterministic schedule is the same as NV-Deterministic expect that we have variable averages for PPH during the shift. NV-Stochastic schedule is the schedule resulting from our proposed formulation when we consider stochasticity in arrivals and PPH but use the same paramteres for PPH of all hours of the shift (i.e. not variable PPH), while in V-Stochastic we use different parameters for

PPH in different hours of the shift (i.e. variable PPH). We record the waiting time of each patient, the total idle time of each shift, and the total number of patients LWBS for each replication. Table 2.4 and Figure 2.3 present the output of the simulation model for each of the schedules. Based on the data and the publicly available reported measures of the ED at FMC, the average waiting time is 2 hours and the average rate of patients who leave without being seen corresponds to 3%, which are almost the same as the results from the simulation model for the FMC Schedule. According to Table 2.4, the lowest average wait time and the lowest rate of LWBS patients are for the V-Stochastic Schedule. It can be concluded that by using the V-Stochastic Schedule, we can decrease the average wait time by more than 16% (or almost 20 minutes) without introducing any new resources. Moreover, the variability of the patient wait times are lower for the V-Stochastic Schedule, which can be helpful in terms of planning and predicting the wait time. Another performance measure that is used by EDs is the 90% percentile wait time, which is only 3.43 hours for the V-Stochastic Schedule and the lowest among other schedules.

Simultaneously, according to Figure 2.3, the shifts are more similar in terms of workload in the Stochastic Schedule compared to the FMC Schedule, and the schedules in which we do not consider the time varying PPH. In the FMC schedule, there are some shifts that on average have almost 50 minutes idle time, and shifts that have on average only 13.7 minutes of idle time, while for the Stochastic schedule these numbers are 22.6 and 14.9 minutes, respectively. The reason can be due to the fact that in the FMC schedule, there are more available physicians during earlier hours of the day when we expect to have fewer arrivals, and consequently, the physicians who work at these hours would on average experience more idle times.

On the other hand, the average idle time for the stochastic schedule is the lowest among other schedules, which suggest the better use of resources and higher utilization rate. By comparing the performance metrics of all the schedules, it can be concluded that considering the time varying PPH has slightly more impact on the improvement of the measures than considering stochasticity of the inputs, which shows the importance of considering the time varying PPH in the staffing and scheduling problem. More particularly, the improvement in average wait time from using Stochastic Schedule as opposed to the NV-Stochastic Schedule is 11% while this reduction is only 8% for using the NV-Stochastic Schedule rather than NV-Deterministic Schedule. By considering the time inhomogenous and stochastic arrivals and PPH of physicians we can improve the wait time by more than 17%, i.e., V-

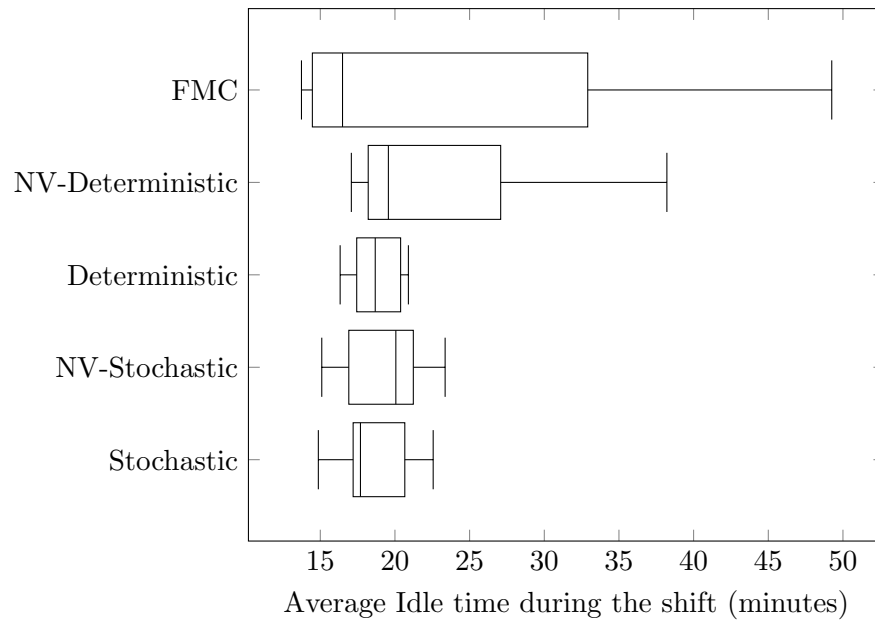


Figure 2.3. Comparison of Schedules with 16 shifts in a day in terms of workload distribution

Stochastic vs NV-Deterministic Schedules. It worth mentioning that the current schedule used in FMC ED, is already the result of many trial and error efforts by the scheduling team. More particularly, they have been working on the physician schedules for many years trying to find the most appropriate schedule that results in superior performance measures. However, based on the simulation results, our proposed model can still outperform the FMC schedule. Moreover, with any changes to the inputs, by using the proposed model, the new optimal schedule can be found more efficiently without the need for trial and errors.

We also study the average wait time in different hours of the day and compare the results for different schedules. Figure 2.4d, represents the average wait time of patients during the day for the FMC and Stochastic Schedules. Based on this figure, in the Stochastic Schedule, we have the lowest average wait time during the hours in which we have the highest number of arrivals. As a result, by allowing longer wait time for the hours with lower number of arrivals, we can decrease the total average wait time. This is done by having lower available capacity in earlier hours of the day and higher capacity just before the peak of arrivals in the Stochastic Schedule, compared to the FMC Schedule, Figure 2.2. By comparing the Stochastic Schedule with the NV-Stochastic Schedule, NV-Deterministic Schedule, and Deterministic Schedule in Figure 2.4, we can see that the Stochastic Schedule results in the lowest average wait time by having the lowest wait time during the hours with the highest number

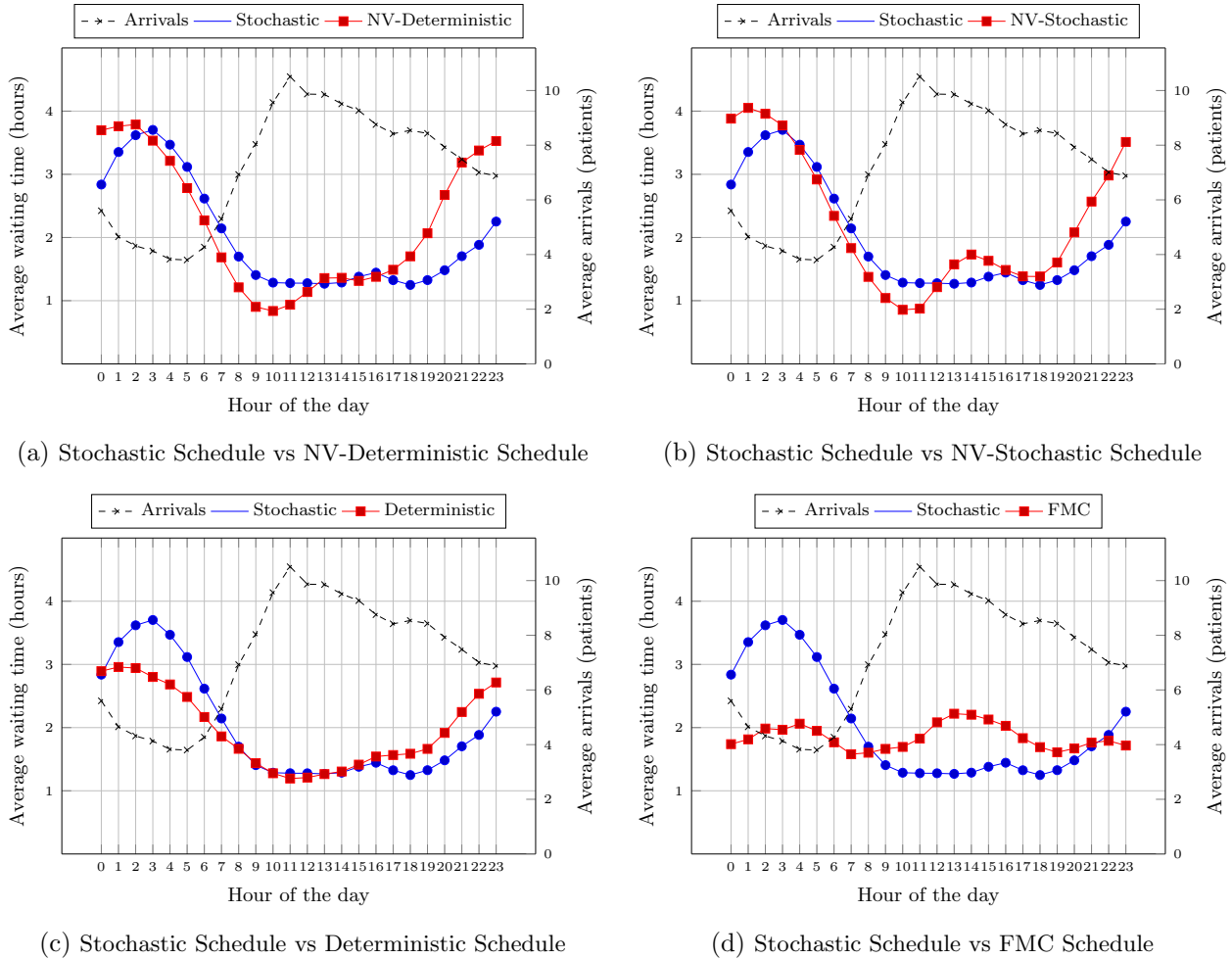


Figure 2.4. Hourly average waiting time of the Stochastic Schedule vs NV-Stochastic, NV-Deterministic, Deterministic, and FMC Schedules compared to the hourly average arrivals to the ED

of arrivals. In terms of average hourly wait time during the day, the Deterministic Schedule, is most similar to the Stochastic Schedule, Figure 2.4c, with the lowest average wait time during the hours of the day with the higher number of arrivals. The lower average wait time results in lower number of patients LWBS. In other words, by using the Stochastic Schedule, not only we achieve a lower average wait time but also we can visit more patients as the number of patients who leave without being seen is lower, Table 2.4.

Number of shifts per day	Schedule	Wait Time (hours)					Average LWBS
		Mean	Median	Std Dev	90% percentile	Max	
15	NV-Deterministic	3.99	3.50	2.48	7.64	12.04	5.9%
	Deterministic	3.99	3.57	2.30	7.35	12.35	5.8%
	NV-Stochastic	3.78	3.40	2.33	7.18	12.06	5.6%
	Stochastic	3.23	2.89	1.82	5.86	9.92	4.5%
16	NV-Deterministic	2.03	1.67	1.63	4.29	9.15	3.1%
	Deterministic	1.85	1.56	1.46	3.85	8.12	2.9%
	NV-Stochastic	1.87	1.49	1.52	3.98	8.00	2.6%
	Stochastic	1.67	1.42	1.33	3.43	7.30	2.2%
17	NV-Deterministic	1.38	1.07	1.25	3.13	8.00	2.1%
	Deterministic	1.25	0.98	1.07	2.74	7.17	1.9%
	NV-Stochastic	1.34	1.07	1.24	2.96	7.06	2.1%
	Stochastic	1.16	0.96	0.98	2.36	6.41	1.6%

Table 2.5. Sensitivity Analysis

2.5.3 Sensitivity Analysis

In order to see the effect of considering time dependent and stochastic physician productivity (or PPH rates) when scheduling physicians in EDs under different scenarios regarding the number of shifts, we run the optimization code with the assumption that we can have 17 or 15 shifts in a day (rather than 16 shifts). We evaluate the resulting schedules using the simulation model and the results are presented in Table 2.5. Based on the results, considering stochasticity and variable physician productivity as oppose to assuming deterministic and time homogeneous productivity (i.e., V-Stochastic Schedule vs NV-Deterministic Schedule) results in 20%, 18%, and 16% reduction in average wait times for the schedules with 15, 16, and 17 shifts in a day, respectively. Therefore, we can conclude that it is more important to have these considerations regarding physician productivity in shift scheduling when resources (or physicians) are more scarce. However, all performance measures are consistently better when using the V-Stochastic Schedules.

2.6 Conclusion

EDs play a pivotal role in providing timely urgent care to patients and availability of physicians is one of the main determinants of patients wait time in EDs. Staffing and scheduling physicians is a

challenging task due to limited budget and stochasticity in the system. In addition to the stochastic nature of patient arrivals and treatment times, the Patients Per Hour (PPH) rate of physicians in EDs fluctuates throughout a shift, with physicians typically seeing more new patients during the initial hours compared to the latter hours. The main goal of this paper is to find an optimal schedule for physicians in EDs to minimize the average wait time of patients by considering all the challenges mentioned above.

In this study, our objective is to devise a schedule aimed at minimizing the number of patients awaiting initial assessment by a physician. To tackle this staffing and scheduling problem, we formulated it as a two-stage stochastic problem. We solve the problem using an enhanced version of the Benders Decomposition Algorithm. We use a simulation model to evaluate the resulting policy and the numerical results show a significant reduction in average wait time of patients compared to the case where we do not consider stochasticity and or variability in PPH of physicians.

The proposed schedule achieved notable improvements in system performance while keeping the daily shift count constant. In other words, comparing the proposed schedule with the current schedule of FMC center in Calgary, we can reduce the average wait time of patients without introducing any additional resources. By strategically adjusting the start times of existing shifts without the need for additional shifts or their extension, and aligning them more closely with the influx of arrivals to the ED, we observed a significant reduction in the average waiting time for patients. Furthermore, our methodology offers inherent adaptability, allowing it to be fine-tuned to optimize scheduling outcomes in response to potential changes in the department's shift allocation. Whether the department opts to increase or decrease the number of shifts per day, or change length of the shifts, our approach remains flexible and capable of delivering optimized scheduling solutions tailored to evolving operational requirements. Another factor for evaluating a schedule is to have shifts that are similar in workload. The idea is that all shifts that are scheduled in a day, should have similar workload or similar average idle time. Using the simulation model, we also investigate this and compare different schedules based on the variability in average idle time of scheduled shifts. Results show that, the proposed schedule that considers stochasticity and the variable PPH, results in shifts that are more similar in terms of workload and average wait time.

In conclusion, our study highlights the critical importance of accounting for the variable PPH

rate in addressing the staffing and shift scheduling challenge. By incorporating this variable, we observed several significant benefits, including a reduction in average patient wait times, decreased variance in patient wait times, and more equitable distribution of workload among physicians across shifts. In essence, by accounting for the PPH of physicians rather than simply the number of available physicians, we can effectively mitigate overstaffing during the initial hours of a shift and alleviate understaffing during the latter hours. This nuanced approach not only enhances operational efficiency but also contributes to improved patient care outcomes within Emergency Departments. Moreover, using the enhanced solution approach, the problem can be solved in a reasonable time and the ED managers can use it as a tool to find optimal schedules, when they want to make any changes to number or lengths of shifts.

In this study we assume that physicians are the only bottleneck of the system and we only focus on availability of physicians. However, in many EDs lack of enough beds or nurse shortage are the main reasons for long wait time in EDs. Even in cases where we need more physicians, we need to consider other resources in the ED, as by improving availability of physicians, we might be moving the bottleneck from physicians to other resources. Considering all the other resources, complicate the problem further and results in moving the boundaries of the system from the ED only to the hospital and even other healthcare systems. In EDs, often the reason for having bed blocks is due to delay in transferring the admitted patients to the hospital. By considering the number of beds in the optimization system, we need to consider the throughput and resources of the hospital as well.

It's important to recognize that EDs are just one component of the broader healthcare system. Research has revealed that a significant proportion of patients who visit EDs in Canada do not actually have urgent medical needs. Often, when other parts of the system, such as access to family physicians, face shortages or challenges, patients find it more convenient to seek care in EDs. Therefore, improving the performance and wait times in EDs might even attract more demand and hence long wait times. To achieve effective improvements within EDs, a holistic approach is necessary. This means studying the entire healthcare system to ensure that expensive resources, specifically designed for urgent cases, are optimally utilized for their intended target patients.

Bibliography

- Ahmed, M. and Alkhamis, T., 2009. Simulation optimization for an emergency department healthcare unit in kuwait, *European Journal of Operational Research* **198**.
- Akbari, M., Zandieh, M. and Dorri, B., 2013. Scheduling part-time and mixed-skilled workers to maximize employee satisfaction, *The International Journal of Advanced Manufacturing Technology* **64**(5): 1–11.
- Al-Najjar, S. and Ali, S., 2011. Staffing and scheduling emergency rooms in two public hospitals: A case study, *International Journal of Business Administration* **2**: 137–148.
- Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y. and Yom-Tov, G., 2015. On patient flow in hospitals: A data-based queueing-science perspective, *Stochastic Systems* **5**: 146–194.
- Atlason, J., Epelman, M. A. and Henderson, S. G., 2008. Optimizing call center staffing using simulation and analytic center cutting-plane methods, *Management Science* **54**(2): 295–309.
- Badri, M. and Hollingsworth, J., 1993. A simulation model for scheduling in the emergency room, *International Journal of Operations & Production Management* **13**: 13–24.
- Batt, R. J., Kc, D. S., Staats, B. R. and Patterson, B. W., 2019. The effects of discrete work shifts on a nonterminating service system, *Production and operations management* **28**(6): 1528–1544.
- Batt, R. J. and Terwiesch, C., 2012. Doctors under load: An empirical study of state-dependent service times in emergency care, *The Wharton School, the University of Pennsylvania, Philadelphia, PA* **19104**.
- Batt, R. J. and Terwiesch, C., 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department, *Management Science* **61**(1): 39–59.
- Batt, R. and Terwiesch, C., 2017. Early task initiation and other load-adaptive mechanisms in the emergency department, *Management Science* **63**: 3531–3551.
- Bechtold, S. E. and Brusco, M. J., 1994. A microcomputer-based heuristic for tour scheduling of a mixed workforce, *Computers & operations research* **21**(9): 1001–1009.

- Benders, J. F., 1962. Partitioning procedures for solving mixed-variables programming problems, *Numerische mathematik* **4**(1): 238–252.
- Berry Jaeker, J. and Tucker, A., 2017. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity, *Management Science* **63**: 1042–1062.
- Birge, J. R., 1997. State-of-the-art-survey—stochastic programming: Computation and applications, *INFORMS journal on computing* **9**(2): 111–133.
- Bodur, M. and Luedtke, J. R., 2017. Mixed-integer rounding enhanced benders decomposition for multiclass service-system staffing and scheduling with arrival rate uncertainty, *Management Science* **63**(7): 2073–2091.
- Brusco, M. J. and Jacobs, L. W., 1998. Personnel tour scheduling when starting-time restrictions are present, *Management Science* **44**(4): 534–547.
- Brusco, M. J. and Johns, T. R., 1998. Staffing a multiskilled workforce with varying levels of productivity: An analysis of cross-training policies, *Decision Sciences* **29**(2): 499–515.
- Campbell, A. M., Gendreau, M. and Thomas, B. W., 2011. The orienteering problem with stochastic travel and service times, *Annals of Operations Research* **186**(1): 61–81.
- Centeno, M., Giachetti, R., Linn, R. and Ismail, A., 2003. Emergency departments ii: A simulation-ilp based tool for scheduling er staff, *Proceedings of the 2003 Winter Simulation Conference*, pp. 1930–1938.
- Chan, D. C., 2018. The efficiency of slacking off: Evidence from the emergency department, *Econometrica* **86**(3): 997–1030.
- Defraeye, M. and Van Nieuwenhuyse, I., 2016. Staffing and scheduling under nonstationary demand for service: A literature review, *Omega* **58**: 4–25.
- Delasay, M., Ingolfsson, A. and Kolfal, B., 2016. Modeling load and overwork effects in queueing systems with adaptive service rates, *Operations Research* **64**: 867–885.
- Delasay, M., Ingolfsson, A., Kolfal, B. and Schultz, K., 2019. The influence of load on service times, *European Journal of Operational Research* **279**: 673–686.

- Erhard, M., Schoenfelder, J., Fügener, A. and Brunner, J. O., 2018. State of the art in physician scheduling, *European Journal of Operational Research* **265**(1): 1–18.
- Ganguly, S., Lawrence, S. and Prather, M., 2014. Emergency department staff planning to improve patient care and reduce costs, *Decision Sciences* **45**(1): 115–145.
- Gans, N., Shen, H., Zhou, Y.-P., Korolev, N., McCord, A. and Ristock, H., 2015. Parametric forecasting and stochastic programming models for call-center workforce scheduling, *Manufacturing & Service Operations Management* **17**(4): 571–588.
- Ghanes, K., Wargon, M., Jouini, O., Jemai, Z., Diakogiannis, A., Hellmann, R., Thomas, V. and Koole, G., 2015. Simulation-based optimization of staffing levels in an emergency department, *Simulation: Transactions of the Society for Modeling and Simulation International* **91**: 942–953.
- Gnanlet, A. and Gilland, W. G., 2014. Impact of productivity on cross-training configurations and optimal staffing decisions in hospitals, *European Journal of Operational Research* **238**(1): 254–269.
- Goodale, J. C. and Thompson, G. M., 2004. A comparison of heuristics for assigning individual employees to labor tour schedules, *Annals of Operations Research* **128**: 47–63.
- Goodale, J. C. and Tunc, E., 1998. Tour scheduling with dynamic service rates, *International Journal of Service Industry Management* **9**(3): 226–247.
- Green, L., Kolesar, P. and Soares, J., 2001. Improving the sipp approach for staffing service systems that have cyclic demands, *Operations Research* **49**: 549–564.
- Green, L., Kolesar, P. and Whitt, W., 2007. Coping with time-varying demand when setting staffing requirements for a service system, *Production and Operations Management* **16**: 13–39.
- Green, L. V., Soares, J., Giglio, J. F. and Green, R. A., 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing, *Academic Emergency Medicine* **13**(1): 61–68.
- Hart, A. and Drall, S., 2007. Productivity: Do 8-9 hour shifts make a difference?, *Annals of Emergency Medicine* **50**: 69–70.

- Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y. and Wu, X., 2007. A survey and experimental comparison of service-level-approximation methods for nonstationary $m(t)/m/s(t)$ queueing systems with exhaustive discipline, *INFORMS Journal on Computing* **19**: 201–214.
- Ingolfsson, A., Haque, M. A. and Umnikov, A., 2002. Accounting for time-varying queueing effects in workforce scheduling, *European Journal of Operational Research* **139**(3): 585–597.
- Izady, N. and Worthington, D., 2012. Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments, *European Journal of Operational Research* **219**(3): 531–540.
- Jeanmonod, R., Jeanmonod, D. and Ngiam, R., 2008a. Resident productivity: Does shift length matter?, *American Journal of Emergency Medicine* **26**: 789–791.
- Jeanmonod, R., Jeanmonod, D. and Ngiam, R., 2008b. Resident productivity: does shift length matter?, *The American journal of emergency medicine* **26**(7): 789–791.
- Jennings, O., Mandelbaum, A., Massey, W. and Whitt, W., 1996. Server staffing to meet time-varying demand, *Management Science* **42**: 1383–1394.
- Kc, D. S., 2014. Does multitasking improve performance? evidence from the emergency department, *Manufacturing & Service Operations Management* **16**(2): 168–183.
- KC, D. and Terwiesch, C., 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations, *Management Science* **55**: 1486–1498.
- KC, D. and Terwiesch, C., 2012. An econometric analysis of patient flows in the cardiac intensive care unit, *Manufacturing & Service Operations Management* **14**: 50–65.
- Kim, K. and Mehrotra, S., 2015. A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management, *Operations Research* **63**(6): 1431–1451.
- Kuntz, L., Mennicken, R. and Scholtes, S., 2014. Stress on the ward: Evidence of safety tipping points in hospitals, *Management Science* **64**: 754–771.
- Kuo, Y., 2014. Integrating simulation with simulated annealing for scheduling physicians in an understaffed emergency department, *HKIE Transactions* **21**: 253–261.

- L'Ecuyer, P., 2016. SSJ: Stochastic simulation in Java, software library. <http://simul.iro.umontreal.ca/ssj/>.
- Li, C., Robinson, E. P. and Mabert, V. A., 1991. An evaluation of tour scheduling heuristics with differences in employee productivity and cost, *Decision Sciences* **22**(4): 700–718.
- Liu, R. and Xie, X., 2018. Physician staffing for emergency departments with time-varying demand, *INFORMS Journal on Computing* **30**(3): 588–607.
- Mass, A. and Moretti, E., 2009. Peers at work, *American Economic Review* **99**: 112–145.
- Moreno, A., Munari, P. and Alem, D., 2020. Decomposition-based algorithms for the crew scheduling and routing problem in road restoration, *Computers & Operations Research* **119**: 104935.
- Rei, W., Cordeau, J.-F., Gendreau, M. and Soriano, P., 2009. Accelerating benders decomposition by local branching, *INFORMS Journal on Computing* **21**(2): 333–345.
- Richardson, D. B. and Bryant, M., 2004. Confirmation of association between overcrowding and adverse events in patients who do not wait to be seen, *Academic Emergency Medicine* **11**(5): 462.
- Robbins, T. R. and Harrison, T. P., 2010. A stochastic programming model for scheduling call centers with global service level agreements, *European Journal of Operational Research* **207**(3): 1608–1619.
- Rossetti, M. D., Trzcinski, G. F. and Syverud, S. A., 1999. Emergency department simulation and determination of optimal attending physician staffing schedules, in P. Farrington, H. Nembhard, D. Sturrock and G. Evans (eds), *Proceedings of the 1999 Winter Simulation Conference*, Vol. 2, IEEE, pp. 1532–1540.
- Saghafian, S., Imanirad, R. and Traub, S., 2021. Do independent and parallel processing physicians influence each other's performance? evidence from the emergency department. Working Paper.
- Santoso, T., Ahmed, S., Goetschalckx, M. and Shapiro, A., 2005. A stochastic programming approach for supply chain network design under uncertainty, *European Journal of Operational Research* **167**(1): 96–115.
- Savage, D. W., Woolford, D. G., Weaver, B. and Wood, D., 2015. Developing emergency department physician shift schedules optimized to meet patient demand, *Canadian Journal of Emergency Medicine* **17**(1): 3–12.

- Schwarz, J., Selinka, G. and Stolletz, R., 2016. Performance analysis of time-dependent queueing systems: Survey and classification, *Omega* **63**: 170–189.
- Shapiro, A., 2003. Monte carlo sampling methods, *Handbooks in operations research and management science* **10**: 353–425.
- Shunko, M., Niederhoff, J. and Rosokha, Y., 2017. Humans are not machines: The behavioral impact of queueing design on service time, *Management Science* **64**: 453–473.
- Sinreich, D. and Jabali, O., 2007. Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance, *Health Care Management Science* **10**(3): 293–308.
- Stolletz, R., 2008. Approximation of the non-stationary $m(t)/m(t)/c(t)$ -queue using stationary queueing models: The stationary backlog-carryover approach, *European Journal of Operational Research* **190**: 478–493.
- Sun, B. C., Hsia, R. Y., Weiss, R. E., Zingmond, D., Liang, L.-J., Han, W., McCreath, H. and Asch, S. M., 2013. Effect of emergency department crowding on outcomes of admitted patients, *Annals of emergency medicine* **61**(6): 605–611.
- Thompson, G. M. and Goodale, J. C., 2006. Variable employee productivity in workforce scheduling, *European Journal of Operational Research* **170**(2): 376–390.
- Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E. and De Boeck, L., 2013. Personnel scheduling: A literature review, *European Journal of Operational Research* **226**(3): 367–385.
- Yang, Y. and Lee, J. M., 2012. A tighter cut generation strategy for acceleration of benders decomposition, *Computers & Chemical Engineering* **44**: 84–93.
- Zaerpour, F., Bijvank, M., Ouyang, H. and Sun, Z., 2022. Scheduling of physicians with time-varying productivity levels in emergency departments, *Production and Operations Management* **31**(2): 645–667.
- Zeltyn, S., Marmor, Y. N., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Shtub, A., Lauterman, T. et al., 2011. Simulation-based models of emergency depart-

ments: Operational, tactical, and strategic staffing, *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **21**(4): 1–25.

2.7 Appendix

Table 2.6. Estimation Results for the effect of various factors on PPH rates. Model 1 includes all variables as discussed in Section 2.5. Models 2-7 gradually remove variables from model 1.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model7
<i>Intercept</i>	1.755*** (0.101)	1.752*** (0.101)	1.747*** (0.097)	1.734*** (0.092)	1.446*** (0.078)	1.176*** (0.016)	0.604*** (0.076)
<i>ShiftHour (base=Hour1)</i>							
<i>Hour 2</i>	-0.448*** (0.025)	-0.447*** (0.025)	-0.447*** (0.025)	-0.447*** (0.025)	-0.450*** (0.025)	-0.446*** (0.025)	
<i>Hour 3</i>	-0.790*** (0.028)	-0.789*** (0.028)	-0.789*** (0.028)	-0.789*** (0.028)	-0.794*** (0.028)	-0.790*** (0.028)	
<i>Hour 4</i>	-0.954*** (0.030)	-0.953*** (0.030)	-0.952*** (0.030)	-0.953*** (0.030)	-0.958*** (0.030)	-0.954*** (0.030)	
<i>Hour 5</i>	-1.027*** (0.031)	-1.026*** (0.031)	-1.026*** (0.031)	-1.026*** (0.031)	-1.030*** (0.031)	-0.026*** (0.031)	
<i>Hour 6</i>	-1.383*** (0.035)	-1.383*** (0.035)	-1.383*** (0.035)	-1.383*** (0.035)	-1.384*** (0.035)	-1.380*** (0.035)	
<i>Hour 7</i>	-2.321*** (0.053)	-2.321*** (0.053)	-2.321*** (0.053)	-2.321*** (0.053)	-2.317*** (0.053)	-2.313*** (0.053)	
<i>Physician (base=MD001)</i>							
<i>MD002</i>	-0.436** (0.150)	-0.449** (0.150)	-0.450** (0.150)	-0.450** (0.150)	-0.454** (0.150)		-0.402** (0.150)
<i>MD003</i>	-0.099 (0.101)	-0.097 (0.101)	-0.097 (0.101)	-0.097 (0.101)	-0.081 (0.101)		-0.029 (0.101)
<i>MD004</i>	-0.406*** (0.112)	-0.407*** (0.112)	-0.408*** (0.112)	-0.407*** (0.112)	-0.385*** (0.112)		-0.332** (0.112)
⋮							
<i>MD109</i>	-0.293* (0.127)	-0.299* (0.127)	-0.299* (0.127)	-0.298* (0.127)	-0.330** (0.127)		-0.278* (0.127)
<i>Night (base=DayShift)</i>	-0.006 (0.022)	-0.009 (0.022)	-0.009 (0.21)				
<i>ED Census</i>	-0.003* (0.001)	-0.004*** (0.001)	-0.004*** (0.001)	-0.004*** (0.001)			
<i>Waiting Room Census</i>	-0.001 (0.002)	0.000 (0.002)					
<i>Boarder Census</i>	-0.002 (0.002)						
Log Likelihood	-11,286	-11,287	-11,287	-11,287	-11,304	-11,511	-13,496
AIC	22,810	22,810	22,808	22,806	22,839	23,037	27,210
BIC	23,650	23,643	23,634	23,625	23,651	23,086	27,980
Observations	8,624	8,624	8,624	8,624	8,624	8,624	8,624

Notes. This table reports estimation results from the Poisson Regression. Robust standard errors are shown in parentheses. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Chapter 3

When to extend a shift in Emergency Departments

3.1 Introduction

The scheduling of physicians in EDs is a challenging task because of a unique combination of characteristics in these systems. EDs pose inherent challenges in management, primarily due to their operation within a highly stochastic environment. This unpredictability stems from the variability in patient arrival rates, the fluctuating urgency levels of patients requiring healthcare intervention, and the consequent impact on physician availability and health resource allocation. On the other hand, ensuring timely service delivery in EDs is paramount as it can have a direct impact on the health outcomes of patients. Therefore, it becomes crucial to have the capability of meeting demand for emergency care within specified wait time parameters. This wait time, known as the target wait time, is defined as the duration from the arrival of the patient to the ED until the initial assessment performed by the physician. Typically, health ministries, in collaboration with clinical experts, determine these time frames based on the severity of patient's conditions.

A commonly employed approach in EDs involves the stratification of patients according to the severity of their condition, with subsequent prioritization based on this assessment. For instance, patients receive triage upon arrival to the ED and their acuity is measured by the CTAS (Canadian

Triage Acuity Scale) score in Canadian EDs. This is a measure on a 1 to 5 point scale, where CTAS 1 corresponds to resuscitation, and CTAS 5 to non-urgent care. A primary objective for both ED managers and physicians is to ensure timely access to physician consultation for a specific proportion of patients within specified timeframes. For instance, at least 98% of all patients who are categorized into CTAS 1 should be seen by a physician immediately. Similarly, the aim is for a physician to assess at least 85% of patients who are classified as less urgent categories such as CTAS 4 within 60 minutes after their arrival to the ED (Cildo et al., 2019).

Determining the optimal physician staffing levels and schedules is a pivotal endeavor in attaining these performance goals regarding wait times within EDs. Given that physicians constitute one of the most significant cost components in ED operations, there exists a financial constraint on the total number of hours allocated for physician coverage. Thus, striking a balance between patient wait times and scheduled physician hours is essential to minimize costs effectively. However, achieving this equilibrium presents a multifaceted challenge. First, the number of patient arrivals to the ED varies over the day and is stochastic in nature. Second, the physicians' productivity (measures as the number of initial assessments performed in an hour, or patient-per-hour (PPH) rate) will vary stochastically as the patient care needs are not the same for all patients, but also the average productivity depends on the hour of the shift. In the literature, it has been shown that physicians typically register fewer new patients during the latter hours of their shifts compared to the initial hours of the shift (Chan, 2018). This can be attributed to a number of different reasons. First, physicians may be occupied with follow-up tasks for patients who they have assessed earlier in their shift (i.e., performing a reassessment). Second, physicians may consciously limit the intake of new patients given the desire to conclude their responsibilities regarding the patients under their care before the end of their shift. In other words, they allow themselves sufficient time to finalize the treatment plans for patients who they have performed initial assessments for. In particular, any patients who are still under the care of a physician when the shift of that physician has ended need to be handed off to a different physician. This is to be avoided as much as possible as hand-offs are known to have a negative impact on the care quality received by the patient. Due to these reasons, the demand for and the supply of services within EDs exhibit high levels of variability. Complicating matters further, physician schedules must be created well in advance. Therefore, these schedules are constructed based on projected rather than actual demand-supply dynamics. This necessitates a careful balance between accommodating potential variations in

workload and providing physicians with adequate foresight to plan their tasks effectively.

In addressing the variability inherent in ED operations, a strategy often employed is the integration of flexibility into physicians' daily schedules. More specifically, schedules are allowed to have minor changes based on the realized demand and capacity of the system in real time. The most commonly used approaches are using overtime, having on call staff, who can be called in when required, or increase the capacity by adding extra shifts or using temporary workers in response to a surge in demand. One of the most straightforward and effective means of introducing flexibility is through the utilization of holdover overtime when the system fails to meet anticipated demand. Holdover overtime is defined as the extension of a worker's regular hours, and it serves as a mechanism to manage fluctuations in demand and uncertainties in service durations (Qin et al., 2015). Throughout this study, the terms "holover overtime" and "overtime" will be used interchangeably.

The clear advantage of overtime over the other two approaches is that physicians are already present at the ED. With on call staff or temporary workers, we need to make the decision well in advance since physicians need to be aware that they should be on call and we need to account for their commute time. However, by using the overtime strategy, the physician is already in the ED and no time will be wasted for the commute. Moreover, we can make the decision closer to the time that we need the extra capacity and we have more information about the demand. We restrict the overtime as a two-hours extension of an existing shift. To make sure that physicians do not work too long, the regular shift (without extension) can be shortened such that it can be extended by an additional two hours (e.g., shifts are scheduled as 6-hours shifts, which can become 8 hours). The most important question is which shifts to extend and what factors should be taken into consideration when making this decision. Other than deciding which shift to extend, it is also important to know when to make this decision for individual shifts.

Most of the studies that use overtime as a practice to incorporate flexible working times try to find a fixed schedule that minimizes total costs of the system, including overtime costs, while assuming that the shifts can be extended whenever the actual realized demand is higher than the expected demand (Fügener and Brunner, 2019; Bürgy et al., 2019; Zhu et al., 2020; Kao and Queyranne, 1985). However, in a complex and busy system like an ED, demand can be high for extended periods of time and the decision regarding the extension of a shift should be based on a policy in which not only the

current status of the system is considered, but the expected supply and demand in future hours should be taken into consideration as well. While there are studies in the literature that employ optimization models to update a scheduling plan in response to unexpected demand surges or diminished service rates (Easton and Goodale, 2005; Mehrotra et al., 2010), these approaches are predominantly situated in settings where demand forecasts are available at least one day in advance. Consequently, there is sufficient time to formulate and implement optimization strategies. However, in ED environments, there are many uncertainties that prevent any updates to the schedule well in advance. Consequently, there is a need for policies and strategies that make physician schedules more flexible and allow for swiftly addressing any fluctuations in the system's demand or capacity. Ideally, such interventions should be easily implementable and yield rapid adjustments within the system.

The decision regarding extending a shift should be based on the anticipated wait time for incoming patients in subsequent hours. It has been shown in the literature that the wait time can be predicted in the ED by variables that account for the workload and capacity in the ED (Ang et al., 2016). The workload in the ED can be captured by the number of arrivals and the number of patients currently waiting to be seen by a physician. The candidate predictor variables for the workforce can be the number of available physicians and the changes that happen to the number of available physicians in the next few hours based on the given shift schedule. While acknowledging the significance of nursing staff and bed availability in ED patient flow dynamics, this study focuses primarily on physician resources. We assume that there are enough beds and nurses in the ED, or otherwise these resources would change proportionally if the decision is made to increase the number of available physicians.

This study aims to demonstrate the efficacy of dynamically extending shifts, as opposed to simply adding extra shifts or permanently lengthening shifts that have been scheduled already. Ultimately, the goal is to reduce the average waiting time for patients through more efficient utilization of the available resources. In essence, we propose that dynamically extending a single shift per day by two hours according to a predetermined policy yields superior results compared to permanently increasing the length of a shift by the same duration. Both approaches result in an equivalent total number of working hours per day. However, dynamic shift extensions offer the advantage of flexibly allocating the additional hours to periods of time that experience the highest demand. By strategically extending the shift that is experiencing the greatest surge in demand, we can mitigate patient wait times effectively.

This approach optimizes resource allocation, and ensures that the extra capacity is utilized precisely when and where it is needed the most. Thereby, it is enhancing the overall operational efficiency and patient satisfaction.

We formulate the problem as an infinite horizon Markov Decision Process and solve the problem using a policy iteration algorithm such that we derive a shift extension policy. Our approach assumes that a schedule is already constructed prior to the extension decision, and the problem formulation will find a policy to extend these shifts (or not) based on the status of the system such that average patient wait times are minimized. We test the performance of the policy through simulation. For our numerical analysis, we use data from an ED in Calgary, Canada. The results show that having the same average number of extensions per day, the dynamic shift extensions based on our policy (i.e., derived from the policy iteration algorithm) always outperforms the fixed schedule (without extensions). We also solve the problem using lookahead policies and conclude that having only a 4-hours lookahead horizon can result in a policy that is as good as the solution found through the policy iteration algorithm.

The remainder of this chapter is organized as follows. Section 3.2 provides an in-depth review of any related literature. Section 3.3 elucidates the problem formulation and structural properties of the model. In Section 3.4, we present our empirical findings, simulation results, and analysis of the implications. Finally, in Section 3.5, we conclude with a summary of key insights and contributions.

3.2 Literature Review

Workforce flexibility has been extensively researched in the operations management literature for decades. The focus of this review is on papers that use overtime as an workforce flexibility method to address fluctuations in demand and uncertainty in capacity while scheduling workforce. According to a review by Qin et al. (2015), research on overtime mainly focuses on two issues: First, finding a schedule through which the amount of overtime is determined by minimizing different objective functions such as total cost, overtime cost, under staffing, and overstaffing . Second, the timing and when is the best time to employ overtime. In this review, we follow a similar categorization and categorize the papers based on whether they perform *Offline Scheduling* or *Online Scheduling*.

In studies that use *Offline Scheduling*, the schedule is generated in advance and overtime happens

when demand is more than the available capacity. The objective in these papers is to find a schedule that minimizes a cost function, where overtime can be used to deal with fluctuations in demand or capacity. On the other hand, in papers that use *Online Scheduling*, usually there exists a base schedule and the objective is to find out if and when to use overtime or to extend a shift in real time. The primary objective in these studies is to find the optimal circumstances when it is beneficial to extend a shifts and implement overtime when confronted with demand levels that exceed expectations or surpass available capacity. In the following paragraphs, we first review the papers that perform offline and online scheduling, and then discuss the positioning and contributions of this research to the existing literature.

Offline Scheduling: In this group of studies, overtime is used as an approach to deal with stochastic demand or capacity when finding a pre determined schedule that minimizes overtime (Fügenger and Brunner, 2019; Gans and Zhou, 2002; Kao and Queyranne, 1985; Bard et al., 2007), understaffing and overstaffing (Parisio and Jones, 2015), or the total costs of the system (Bürgy et al., 2019; Zhu et al., 2020; Fry et al., 2006; Campbell, 2012; Liao et al., 2012; Silva and de Souza, 2020). Using a probabilistic model, Kao and Queyranne (1985) show that ignoring demand uncertainty can result in underestimates of budget needs, which (to some extent) can be solved by using overtime. Extending the work of Kao and Queyranne (1985), Campbell (2012) allows for the use of pre-scheduled on-call overtime, which is shown to be cost effective only if the level of demand uncertainty is low or holdover overtime is not available. In their model, pre-scheduled overtime costs the same as holdover overtime when an on-call worker is called in. Otherwise, they are paid a standby rate. Pre-scheduled on-call overtime is mainly used when there is a probability that a worker does not accept to stay longer upon request. A similar concept is applied by Fügenger and Brunner (2019), where variable shift extensions are allowed and if a variable shift extension is scheduled, the physician might have to stay a few periods longer with a given probability. This results in a lower amount of unplanned overtime for physicians. While most of the studies use different optimization models to schedule workforce, Dynamic Programming has been used in only two studies in which the objective is to find a policy for the number of employees to hire and how to schedule arriving surgeries, Gans and Zhou (2002) and Silva and de Souza (2020), respectively, such that the overall costs of the system (including overtime costs) are minimized.

Online Scheduling: An algorithm is called online if exogenous future information is unknown and the algorithm makes decisions based on past and current information (Keyvanshokoo et al., 2021).

In the studies in this group, an optimization model is used to find the optimal schedule adjustment plan once the demand happens (if necessary). Easton and Goodale (2005) model optimal responses to unplanned employee absences in a system with multi servers that provide discrete pay per use services with customer abandonment. Initially they find a schedule where absences are anticipated by considering an expected absenteeism rate. Having holdover overtime, call-ins, and temporary workers as possible responses, they find that the best recovery decision using an optimization model that considers the costs of lost sales resulting from customer abandonment and the wages paid for the extra working hours of the staff. They conclude that holdover overtime generally outperforms the other two options, as it allows managers to apply exact amounts of overtime labour at the most advantageous times. Similarly, in another study for a call center setting (Mehrotra et al., 2010), an integer programming model is used to find a cost effective agent re-scheduling plan (including shift extensions) once a change in agent demand has been determined. Overtime has also been considered in the field of patient scheduling. In a recent study by Keyvanshokoh et al. (2021), once a request is made from a patient to book an appointment, an optimization problem is solved to make two instantaneous and irrevocable decisions: an allocation decision for each incoming customer and an overtime decision for each server on each day with the objective of maximizing the cumulative reward of allocating a patient to a server minus the cumulative overtime cost over a planning horizon.

Finally there are some papers in the literature that perform both *offline* and *online scheduling*. Bandi and Gupta (2020) use a 2-step robust optimization approach for staffing and scheduling of operating rooms. In the first step, a baseline staffing level is found that minimizes the total costs including overtime costs while assuming that staffed operation room demand that is not met by the available staff is satisfied with the help of overtime. In the next step, knowing the exact surgery mix, a detailed schedule is generated that minimizes the cost of overtime where the length of the surgeries are stochastic. In another study by Wright and Bretthauer (2010), nursing shortage has been addressed using a 3-step approach. While the first step is similar to the first step by Bandi and Gupta (2020), in the second step float nurses are used first to account for the difference between the original forecast and the patient demand actually experienced for each shift. If the target staffing level is not met after the second step, then overtime is used. Finally, Pinker and Larson (2003) combine *offline* and *online scheduling* to determine the regular and contingent worker pool sizes over a finite planning horizon. They formulate the model as an optimization problem that minimizes regular labour costs

and the expected cost of backlog and any used overtime. The expected cost of overtime and backlog is included by an embedded dynamic programming problem that finds the optimal decisions regarding the utilization of overtime resources.

Positioning of this study: In this study we find an optimal policy for when to extend a shift in an emergency department. We show that dynamic shift extensions based on the state of the system can result in lower average waiting times for patients. As far as we know, this is the first study that investigates the question of when to extend a shift in a setting where the demand for services and the service rate are both stochastic, and the system is so volatile that this decision should be made very frequently. Formulating the problem as an infinite horizon MDP, the shift extension decision is made by considering the current state of the system as well as the expected number of arrivals and anticipated capacity already scheduled for future periods. As discussed above, the timing of using overtime is only considered in the papers that work on *On-line Scheduling* and in those papers the future is not considered in the decision making process. However, the average number of arrival and the available number of physicians in EDs are non-homogeneous during the day, and it is very important to take that into consideration when deciding on extending a shift.

3.3 Problem Formulation

The objective of this study is to find an optimal policy for extending physician shifts in EDs based on the state of the system (i.e., based on ED census). We assume that a physician schedule is provided, which specifies the number of shifts that are scheduled in a day and their start times. Here we assume that all shifts have the same length of six hours. The schedule is represented by $L = \{L_h\}$ for $h = 0, \dots, 23$, where L_h is the number of shifts that start at hour h . The objective is to determine how many shifts to extend based on the number of patients waiting to be seen and the parameter values for the upcoming hours, including the average number of patient arrivals as well as the total productivity of the ED system. The total productivity depends on the available number of physicians that are scheduled to work as well as the hour of the shift that each physician is in. The latter is important because we assume that physicians have time dependent productivity rates. We also assume that the decision to extend a shift should be made two hours before the end of a shift and once the decision is made the average productivity of the physician will change, since there are two

additional hours that the physician can work. Therefore, the total productivity of the ED (as a sum of the PPH rates for individual physicians) would also depend on the actions that are taken in the preceding hour, two hours and three hours, which is explained in more details in this section.

We formulate the problem as a discounted infinite horizon MDP model by providing the decision epochs, state space, action sets, transition probabilities, and costs in this section. We chose a discounted model with a discount factor close to one, rather than an average reward model, as it reflects the medium term planning horizon that is most applicable in the hospital and ED setting (Patrick et al., 2008). By discounting only slightly, we can make sure that costs are relatively similar in the short term, while far distant costs are less valued. We use the schedule in Table 3.1 as an example to describe the dynamics of our MDP model. Based on this schedule, there are ten shifts in a day with starting times ranging from midnight to 9 pm. The regular length of all shifts is 6 hours which is shown in light grey. Each shift may or may not be extended for two hours. The possible shift extensions are shown in dark grey. Hours 5 and 6 of each shift are differentiated as the physician’s productivity depends on whether the shift is actually extended or not.

Shift	Hour of the day																							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	1	2	3	4	5	6	7	8																
2					1	2	3	4	5	6	7	8												
3					1	2	3	4	5	6	7	8												
4									1	2	3	4	5	6	7	8								
5										1	2	3	4	5	6	7	8							
6											1	2	3	4	5	6	7	8						
7														1	2	3	4	5	6	7	8			
8	8																	1	2	3	4	5	6	7
9	6	7	8																1	2	3	4	5	
10	4	5	6	7	8																	1	2	3

Table 3.1. Example Schedule, all shifts are initially 6 hours and might be extended by two hours

3.3.1 Decision Epochs

At the beginning of each hour, the decision maker should decide if a shift that is about to end in two hours should be extended for an additional two hours or not based on the ED’s current status. We assume that this decision to extend a shift is made two hours before the end of the shift, such that the physician has sufficient time to plan their tasks that can be done during the remainder of the shift. At the beginning of each hour, the decision maker examines the schedule to determine if any shift would end within the next two hours. During hours where no shift is ending within this timeframe,

no decision needs to be made. For example, this is the case at the beginning of hours 0, 2, and 3 for the schedule illustrated in Figure 3.1. However, during the remaining hours, the decision maker can determine whether to extend a shift (or not) based on the system's state. Moreover, it is assumed that extending the duration of a shift by two hours would be considered optimal. Any duration less than two hours would likely not allow physicians adequate time to accommodate additional patients to be treated, whereas exceeding two hours could lead to fatigue and compromising the quality of patient care (Frank and Ovens, 2002), particularly as it will extend the total length of the shift beyond eight hours.

3.3.2 State Space

The decision to extend a shift in an ED is mostly dependent on the expected waiting time of the patients as it is one of the most important performance indicators of the system. The expected waiting time of patients in an ED can be calculated using the general formula of dividing the workload to the processing time (Ang et al., 2016). Therefore, we need to capture these two factors in our state space in order to have the necessary and sufficient information to make the extension decision and compute the transition and cost functions. For workload, we can use the number of patients waiting in the ED and the expected number of arrivals in the next hour. The number of arrivals to the system in each hour is considered to be stochastic where the average is dependent on the hour of the day. To determine the processing time, we need to know the number of physicians available at any given hour, their designated shift hour according to the provided schedule, and whether any shifts have been extended within the last three hours (as such extensions will contribute to the productivity of the ED). The productivity of individual physicians also is a random variable with an average that depends on the hour of the shift.

Consider for example that the decision maker has to decide what to do at hour $t = 1$ for the schedule provided in Table 3.1: there is one physician scheduled to work in the second hour of their shift, another in their fifth hour, and potentially one more in the seventh hour (if shift 10 was extended). The average productivity rate of the physician who is in the fifth hour of their shift depends on our decision at the current hour (i.e., at $t = 1$) regarding whether or not to extend the shift. If we choose to extend it and the shift is scheduled to end in four hours, the productivity will be higher compared

to the scenario where the shift concludes in two hours (since the physician will start to wrap up if the shift is not extended). Additionally, the availability of the third physician who might be in the seventh hour of their shift is contingent upon the decision that has been made two hours ago (regarding the extension of the ninth shift). Therefore, a state of the system, denoted by $\vec{s} \in S$, takes the following form:

$$\vec{s} = (h, W, E^1, E^2, E^3)$$

Here, h is the hour of the day, which is used to determine the available number of physicians that are working according to the schedule that is provided, the hour of their shift, and the expected number of arrivals to the system during hour h . Each time unit is considered to be one hour starting from the beginning of the day at hour 0. Therefore, the hour of the day in the next state would be hour $h + 1$, and the hour of the day becomes hour 0 of the next day if h equals 23. Let W represent the number of patients waiting for their initial assessment at the beginning of hour h . The schedule of the ED determines how many physicians are available at hour h as well as the hour of the shift that they work in at hour h . However, in order to know the exact number of available physicians at each hour, we also need to know if we have decided to extend any shift in the past three hours. E^1, E^2 , and E^3 are integer numbers representing the number of shifts that have been decided to get extended 1, 2, and 3 hours ago, respectively. If we decided to extend one shift 3 hours ago (i.e., $E^3 = 1$), the last hour of this extension will start at hour h , which means that we have one extra physician available due to this extension. Similarly, if we decided to extend a shift two hours ago (i.e., $E^2 = 1$), its extension would start at hour h . In this study we assume different average productivities for the physicians during their shift. We also assume that the productivity in the 5th and 6th hours of the shifts that will be extended would be different from the productivity of the 5th and 6th hours of the shifts that will end in their regular length. Consequently, we need to know the decision whether any shift were extended one hour ago (that is E^1), as it will affect the (average) productivity of the 6th hour for that shift, and consequently, the total productivity of all physicians working at the ED in hour h . For example, if we are at hour 16 of the day in our example above, we need to know the value of E^3 and E^2 to know if shifts 4 and 5 have been extended as the 8th and 7th hour of those shifts would be at hour 16, respectively. We also need to know the value of E^1 as it will affect the productivity of the physician working at the 6th hour of shift six. Based on these variables, we can calculate the total productivity

of the ED and the total number of patients who can be seen (as an average) in hour h . We also assume that any unmet demand would be transferred to the next hour.

In order to have a finite state space, we assume an upper bound on the number of patients waiting in the waiting room ($W \leq WUB$). This upper limit WUB is set sufficiently large such that it is only for practical purposes and it has no significant impact. Moreover, we will see later that the optimal policy is of a control-limit type, and it is based on the parameters used in the model. In particular, there will be a threshold on W for each combination of h , E^1 , E^2 , and E^3 , such that it is always optimal to extend the shift if the state space exceeds that threshold value. Therefore, the upper limit for W should be set sufficiently higher than this threshold.

3.3.3 Action Sets

At the beginning of each hour, we need to decide if we want to extend a shift that is at the beginning of its 5th hour, if any. The action taken in each state is denoted by $a_{\vec{s}}$, which represents the number of shifts to extend. The set of feasible actions compatible with state $\vec{s} \in S$ is denoted by $A_{\vec{s}}$, and must satisfy the following constraint:

$$a_{\vec{s}} \leq L_{h-4}$$

Based on this constraint, $a_{\vec{s}}$ should be less than or equal to the number of shifts that have started exactly four hours ago. The value of $h - 4$ is less than 0 for the hours $h = 0, 1, 2$, and 3. In these cases, $h - 4$ would be the last hours of the previous day (i.e., the hours 20 to 23 of the last day, respectively). All actions $a_{\vec{s}}$ are constrained to be positive and integer.

3.3.4 Transition Probabilities

Once the shift extension decisions are made, the only source of uncertainties in the transition to the next state of the system include the number of arrivals at hour h (denoted by random variable N) and the total productivity of the physicians working at hour h (denoted by T), which depends on the hour of the shift for each physician that is working at hour h . The number of arrivals to the system follows a Poisson process with different hourly averages during a day. For the productivity of

the physicians we use the empirical probability distribution from the data with varying hourly averages during a shift. The average productivity for the last two hours of an extended shift is assumed to be different compared to the average productivity of the last two hours of a regular shift. Based on the expected total productivity and the expected number of arrivals at hour h , we can find how many patients will be seen and removed from the waiting room at hour h . Once we determine how many patients are removed or added to the waiting room at hour h , we know the state of the ED in the next hour.

When the current state equals $\vec{s} = (h, W, E^1, E^2, E^3)$, we will go to state $\vec{j} = (h + 1, W + N - T, a_{\vec{s}}, E^1, E^2)$ with probability $p(\vec{j}|\vec{s}, a_{\vec{s}}) = p(N - T|\vec{s}, a_{\vec{s}})$, or the probability of having $N - T$ patients added (or removed, if negative) from the waiting room at hour h . We define a period in our model to be one hour, such that the hour of the day for the next state becomes $h + 1$ or 0 when $h = 23$. The number of patients in the waiting room in the current state (i.e., W) would be increased by the number of arrivals N during hour h , and decreased by the number of patients seen by the available physicians T during hour h . Consequently, we can say that the number of patients in the next state (i.e., in state \vec{j}) would equal the maximum of $W + N - T$ and zero. The decision $a_{\vec{s}}$ that we make in the current state will be the decision that we made one hour ago in the next state. Similarly, the decisions we made in the previous two hours from hour h (i.e., E^1 and E^2 , respectively) will become the decisions that we made 2 and 3 hours ago when considering the next state at hour $h + 1$.

The probability of going to state \vec{j} from state \vec{s} given that we make the decision $a_{\vec{s}}$ is provided by $p(N - T|\vec{s}, a_{\vec{s}})$. The probability of having N arrivals depends on the hour of the day (or h). However, more factors have to be considered when calculating the probability of having a total productivity of T in a particular hour. In this study, we assume that the average PPH rates vary during the shift. For example, the average PPH rate can be higher in the first hours of the shift, and it then decreases to lower values for later hours of the shift. Furthermore, we assume that our action regarding extending the shift would affect a physician's productivity in the 5th and 6th hours of their shift, since they will be working in the ED for four more hours instead of just two more hours if they are notified that their shift is extended. Therefore, in order to calculate the probability $p(N - T|\vec{s}, a_{\vec{s}})$, we need to know how many physicians are available and in which hour of their shift they are working, while differentiating between the 5th and 6th hours of the extended and not extended shifts. Thus, we define the following vector for each state:

$$\vec{O}_{\vec{s}} = \{o_1, o_2, o_3, o_4, o_5, o_{5e}, o_6, o_{6e}, o_{7e}, o_{8e}\}$$

In this vector, o_1 to o_4 are the number of physicians working in the 1st to 4th hour of their shifts, respectively, at hour h . There are different variables for the number of physicians who are working at the 5th and 6th hours of shifts that are not extended (i.e., o_5 and o_6), and shifts that are extended, (i.e., o_{5e} and o_{6e}). Lastly, o_{7e} and o_{8e} are the number of physicians who are working in the 7th and 8th hour of the extended shifts. Once we know the value of vector $\vec{O}_{\vec{s}}$, we can calculate the probability distribution for the total productivity at state \vec{s} . To clarify this description, consider the schedule in Table 3.1. Based on this schedule and possible actions taken in the past, we can see that we will have several options for the number of available physicians working in the ED at hour 0. At this hour, we will have one physician working in the 1st hour of their shift and a physician working in the 4th hour of their shift. We have also a physician working in the 6th hour of their shift, but their expected PPH rate depends on the action taken in the previous period. If one hour ago, at hour 23 of the last day, it was decided to extend shift 9, then o_{6e} of the vector $\vec{O}_{\vec{s}}$ will be 1 and o_6 will be zero. Finally, we may or may not have a physician working in the 8th hour of shift 8, depending on the decision made three hours ago (i.e., E^3). Therefore, the vector $\vec{O}_{\vec{s}}$ for state $\vec{s} = \{h = 0, W, E^1, E^2, E^3\}$ will be $\vec{O}_{\vec{s}} = \{1, 0, 0, 1, 0, 0, 1 - E^1, E^1, 0, E^3\}$, which can have the following four possible values:

$$\vec{O}_{\vec{s}}^1 = \{1, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$$

$$\vec{O}_{\vec{s}}^2 = \{1, 0, 0, 1, 0, 0, 1, 0, 0, 1\}$$

$$\vec{O}_{\vec{s}}^3 = \{1, 0, 0, 1, 0, 0, 0, 1, 0, 0\}$$

$$\vec{O}_{\vec{s}}^4 = \{1, 0, 0, 1, 0, 0, 0, 1, 0, 1\}$$

Based on data, we have the probability distribution for the productivity of physicians in the first to the last hour of the shift. Therefore, we can calculate the total productivity of the ED based on the vector $\vec{O}_{\vec{s}}$. For instance, the corresponding vector for state $\vec{s} = (h = 0, W, E^1 = 0, E^2 = 0, E^3 = 0)$ would be $\vec{O}_{\vec{s}}^1 = \{1, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$, which states that there are three physicians available in the system who are at the beginning of the 1st, 4th, and 6th hour of their shifts. Moreover, the physician who is at the beginning of hour six of their shift will not have an extended shift. Based on this information, we are able to find the corresponding probability distribution for the productivity from the data. Therefore, once we know the current state and the action we take in that state, we can find the probability distribution for the total productivity of the system when solving the MDP.

3.3.5 Costs

The cost associated with each state action pair takes into consideration the relative cost of extending a shift to the cost of one hour of patients' wait time. We write the total cost for a given state as:

$$c(\vec{s}, a_{\vec{s}}) = W + \alpha a_{\vec{s}}$$

The objective of the MDP is to minimize the discounted total cost. The cost function balances the cost of the patients who are waiting for an initial assessment and the cost for the ED in having to extend a shift and to pay physicians. The multiplier α is a positive number defined as the cost of extending a shift relative to the wait cost of patients. By giving more weight to the wait time of patients (or having a lower value of α), we will have a policy in which we would extend a shift more frequently. On the other hand, by having higher values of α we will have an optimal policy that results in fewer number of extensions and consequently longer wait times for patients. According to our simulation model, the average wait time and the number of extensions per day are related to the value of α through the optimal policy corresponding to that α . Therefore, we can choose the value of α , and consequently the optimal policy, based on the target wait time of the system or based on the available budget for the average number of shift extensions per day.

3.3.6 Optimality Equations and Solution Approaches

The optimal value function in our formulation is denoted by $v^*(\vec{s})$, and corresponds to the minimum discounted cost over the infinite horizon for each state and satisfies the following optimality equations for all states:

$$v^*(\vec{s}) = \min_{a_{\vec{s}} \in A_{\vec{s}}} \{W + \alpha a_{\vec{s}} + \lambda \sum_{N-T=LB}^{UB} p(N-T|\vec{s}, a_{\vec{s}}) v^*(h+1, W + N - T, a_{\vec{s}}, E^1, E^2)\} \quad \forall \vec{s} \in S \quad (3.1)$$

where λ is the discount rate, and $p(N-T|\vec{s}, a_{\vec{s}})$ is the probability that the number of arrivals to the system minus the number of patients seen by the available physicians at hour h is equal to $N-T$. Based on the data, we can find a lower bound LB and an upper bound UB for the value of $N-T$.

We solve this problem using policy iteration algorithm to derive the optimal policy $a_{\vec{s}}^*$:

$$a_{\vec{s}}^* = \arg \min_{a_{\vec{s}} \in A_{\vec{s}}} \{W + \alpha a_{\vec{s}} + \lambda \sum_{N-T=LB}^{UB} p(N-T|\vec{s}, a_{\vec{s}}) v^*(h+1, W + N - T, a_{\vec{s}}, E^1, E^2)\} \quad \forall \vec{s} \in S \quad (3.2)$$

It can be shown that the policy iteration algorithm terminates in a finite number of iterations with a solution of a discount optimal policy and the optimality equation since the state space is finite and action space $A_{\vec{s}}$ is also finite for each state $\vec{s} \in S$ (Puterman, 2014). In the next section, we show that the optimal policy has a monotone non-decreasing property and we can use a monotone policy iteration algorithm to solve the problem, which is presented in Algorithm 1. The difference between the monotone algorithm with the regular policy iteration algorithm is that because of the non-decreasing property of the optimal policy in W , we do not need to enumerate over all states. Once the optimal policy is to extend all possible shifts at an hour for a specific value of W , we know that the optimal policy would be the same for all states with the same values of h , E^1 , E^2 , and E^3 but higher values of W . We also solve the problem using a lookahead model presented by Powell (2014) and compare the results with the monotone policy that results from the policy iteration algorithm.

3.3.7 Structural Properties

In this section, we provide some structural properties of the policy that follows from the MDP model presented in the previous section. One of the principle uses of MDP formulations is to establish the existence of optimal policies with special structure, which can be appealing to decision makers, easy to implement, and enabling efficient computations (Puterman, 2014). Based on numerical results, we observe that the optimal policy for state $\vec{s} = \{h, W, E^1, E^2, E^3\}$ is of a control limit type, which is a deterministic Markov policy composed of decision rules of the form:

$$a_{\vec{s}} = \begin{cases} 0, & \text{if } W_{\vec{s}} < B_{h,E^1,E^2,E^3} \\ 1, & \text{otherwise} \end{cases} \quad (3.3)$$

Equation 3.3 can be interpreted as follows: In each hour, assuming that there is a shift at the beginning of its 5th hour, it is optimal to extend the shift only if the number of patients waiting for an initial assessment is more than or equal to threshold B_{h,E^1,E^2,E^3} , and the shift should not be extended otherwise. If there is no shift that begins of its 5th hour at the start of hour $h_{\vec{s}}$, then there is no decision to be made. Moreover, if there are multiple shifts that begin their 5th hour at hour $h_{\vec{s}}$, then there will be different thresholds for each of the shifts. Having the control-limit type optimal policy enables us to solve the problem with more efficient algorithms such as the monotone policy iteration algorithm (Puterman, 2014). This is illustrated in Algorithm 1. It is much easier to implement these

Algorithm 1: The Monotone Policy Iteration Algorithm

Initially set the default optimal policy of all states to be 0 or $a_{\vec{s}}^{*0} = 0 \quad \forall \vec{s} \in S$;
 $counter = 0$;
The total number of states= NS ;
while $counter < NS$ **do**
 (Policy Evaluation) Using $a_{\vec{s}}^{*0}$ and equation 4.1, calculate the values of $v^*(\vec{s}) \quad \forall \vec{s} \in S$;
 for $h = 0, \dots, 23$ **do**
 for $E^1 = 0, \dots, L_{h-5}$ **do**
 for $E^2 = 0, \dots, L_{h-6}$ **do**
 for $E^3 = 0, \dots, L_{h-7}$ **do**
 for $W = 0, \dots, WUB$ **do**
 (Policy Improvement) Using the values of $v^*(\vec{s})$ from Policy Evaluation
 step and equation 3.2, calculate the values of $a_{(h,W,E^1,E^2,E^3)}^*$;
 if $a_{(h,W,E^1,E^2,E^3)}^* = L_{h-4}$ **then**
 $a_{\vec{s}}^* = L_{h-4}$ for states (h, i, E^1, E^2, E^3) where $W < i \leq WUB$;
 break the loop for W ;
 end
 end
 end
 end
 end
 end
 counter=0;
 for all states do
 if $a_{\vec{s}}^* = a_{\vec{s}}^{*0}$ **then**
 counter=**counter**+1;
 end
 Replace $a_{\vec{s}}^{*0}$ with $a_{\vec{s}}^*$;
 end
end

types of policies in an ED setting.

Numerical results show that if it is the optimal decision to extend a shift for state $\vec{s} = (h, W, E^1, E^2, E^3)$, then it is also the optimal decision to extend a shift for all other states with similar values of h , E^1 , E^2 , and E^3 , but a higher value of W . This can be explained as follows: if for a state $\vec{s} = (h, W, E^1, E^2, E^3)$ with at least two possible actions of a_1 and a_2 , the optimal decision a_2 which is greater than or equal to a_1 , then for the state $\vec{j} = (h, W + m, E^1, E^2, E^3)$ where m is an integer greater than or equal to 1, the optimal decision would be at least a_2 as well. As mentioned before, the possible set of actions for states \vec{j} and \vec{s} are the same as they are both at hour h of the day and the possible actions at any state only depend on h .

When the optimal policy is non decreasing and there are only two actions to take at each state, then a control limit policy will be optimal. In the shift extension policy of our MDP formulation, we make a decision at the beginning of each hour of the day and not for each shift. Therefore, there would be three possible situations at the beginning of each hour: First, there can be no shift that is about to start its 5th hour and we actually have no decision to make (such as hour 0 of the example schedule). Second, we can have exactly one shift that is about to start its 5th hour and we have only two possible actions to choose from: to extend that shift or not to extend that shift (such as hour 1 of the example schedule). Lastly, There can be some hours in which we have more than one shift that is about to start their 5th hour (such as hour 8 of the schedule in Table 3.1). In these hours, we have more than two choices: to extend none, one or both shifts. Based on numerical results, the optimal policy would be to extend one shift if the number of patients waiting for an initial assessment is more than a threshold B , and to extend both shifts if the number of patients waiting for an initial assessment is more than $B + b$, where b is an integer greater than or equal to zero. Therefore, the optimal policy can be translated to optimal policies for two decisions for the two shifts (i.e., shift 2 and 3 of the schedule in Table 3.1) that are at the beginning of their 5th hour at hour 8; Extend shift 2 if the number of patients waiting is more than B and extend shift 3 if the number of patients waiting is more than $B + b$. Therefore, for each decision that we make for each shift, we have two possible choices and since our optimal policy is monotone and non decreasing, it is of a control-limit type.

3.4 Numerical Results

In this section, we provide the results of the monotone policy iteration algorithm and the lookahead policy model for solving the shift extension problem using data from a Canadian ED. The data includes the hourly number of arrivals and the PPH of physicians at all hours of the shift for the period of 2017-2018 which covers total of more than 100,000 visits to the Foothills Medical Center in the city of Calgary. Furthermore, we use this data in an optimization model to find a daily schedule that minimizes the mismatch between number of patients waiting to be seen and the number of patients who can be seen by available physicians based on their PPH rate at any hour during a one month planning horizon. In the optimization model we assume that we have total of 15 shifts per day with length of 6 hours each. Table 3.2 shows the resulting schedule from the optimization model. We use the result of chapter one to optimally find a schedule with 15 6-hour shifts.

Shift	Hour of the day																							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								
8																								
9																								
10																								
11																								
12																								
13																								
14																								
15																								

Table 3.2. The Schedule resulting from the optimization model

Using the schedule in Table 3.2, we solve the MDP for $\lambda = 0.99$ to derive the optimal policy for different values of α . Figure 3.1, represents the optimal policy for α of 5 and 30. The horizontal axis shows the hours of the day at which there is at least one shift at the beginning of its 5th hour, so there is a decision to make, and the possible three previous actions for each of those hours. For example, at hour 0 we have one shift that has started 4 hours ago so we should decide about its extension. There are two options for values of E^1 and E^2 for states that are at the beginning of hour 0 of a day as we have two shifts that were at the beginning of their 5th hour, 1 and 2 hours ago. However, the value of E^3 will always be 0 for states at the beginning of hour 0 as no shift has been at the beginning of its 5th

hour at hour 21 of the previous day. Therefore, for hour 0, we have 4 different options for the possible values of the previous actions. Consequently, on the horizontal axis we have all possible states in which we should make a decision regarding extending a shift, with the values of W on the vertical axis. The circle and square marks show the minimum number of patients waiting for which the optimal decision is to extend a shift at different states for α of 5 and 30, respectively. For instance, at midnight, if shifts 12 and 13 have been extended 1 and 2 hours ago, respectively, and E^1 and E^2 are both equal to 1, then the optimal decision would be the last column of hour 0 on Figure 3.1, which is to extend shift 14 if the number of patients waiting is greater than or equal to 6 and 19 patients, for α of 5 and 30, respectively. For hour 17 of the day, based on the given schedule, we have two shifts to decide about their extensions. Hence, we have one mark for the optimal decision of each shift and the optimal decision is to extend one shift according to the lower mark and both shifts based on the upper(darker) mark. We can see that for higher values of α , the optimal decision is to extend for higher number of patients waiting as it is more costly to extend a shift than to have a patient waiting compared to lower values of α . Based on Figure 3.1, we can derive some more general policies. For example for $\alpha = 30$, the optimal policy is to always extend if the number of patients waiting is more than 26, and never extend if it is less than 16, no matter what is the state of the system. For the number of patients waiting in between 16 and 26, the optimal decision depends on the hour of the day and the previous actions and we would have different thresholds for the minimum number of patients to extend a shift for each state. We evaluate and compare the policies with different values of α and also policies derived from the lookahead model with different horizons using a simulation model presented in Section 3.4.1. In the lookahead model, instead of the cost to go function, we only look at the immediate costs of k periods ahead. For example, in the one period lookahead model we only consider the immediate cost while in the 4 period lookahead model, we look at the costs in the next four periods (hours).

3.4.1 Simulation

Our discrete event simulation model mirrors the intricate patient flow dynamics within a Calgary, Canada ED. Following the protocols typical of North American EDs, upon arrival, patients are swiftly assessed by a triage nurse who conducts a concise assessment and initiates an electronic record. Patients then await their turn in a designated waiting area, during which their electronic records are organized in a virtual queue, officially marking the onset of their waiting period. Upon a physician's availability,

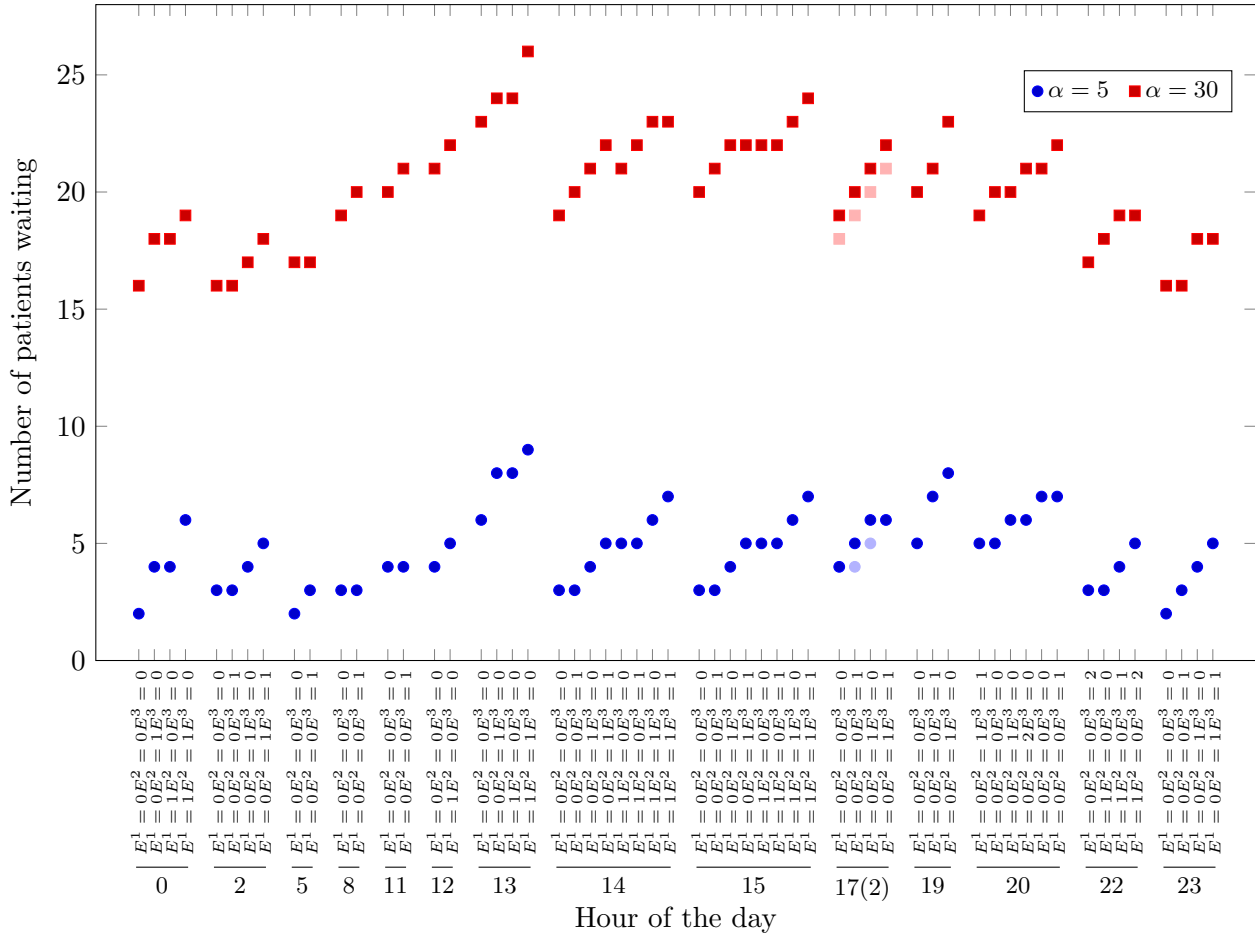


Figure 3.1. The minimum number of patients waiting to be seen for which the optimal policy is to extend a shift for different hours of the day and different values of the previous actions.

they select the next patient to attend to from this virtual queue. The patient’s initial consultation with the physician signifies the end of their waiting period.

The inputs to the simulation model includes: The arrival rate of patients which follows the Poisson process and has different averages for different hours of the day; The patient inter initial assessment time for different hours of the shift, which has exponential distribution and captures the PPH of physicians at each hour of the shift; The schedule of the ED; and the shift extension policy. The outputs of the simulation model are the waiting time of each patient and the total number of extensions during the simulation horizon. At the onset of hour zero, there are no patients awaiting service. Therefore, we incorporate a 2-day warm-up period before each replication of the simulation. We conduct 100 replications and calculate the average waiting time across all replications as our principal output metric. The simulation model is implemented using the Stochastic Simulation in Java

(SSJ) library (L'Ecuyer, 2016). The average waiting time and average number of extensions per day are the metrics for evaluating the policies. The simulation results are presented in Figures 3.2 and 3.3 and explained in the following paragraphs.

We first run the simulation for the schedule of Table 3.2 and optimal policies resulting from the monotone policy iteration algorithm with different values of α . The policy resulting from the monotone policy iteration algorithm, where the shift extension decision is made 2 hours before the end of the shift and there are 15 shifts in a day is called *PI-2h-15*. Figure 3.2, shows the changes in the average waiting time and the average number of extensions per day for different values of α for the model *PI-2h-15*. As it can be expected, as the value of α increases, the average number of extensions per day decreases and consequently the average waiting time increases. The average number of extensions per day can be from 0 to 15, where we always extend all shifts. This graph can be used to find the most suitable policy for the ED based on the target wait time or the available budget for number of shift extensions per day. For example, if the target wait time is 2 hours, then the corresponding α would be 60, for which the average daily number of extensions is almost 2 per day. On the other hand if we have allocated a fixed budget for shift extensions, for example 1 extensions per day, then the corresponding α will be 5, for which the average wait time will be more than 2 hours. Therefore, we can use this graph to know the proper value of α based on the budget and target wait time of the ED, without actually giving a numerical value to wait time of patients versus cost of extending a shift.

In order to better evaluate the effectiveness of having dynamic schedule, we compare it to static schedules, in which none to all shifts have 8 hours length. We solve the optimization model and allow one to all of the shifts of the schedule of Table 3.2 to have 8 hours length. As a result, we will have 16 schedules in which 0 to 15 of the shifts have fixed length of 8 hours and the rest have 6 hours length. We call these schedules *Fixed Schedules 15* with n number of shifts with 8 hours length. For PPH of the shifts with regular length of 6 hours and the shifts with 8 hours length in the *Fixed Schedules 15*, we use the empirical data for PPH of the 6-hour and 8-hour length shift. However, with a dynamic policy, physicians will get informed about the extension of their shifts at the beginning of hour 5 of the shift and presumably their PPH will be higher in hours 5th and 6th of the shift if it is going to be extended compared to when it will end in its regular length of 6 hours. If the shift is not extended, the PPH will be entirely from data for 6-hour length shift. However, if the shift is going to be extended, for the first 4 hours we use PPH rate of the first 4 hours of the 6-hour length shifts from the data and

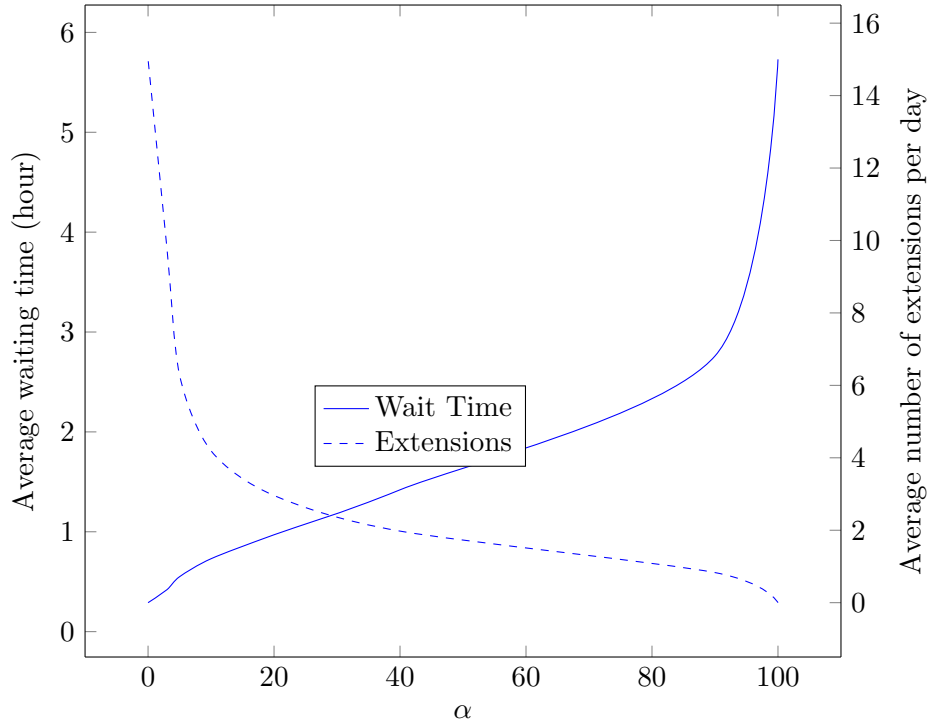


Figure 3.2. Simulation Results: Average waiting time and average number of extensions per day for different values of α for the model PI-2h-15

for the last 4 hours, we use PPH rate of the last 4 hours of the 8-hour length shift in the data. The hourly average PPH of 6 hour and 8 hour length shifts alongside the hourly average PPH used for the extended shifts is provided in Table 3.3.

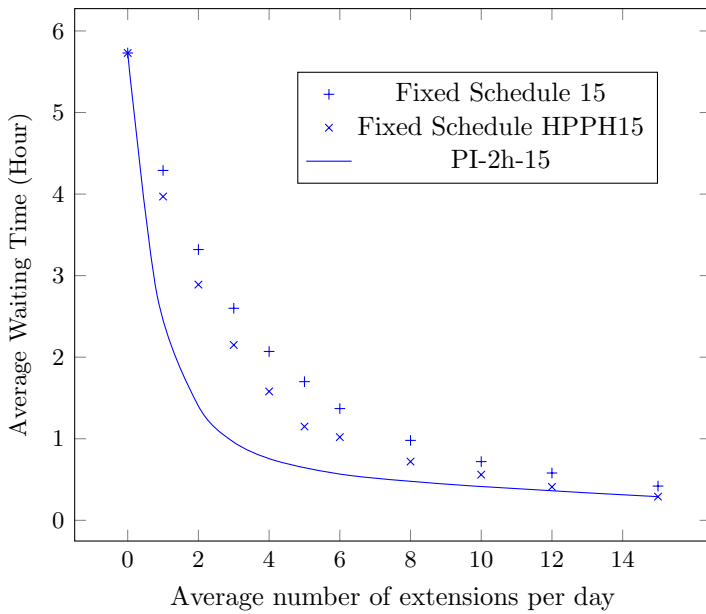
Based on Table 3.3, the PPH rate of the first 4 hours of the 6-hour length shifts is slightly higher than the PPH of the first 4 hours of a 8-hour length shift. Therefore, with dynamic policy, not only we can use the extra hours at the most advantageous times, but also, the average PPH of an extended shift is higher than the average PPH of a shift that has the original length of 8 hours. To separate the effect of increased average PPH from the effect of dynamically extending shifts on the improvements in the average waiting time, we also run the simulation for the *Fixed Schedules 15* with 0 to 15 8-hour shifts with the PPH rates used for the extended shifts (the 3rd row of the Table 3.3), instead of PPH of 8-hour shifts (the 2nd row of the Table 3.3). We call these model *Fixed Schedules HPPH15* (High PPH). In this way, the average PPH values would be the same for 8-hour shifts in the *Fixed Schedule HPPH15* and the extended shifts in the dynamic model, and any difference in the average waiting time of the model *PI-2h-15* and *Fixed Schedules HPPH15* will be only due to having dynamic shift extension policy and not because of the higher average PPH.

Shift Length	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	Hour 7	Hour 8	Average
6	3.48	2.39	1.79	1.52	1.24	1.31			1.95
8	3.17	2.45	1.83	1.57	1.45	1.39	1.01	0.53	1.67
8 (Extended)	3.48	2.39	1.79	1.52	1.45	1.39	1.01	0.53	1.70

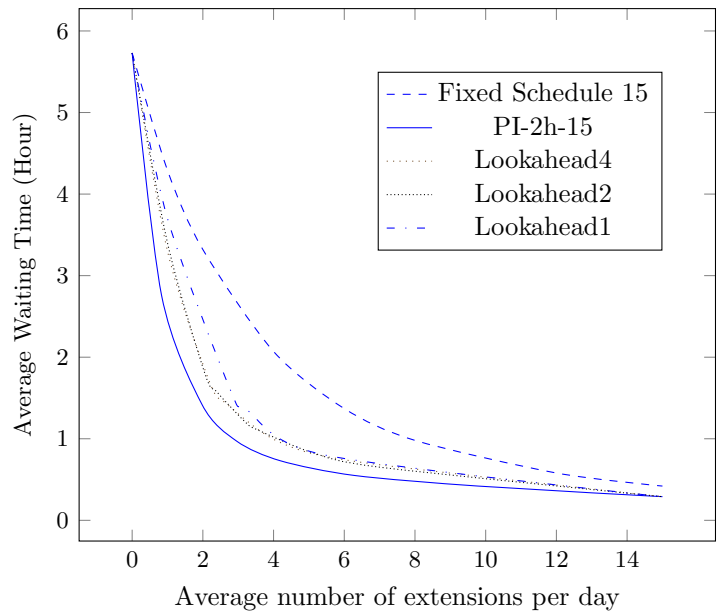
Table 3.3. The average PPH rate of shifts with lengths of 6 and hours and the average PPH used for the shift that is extended at the beginning of its 5th hour

The simulation results for comparing the dynamic and static schedules are presented in Figure 3.3. The Y-axis is the average waiting time while the X-axis is the average number of extensions per day for the dynamic schedule and the number of shifts with 8-hours length for the static schedules. In the first graph, Figure 3.3a, the average waiting time of dynamic schedule, *PI-2h-15*, *Fixed Schedules 15*, and *Fixed Schedule HPPH 15* are compared. Based on this figure, with having dynamic shift extensions, model *PI-2h-15*, it is possible to achieve the same average waiting time with fewer number of extensions per day compared to the *Fixed Schedules 15* with different number of 8-hour length shifts per day. It can also be said that with the same average number of extensions per day, having dynamic extension always outperform the *Fixed schedule 15* with the same number of 8-hour shifts per day in terms of average waiting time. This improvement is more considerable when the average number of extensions per day is smaller and as it increases, the performance of the dynamic extensions and the Fixed Schedule with 8-hour length shifts gets closer. The waiting time of the schedule with average of 3 dynamic shift extensions per day is almost 60 percent lower than the waiting time of the *Fixed Schedule 15* with 3 8-hour length shifts in a day. This should be noted that, this improvement is without addition of any resources and only by dynamically extending shifts and by better use of existing resources. Furthermore, we can see that by having the same PPH rates for the 8-hour length shifts in the *Fixed Schedule HPPH 15* and extended shifts of the model *PI-2h-15*, the average waiting time is the same when none or all shifts are always extended. For other average number of extensions per day, most part of the difference between the average wait time of the model *PI-2h-15* and the *Fixed Schedule 15* is due to having dynamic shift extensions rather than higher PPH. For example, when the average number of extensions per day is 3, 25% of the reduction in wait time is due to higher PPH and the rest is due to dynamically extending shifts.

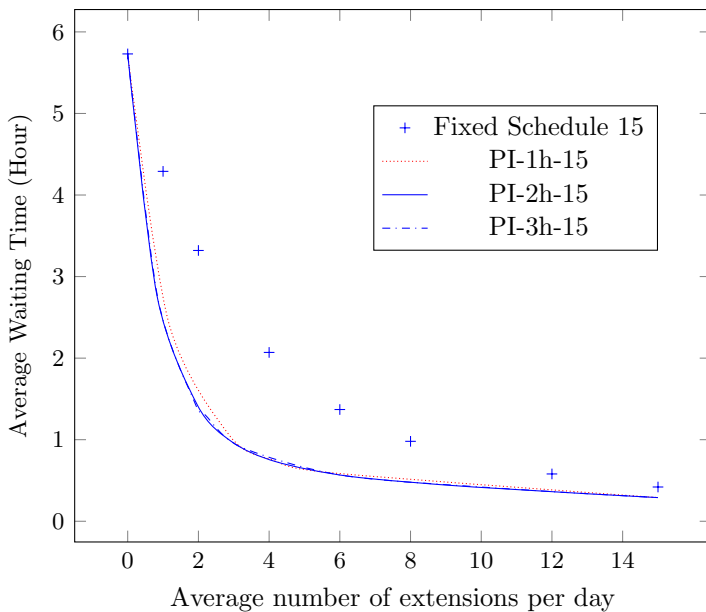
In Figure 3.3b, the waiting time and average number of extensions per day for policies derived from the 1, 2, and 4 step lookahead models, called *Lookahead1*, *Lookahead2*, and *Lookahead4*, respec-



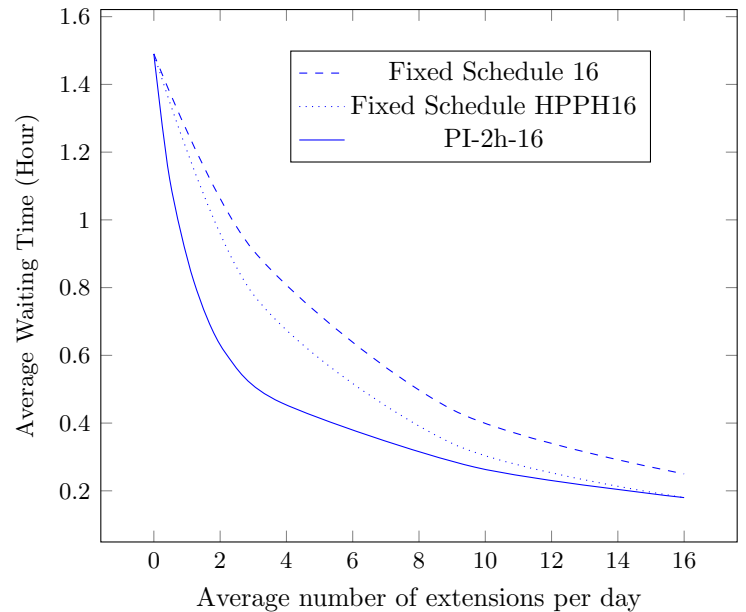
(a) Comparison of the dynamic model with the Fixed Schedule



(b) Comparison of the Policy iteration algorithm with the Lookahead models



(c) Comparison of the dynamic models where the decision is made 1, 2, and 3 hours before the end of the shift



(d) Comparison of the dynamic model with the Fixed Schedule with 16 shifts in a day

Figure 3.3. Simulation Results: The hourly average wait time of patients versus the average number of extensions per day for different schedules and policies

tively, for different values of α , are presented. In the lookahead models the number of steps is the number of hours in the future that is considered for the cost to go function while making the decision. It can be seen that using lookahead models can result in reduction of average wait time. However, this reduction is not as good as the improvement resulting from the policy from the policy iteration

algorithm. For example, for the average wait time of 2 hours, the average number of extensions per day is 1.8, 2.2, 2.21, and 2.8 for *PI-2h-15*, Lookahead4, Lookahead2, and Lookahead1 models. The policy from the iteration algorithm outperforms the policies from lookahead models as in policy iteration algorithm, we have an infinite horizon model while in the lookahead model we only consider one to 4 periods ahead. However, the structure of policies are similar in all models, Figure 3.4. The policies presented in Figure 3.4 result in the same average waiting time of 27 hours. Since the processing time of both algorithms is less than an hour (a couple of minutes for the Lookahead model and almost an hour for the monotone policy iteration algorithm), and the problem will be solved infrequently, it can be concluded that even though the improvement from using the policy iteration algorithm is small compared to the lookahead model, it is still considered to be the better approach for this problem.

In our model, the assumption is that the decision for extending a shift is made 2 hours before the end of the shift. In order to see how sensitive the results are to the timing of the decision, we resolve the MDP assuming that the decision is made 1 hour and 3 hours before the end of the shift. By making the decision later in the shift, we will have more information about the number of patients in the waiting room at the time that the extension will start. On the other hand, as discussed above, we assume that once the physician finds that his shift will be extended, his PPH will slightly increase as he knows that he will have more time in the ED and he can sign up for more new patients, instead of just wrapping up his existing patients. Therefore, by making the decision later, we lose that extra PPH. Examining this tradeoff by comparing the average waiting time of patients when the decision is made 1 and 3 hours before the end of the shift compared to the case when the decision is made 2 hours before the end of the shift, Figure 3.3c, we can conclude that even though the average waiting time can be marginally higher when the decision is made 1 hour before the end of the shift, the result is not very sensitive to the time of the decision.

We also test the effect of having dynamic shifts extension as oppose to having fixed schedules with different number of 8-hour shifts in a setting where we have more shifts in a day. In other words, we investigate the effect of dynamically extending shifts when the resources are less scarce. We assume that we have 16 shifts in a day, as oppose to 15 shift, and use the optimization model to find the schedule that minimizes the mismatch between the number of patients in the waiting room and available capacity. Similar to what we did with 15 shifts in a day, we find fixed schedules with total of 16 shifts in a day in which 0 to 16 shifts have 8-hour length instead of 6-hour length, *Fixed Schedule*

16. Then, we solve the MDP using the monotone policy iteration algorithm for different values of α for the optimal schedule with 16 6-hour length shifts and using the simulation model, compare the schedules based on their average waiting time. We assume that the decision is made 2 hours before the end of the shift and the model is called *PI-2h-16*. The results are presented in Figure 3.3d.

Similar to the case with 15 shifts in a day, we will most benefit from having dynamic shift extension if the average number of extensions per day is neither too few nor too many, 3.3d. For example, when this number is 3, the difference between the average waiting time of the model with dynamic shift extension and the Fixed Schedule with 6 8-hour length shifts is more than half an hour. However this difference approaches 0 when the average number of extension per day is close to zero or to 16. Moreover, with 16 shifts in a day, the peak of the difference between the dynamic model and the fixed schedule is more toward the left while it is more toward the right of the graph in the case with 15 shifts in a day, Figure 3.3a. The reason is that with 15 shifts in a day, the daily average capacity of the ED to visit new patients is ranging from 176 to 204 patients depending on number of extensions per day. However, this number is from 187.2 to 213.76 for the case with 16 shifts in a day. The average daily number of arrivals to the ED is 161 according to the data. Therefore, for the case with 15 shifts in a day, the capacity is not much more than demand and as a result the improvement in the average wait time is more, close to 2 hours, and the difference between the average waiting time of the models *Fixed Schedule 15* and *PI-2h-15* decreases as the average number of extension per day increases. Nevertheless, when the number of shifts per day is 16, the capacity is well beyond the average number of arrivals and as result, we see less improvement in the average wait time by using the dynamic policy.

Lastly, to have an easier to implement policy, we can use one threshold for each hour of the day, for which there is a shift that will end in 2 hours, instead of having different thresholds for each combination of values for the decisions made in the last three hours. For example, for hour 14 of the day for the schedule with 15 shifts in a day, Figure 3.1, we have 8 different possible combinations for values of E^1 , E^2 , and E^3 and therefore, 8 different thresholds. We can use the average of these 8 thresholds and call it the threshold for hour 14 of the day for the minimum number of patients in the waiting room to extend the shift. By using this average threshold, with the same average number of extensions per day, the average hourly waiting time is only 2% higher than the case where we have different thresholds depending on our previous actions. Moreover, we can just have 1 overall threshold,

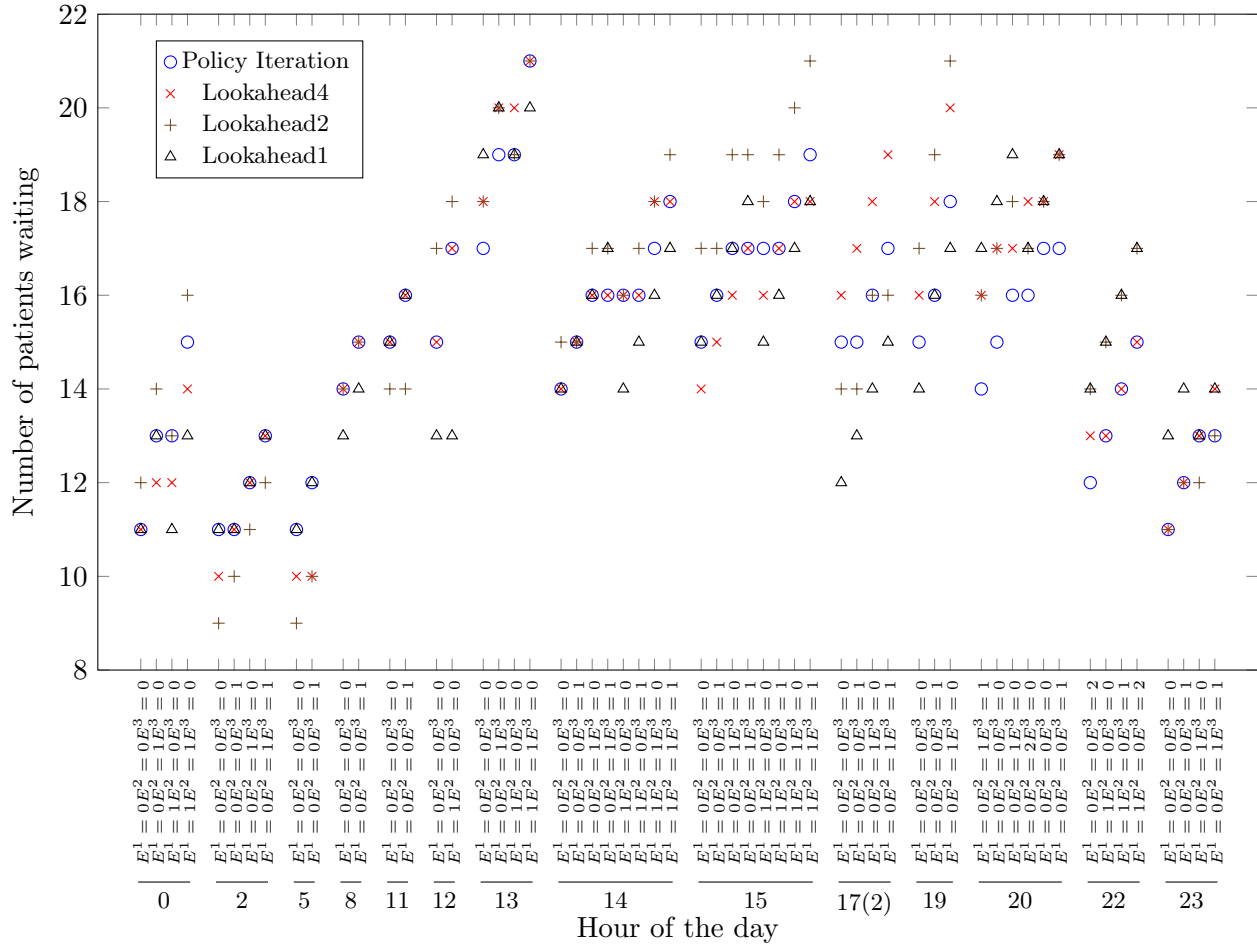


Figure 3.4. The minimum number of patients waiting to be seen for which the optimal policy is to extend a shift for different hours of the day and different values of the previous actions for the policies derived from the Policy Iteration algorithm and 4-step, 2-step, and 1-step lookahead models. All policies result in the average wait time of 0.7 hours

which is the average of all thresholds for all hours of the day and all values of previous actions. For example, we can say always extend a shift if the number of patients in the waiting room is more than 15 patients, no matter what is the hour of the day or what were our previous actions. In this case, the waiting time would be only 3% higher for the same average number of extensions per day, compared to having different thresholds for each combinations of $h, E^1, E^2,$ and E^3 . As a result, we can use this average value and have an easy to implement policy that can achieve a waiting time very close to the waiting time achieved by having different policies for each state.

3.5 Conclusion

Overcrowding is a pervasive challenge in EDs. Despite meticulous scheduling efforts, the inherent variability in patient arrivals and physician productivity often leads to prolonged wait times. Long wait times in EDs results in increased risk of medical errors, increased number of patients who leave the ED without being seen, and fatigue in physicians. To mitigate this issue, many EDs resort to overtime shifts, extending the scheduled duration of physician shifts. Particularly, whenever there is a surge in demand, or number of patients waiting to be seen by a physician, physicians are asked to stay longer after their shift is over. However, there exists no standardized protocol dictating when such extensions should occur and the decision is made subjectively.

The main goal of this study is to find an optimal policy on when to extend physicians' shifts. This study addresses this issue by formulating the problem as a Markov Decision Process and employing a policy iteration algorithm for resolution. The resulting optimal policy is a threshold policy and shows when to extend a shift based on the number of patients waiting to be seen, hour of the day, and the extension decisions made in the previous hours. We use simulation model to compare the case where we use the resulting policy and dynamically extend shifts to the case where we statically increase length of few or all shifts. Simulation results demonstrate the superiority of dynamic scheduling over static schedules as with dynamic schedules we can use the extra capacity at the most advantageous times. The resulting policy is of control limit type and based on the hour of the day and number of available physicians, the decision would always be to extend the shift if the number of patients waiting is more than a threshold.

We also compare the policy iteration algorithm with lookahead policies and even though the policies resulting from the lookahead models result in significant improvement in patients' average wait times, due to short implementation time, the policy iteration algorithm would still be the best approach. Moreover, the sensitivity analysis shows that the results are not very sensitive to the time of decision and whether it is made 1, 2, or 3 hours before the end of the shifts, the results would almost be the same.

This research, to the best of our knowledge, is the first study that investigates the problem of finding an optimal policy for extending shifts in EDs. In our formulation we consider the stochasticity

in number of arrivals and productivity of physicians. The number of arrivals to the ED follow a non homogeneous Poisson process, while for the PPH of physicians we use the empirical distribution. We also consider that the average PPH of physicians decreases during the shift. The resulting policy is straightforward to implement and can effectively reduce wait times in EDs. Depending on the budget and target wait time, ED managers can select a suitable policy to address overcrowding. At the start of each hour, they simply need to decide on shift extensions based on the number of waiting patients, which is easily accessible information. They do not need to solve an optimization problem or go through complicated algorithms and procedures. It can be even done automatically: An alarm system can be easily integrated into the ED system based on the policy and real-time waiting room census to issue a message to extend a shift.

One of the limitation of this study is that we only consider the availability of physicians and assume that the longer wait time resulting from a surge in demand is due to having not enough physicians on site. However, in many cases, bed availability or nurse shortage are the causing factor for longer wait time. It would be best if all such factors could be integrated in the state space to have a more realistic model for the ED. Nevertheless, adding more dimensions to the state space with their own stochasticities would make the optimality equations and transition probabilities more challenging.

Another idea for future studies is to integrate the scheduling and staffing problem from chapter two and the shift extension problem from current chapter. More particularly, if we find an optimal fixed schedule knowing that we can extend shifts when necessary, and simultaneously, find the optimal policy to extend shift, it can be expected that we can have more improvement in the average patients' wait time. Currently, we assume that the schedule is fixed and we find an optimal policy on when to extend shifts. But, it is possible that we find a different fixed schedule if we include the option of shift extension in the staffing and shift scheduling problem.

Bibliography

- Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L. and Aratow, M., 2016. Accurate emergency department wait time prediction, *Manufacturing & Service Operations Management* **18**(1): 141–156.
- Bandi, C. and Gupta, D., 2020. Operating room staffing and scheduling, *Manufacturing & Service*

- Operations Management* **22**(5): 958–974.
- Bard, J. F., Morton, D. P. and Wang, Y. M., 2007. Workforce planning at usps mail processing and distribution centers using stochastic optimization, *Annals of Operations Research* **155**(1): 51–78.
- Bürgy, R., Michon-Lacaze, H. and Desaulniers, G., 2019. Employee scheduling with short demand perturbations and extensible shifts, *Omega* **89**: 177–192.
- Campbell, G. M., 2012. On-call overtime for service workforce scheduling when demand is uncertain, *Decision Sciences* **43**(5): 817–850.
- Chan, D. C., 2018. The efficiency of slacking off: Evidence from the emergency department, *Econometrica* **86**(3): 997–1030.
- Cildoz, M., Ibarra, A. and Mallor, F., 2019. Accumulating priority queues versus pure priority queues for managing patients in emergency departments, *Operations Research for Health Care* **23**: 100224.
- Easton, F. F. and Goodale, J. C., 2005. Schedule recovery: Unplanned absences in service operations, *Decision Sciences* **36**(3): 459–488.
- Frank, J. R. and Ovens, H., 2002. Shiftwork and emergency medical practice, *Canadian Journal of Emergency Medicine* **4**(6): 421–428.
- Fry, M. J., Magazine, M. J. and Rao, U. S., 2006. Firefighter staffing including temporary absences and wastage, *Operations research* **54**(2): 353–365.
- Fügener, A. and Brunner, J. O., 2019. Planning for overtime: The value of shift extensions in physician scheduling, *INFORMS Journal on Computing* .
- Gans, N. and Zhou, Y.-P., 2002. Managing learning and turnover in employee staffing, *Operations Research* **50**(6): 991–1006.
- Kao, E. P. and Queyranne, M., 1985. Budgeting costs of nursing in a hospital, *Management Science* **31**(5): 608–621.
- Keyvanshokoo, E., Shi, C. and Van Oyen, M. P., 2021. Online advance scheduling with overtime: A primal-dual approach, *Manufacturing & Service Operations Management* **23**(1): 246–266.

- L'Ecuyer, P., 2016. SSJ: Stochastic simulation in Java, software library. <http://simul.iro.umontreal.ca/ssj/>.
- Liao, S., Koole, G., Van Delft, C. and Jouini, O., 2012. Staffing a call center with uncertain non-stationary arrival rate and flexibility, *OR spectrum* **34**(3): 691–721.
- Mehrotra, V., Ozlü, O. and Saltzman, R., 2010. Intelligent procedures for intra-day updating of call center agent schedules, *Production and operations management* **19**(3): 353–367.
- Parisio, A. and Jones, C. N., 2015. A two-stage stochastic programming approach to employee scheduling in retail outlets with uncertain demand, *Omega* **53**: 97–103.
- Patrick, J., Puterman, M. L. and Queyranne, M., 2008. Dynamic multipriority patient scheduling for a diagnostic resource, *Operations research* **56**(6): 1507–1525.
- Pinker, E. J. and Larson, R. C., 2003. Optimizing the use of contingent labor when demand is uncertain, *European Journal of Operational Research* **144**(1): 39–55.
- Powell, W. B., 2014. Clearing the jungle of stochastic optimization, *Bridging data and decisions*, Informs, pp. 109–137.
- Puterman, M. L., 2014. *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons.
- Qin, R., Nembhard, D. A. and Barnes II, W. L., 2015. Workforce flexibility in operations management, *Surveys in Operations Research and Management Science* **20**(1): 19–33.
- Silva, T. A. and de Souza, M. C., 2020. Surgical scheduling under uncertainty by approximate dynamic programming, *Omega* **95**: 102066.
- Wright, P. D. and Bretthauer, K. M., 2010. Strategies for addressing the nursing shortage: Coordinated decision making and workforce flexibility, *Decision Sciences* **41**(2): 373–401.
- Zhu, S., Fan, W., Liu, T., Yang, S. and Pardalos, P. M., 2020. Dynamic three-stage operating room scheduling considering patient waiting time and surgical overtime costs, *Journal of Combinatorial Optimization* **39**(1): 185–215.

Chapter 4

Enhancing Efficiency and Applicability of Surge Calls in EDs

4.1 Introduction

Overcrowding in EDs is an unpredictable, inevitable, and recurrent challenge. Overcrowding in ED is a situation in which the demand for health services exceeds the capacity of the ED to deliver quality care in a reasonable amount of time. ED overcrowding adversely affects the ability of the hospital to provide appropriate services, results in increased patient morbidity and mortality, and increases burnout among physicians. Overcrowding within EDs often arises from unpredictable events such as adverse weather conditions, mass casualty incidents, and disease outbreaks. The surge in demand is typically unforeseeable, necessitating immediate response and adaptation. Despite efforts to align capacity with anticipated demand, the inherent stochasticity in both patient influx and resource availability ensures that instances of extended wait times and a backlog of patients awaiting physician attention persist. A critical factor contributing to prolonged patient wait times is the insufficient availability of physicians within the ED. Given that physicians constitute one of the most costly resources in EDs, it is important to find an approach to efficiently increase the availability of physicians for mitigating overcrowding while maintaining fiscal responsibility.

Various strategies are employed across hospitals and EDs to address overcrowding and to make

efficient use of nurses and physicians. Having flexibility in the schedule has shown to help reduce costs at hospitals by reacting to changes and surge demand (Rath and Rajaram, 2022). The most common approaches for incorporating flexibility in schedule include overtime utilization, on-call physician rotations, and the implementation of surge calls. The Foothills Medical Center (FMC) ED in Calgary, Alberta, employs the latter strategy; however, its success has been hampered by specific challenges. More particularly, when there is a surge in demand in FMC ED, a surge call is issued to all physicians who are not currently on schedule. The department hopes that physicians can make adjustments to daily plans and can come to the ED as soon as possible. However, it has been shown that the implementation of this strategy has not been as successful as desired. Chief among these challenges is the absence of a clear and optimal policy dictating when a surge call should be issued. Currently this decision is made subjectively based on the status of the ED and the number of patients waiting to be seen. Additionally, despite financial incentives designed to encourage physicians involvement, the response rate to these calls are too low, resulting in sub optimal physician participation. While surge staffing offers some flexibility and the capacity to allocate additional resources, a notable gap exists in the absence of systematic guidelines for optimizing this strategy (Hu et al., 2021). It also worth mentioning that surge call strategy in this paper refers to flexible, high-cost capacity that is used to deal with short-term surges in demand that need only minor adjustments in staffing level and is not meant for major disaster response.

This research addresses identified shortcomings by focusing on two primary issues. Firstly, we tackle the absence of a well-defined policy for issuing surge calls by formulating the problem as a Markov Decision Process (MDP). Through this approach, we aim to establish an optimal policy for issuing surge calls that considers the dynamic and probabilistic nature of ED operations. The objective of the model is to minimize the weighted average of the number of patients waiting to be seen by a physician and the cost incurred by issuing surge calls and adding extra, more expensive shifts to the schedule. We encounter two sources of stochasticity in our model: one from the number of arrivals to the ED and one from the capacity of the system, measured by the Patient Per Hour (PPH) rate of physicians in the ED. Moreover, our model incorporates the likelihood that physicians may not respond to a surge call, a parameter intricately linked to the incentives and consequences associated with their response or non-response. Secondly, by manipulating this probability (probability of having a surge call responded by a physician) within our MDP framework, our research offers decision-makers a tool to discern the

most effective policy given the available incentives and constraints.

We evaluate the effectiveness of our policy using a simulation model. To ensure a fair comparison, we compare two cases with the same average number of shifts scheduled in a day: having n fixed shifts in a day versus having fewer than n fixed shifts in a day but compensating by adding the average number of surge calls issued in a day, resulting in an average of n shifts per day. The results show that the flexibility resulting from issuing surge calls, rather than solely adding fixed shifts to the schedule, can improve the average and 90th percentile wait time by almost 13 and 40 minutes, respectively. The improvement in wait time is even more pronounced when resources become scarcer and fewer fixed shifts are scheduled. However, as resources become less scarce and more fixed shifts are scheduled in a day, using surge calls instead of adding more fixed shifts can have a negative impact on ED wait times. This occurs because issuing a surge call is a strategy employed in response to overcrowding and is not an optimal scheduling approach. Consequently, when more physicians are available in a day and the probability of overcrowding is low, using surge calls instead of optimizing the schedule by adding more shifts can yield suboptimal results.

The remainder of this chapter is organized as follows. Section 4.2 provides an in-depth review of any related literature. Section 4.3 presents the Markov decision process problem formulation. In Section 4.4, we discuss our empirical findings, and in Section 4.5, we present simulation results and analysis of the implications. Finally, in Section 4.6, we conclude with a summary of key insights and contributions.

4.2 Literature Review

There are three main streams of research in the literature related to this study. The first stream focus on importance of considering the variable productivity of physicians in the staffing and scheduling of physicians, which is extensively discusses in Section 2.2.2 of this thesis. The second stream is about strategies employed in EDs to address overcrowding or surge demand. The other stream includes papers that are about finding the optimal policy about when and how to use surge capacity in different settings. In the next paragraphs, we study these two last streams and then we position our work in the literature and discuss how this study can fill the existing gaps.

The strategies to mitigate the overcrowding problem in ED can be categorized into two main groups. The first group of studies focuses on strategies to reduce and manage the demand and number of arrivals to the ED. This can be done by diverging the arrivals to other EDs (Xu and Chan, 2016), by delay announcement (Park et al., 2023), by different queueing configurations (Ferrand et al., 2018), by faster transferring admitted patients to the hospital and reducing length of stay (Konrad et al., 2013; Somanchi et al., 2022), or by discharging patients early (Mills et al., 2021). The second group, on the other hand, which is also the focus of this study, is about improving the capacity and more specific to our work, total available PPH of ED by somehow increasing the number of available physicians. This can be done by extending the shift length, calling in the on call physicians, or issuing surge calls. The first strategy, which is physician shift extensions, is the focus of chapter 3 of this thesis and the related literature is extensively reviewed in section 3.2. The latter strategy, surge calls, is the main focus of this chapter and this review.

calling in on-call physicians and issuing surge calls differ significantly to the shift extension approach in terms of timing and the amount of extra payment offered to physicians, as well as their response rates. Specifically, on-call physicians receive compensation for their availability regardless of whether they are needed in the ED, leading to a guaranteed response rate of 100%. Conversely, with the surge call strategy, physicians are not compensated for being on call, and their response is not mandatory when a surge call is issued, resulting in a generally lower response rate, despite the higher compensation compared to a regular shift. This review focuses primarily on papers examining the utilization of surge calls for staffing in emergency departments.

Grant et al. (2022) explore the optimal scheduling of appointments in scenarios with both standard and surge capacity. It focuses on dynamically allocating resources to meet fluctuating demand, particularly in service systems such as healthcare or emergency services. They find an optimal scheduling policy through a stochastic dynamic programming formulation. Our research diverges in its emphasis on capacity augmentation during periods of overcrowding, with a focus on real-time capacity enhancement. This stands in contrast to their focus on optimizing capacity allocation for appointments scheduled in advance or through offline processes. Due to recent government regulations in the US that hospitals need to improve emergency preparedness to manage surge capacity, Mills et al. (2021), propose coordinated early discharge of patients and inpatient workload soothing as a response to a surge in demand. Formulating the problem as an optimization model they show that coordinated

discharge and workload smoothing are complementary strategies to improve surge response.

Surge in demand occurs in manufacturing settings as well. In a study by Huang et al. (2016), the problem of supply chain planning in presence of demand surges is investigated. They define demand surges as random significant increases in demand in an otherwise relatively stable demand environment. They propose two possible sourcing strategies: reactive capacity and safety stock. Even though EDs are different in nature, there are some similarities between surge call and on call physicians vs reactive capacity and safety stock, in reactive capacity, they can contract with an outside firm and to keep safety stock for demand surges, a firm can set aside extra inventory in their own warehouse. With safety stock they have to pay for storage even if there is no surge in demand and that is similar to the idea of having on call physicians. With reactive capacity, there is a possibility of not finding an outside firm in a reasonable time, which is similar to the idea of surge calls in EDs. Huang et al. (2016), conclude that the best strategy depends on the magnitude and predictability of surge demand duration, intensity, compactness and frequency.

Positioning of this study: To the best of our knowledge, this research is one of the first studies that finds an optimal policy for issuing surge call in EDs by formulating the problem as a MDP to provide guidance to decision makers to effectively design and implement surge calls in EDs. The policy also indicate what should be the response rate to have an effective strategy.

4.3 Problem Formulation

The aim of this study is to formulate an optimized policy for dispatching surge calls to physicians within EDs based on real-time system conditions. In this paper we assume that there exist a schedule for physicians that represents the number of shifts in a day, their start time and their length. This schedule is denoted by L_h for $h = 0, \dots, 23$, where L_h represents the number of shifts commencing at hour h . The objective is to determine when to issue a surge call for physicians based on both the current patient load awaiting treatment and the overall capacity of the system, quantified as the total Patients Per Hour (PPH). The total PPH is contingent upon the number of available physicians and their respective shift hours, considering that physicians may exhibit varying patient throughput rates at different hours during their shifts. More particularly, to find the total PPH, we use the PPH

distribution of each available physicians which depends on the hour of their shift. Furthermore, we assume that after a surge call is initiated, physicians will arrive at the ED two hours later. This assumption is made as two hours would give enough time to physicians to commute to the ED and is short enough to be considered as a timely response to a surge in demand. However, in different settings and under different circumstances, we can changes this assumption. Consequently, the total PPH at any given hour is not solely determined by the present schedule, but also by decisions made 2 to 8 hours earlier, illustrating a dynamic interplay explained in detail within this section.

We formulate the problem as a discounted infinite horizon MDP model by providing the decision epochs, state space, action sets, transition probabilities, and costs below. We chose a discounted model, rather than an average reward model, with a discount factor close to one as it reflects the medium term planning horizon that is most applicable in the hospital and ED setting (Patrick et al., 2008). By discounting only slightly, we can make sure that costs are relatively similar in short term, while far distant costs are less valued. We use the schedule in Table 4.1 as an example to better present the MDP model. Based on this schedule, we have 10 shifts in a day with starting times ranging from midnight to 9 pm. The regular length of all shifts is 7 hours.

Shift	Hour of the day																							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	1	2	3	4	5	6	7																	
2					1	2	3	4	5	6	7													
3					1	2	3	4	5	6	7													
4										1	2	3	4	5	6	7								
5										1	2	3	4	5	6	7								
6											1	2	3	4	5	6	7							
7												1	2	3	4	5	6	7						
8													1	2	3	4	5	6	7					
9	6	7																	1	2	3	4	5	
10	4	5	6	7																		1	2	3

Table 4.1. Example Schedule

4.3.1 Decision Epochs

At the beginning of each hour during the day, the decision maker should decide if he needs to place a surge call based on the status of the system. We assume that the time required for a physician to respond to a surge call and arrive to the ED is two hours.

4.3.2 State Space

The decision to place a surge call and call in more physicians in ED is mostly dependent on the expected wait time of the patients as it is one of the most important performance indicators of the system. The expected wait time of patients in ED can be calculated using the general formula of dividing the workload to the processing time (Ang et al., 2016). Therefore, in our state space we need to capture these two factors in order to have the necessary and sufficient information to make the decision and compute the transition and cost functions. For workload, we can use the number of patients waiting in the ED and the expected number of arrivals in the next hour. The number of arrivals to the system in each hour is considered to be stochastic with variable averages during the day. For the processing time, we need to know the number of available physicians at any hour, the hour of their shift based on the given schedule, and if we have any extra physicians who have responded to surge call placed in the last hours. PPH of physicians also is stochastic with variable averages for different hours of the shift. For example, if we are at hour 1 of the day, based on the schedule in Figure 4.1, we have one physician at the 2nd hour of his shift, one at the 5th hour of his shift, and we may have one or more physicians working on surge shifts. Therefore, a state of the system, denoted by $\vec{s} \in S$, takes the following form:

$$\vec{s} = (h, W, E^1, E^2, E^3, E^4, E^5, E^6, E^7, E^8)$$

In this context, h denotes the hour of the day, which governs the available number of physicians, their shift hours, and the expected number of arrivals to the system during that hour. Each time unit is equivalent to one hour, starting from hour 0. Thus, in the subsequent state, the hour of the day would be $h + 1$. However, if h equals 23, the next state would transition to hour 0 of the following day.

W represents the number of patients waiting to be served at the beginning of hour h . The schedule of the ED determines how many physicians are available at hour h and what is the hour of their shift. However, in order to know the exact number of available physicians at each hour, we must also consider whether any physicians in the department have responded to a surge call within the last 8 hours. Binary variables E^1 to E^8 are set to one if, respectively, a surge call was placed and a physician responded to it 1 to 8 hours ago. For instance, if a surge call was placed 2 hours ago and a physician has responded positively, i.e. $E^2 = 1$, then an additional physician is available during

that hour. Similarly, if a shift was extended 8 hours ago, $E^8 = 1$, then the current hour marks the conclusion of the surge call that has started 6 hours ago. Utilizing these variables, we can compute the total Patient Per Hour (PPH) of the system, representing the total number of patients that can be attended to within hour h . Additionally, any unmet demand during a particular hour is assumed to carry over to the subsequent hour.

In order to have a finite state space, we assume an upper bound, WUB , on the number of patients waiting in the waiting room, W . The bound is set sufficiently high so it is of little practical significance. Moreover, we will see later that the optimal policy is of control-limit type and based on the parameters used in the model, there is a threshold for W , for each combination of h , E^2 to E^8 , that beyond that, the optimal policy is always to place a surge call. Therefore, the upper bound for W , should only be sufficiently higher than this threshold.

4.3.3 Action Sets

At the start of each hour, a decision must be made regarding whether to initiate a surge call and summon additional physicians. This decision is represented by the action taken in each state, denoted as $a_{\bar{s}}$, which is a binary variable equal to one if a surge call is initiated. However, despite the incentive of extra payment, it's common for physicians not to respond to surge calls which is another source of uncertainty in this problem. In other words, it is possible that after placing a surge call, no physician responds to the call. We call this probability R . Therefore, in our model, we incorporate a probability R representing the likelihood of physicians responding to a surge call. More particularly, when we make a decision to place a surge call $a_{\bar{s}}$, the probability that we can have one extra physician in 2 hours is R , which means that when $a_{\bar{s}} = 1$, in the next state E^1 will be one with probability of R and will be zero with probability of $1 - R$.

4.3.4 Transition Probabilities

After surge call decisions are finalized, the primary sources of uncertainty in transitioning to the next state of the system include three key factors. Firstly, it's the number of arrivals at hour h , referred to as N . Secondly, the total Patient Per Hour (PPH) of physicians on duty at hour h , denoted as T , which varies based on each physician's shift hour. Lastly, the probability of responding to a

call, represented by R . Therefore, the main determinants of transition probabilities is the probability of having $N - T$ patients added to or subtracted from the waiting room and the physicians' response rate, R , to the surge calls. Note that $W + N - T$ will be an integer number, greater than or equal to zero.

With current state being $\vec{s} = (h, W, E^1, E^2, E^3, E^4, E^5, E^6, E^7, E^8)$, we will go to state:

$$\vec{j}^1 = (h + 1, W + N - T, a_{\vec{s}}, E^1, E^2, E^3, E^4, E^5, E^6, E^7) \text{ with probability } p(\vec{j}^1|\vec{s}) = R.p(N - T|\vec{s}),$$

and will go to state:

$$\vec{j}^2 = (h + 1, W + N - T, 0, E^1, E^2, E^3, E^4, E^5, E^6, E^7) \text{ with probability } p(\vec{j}^2|\vec{s}) = (1 - R).p(N - T|\vec{s}) .$$

The system's arrival rate follows a Poisson process with varying hourly averages throughout the day. To model the Patient Per Hour (PPH) of physicians, we utilize empirical probability distributions derived from data, accounting for fluctuating hourly averages during shifts. The total PPH of the system in any given hour depends on the original schedule and the decisions made in preceding hours. Consequently, for each hour, we establish a base schedule or total PPH, and based on decisions from the past 8 hours, there are 2^7 potential total PPH scenarios. More particularly, based on the fixed schedule we will find the minimum total PPH and based on values of E^2 to E^8 , we might add more to it. Since E^2 to E^8 are binary variables, in total we will have 2^7 possible combinations. Note that if E^2 is one it means that we have one extra physicians (due to the surge call) who is at hour one of their shift, while if E^8 is one, it means that we have one extra physician at hour 7 of the shift. It should also be mentioned that E^1 will not affect current physician availability as the extra physician will start their shift in the subsequent hour.

Additionally, within each scenario, we must ascertain the probability of physicians seeing anywhere from 0 to a maximum of 20 patients. For example, assuming the schedule of Table 4.1, if we are at hour 0 of the day, then there is one physician at the first hour of the shift, one physician at the 6th hour of the shifts and one physician at the 4th hour of the shift. Depending on our previous decisions we might have 0 to 7 more physicians at first to seventh hour of their (surge) shift as well. For each of these available physicians, depending on the hour of their shift, we have a probability distribution for the number of patients they might seen in that hour. More specifically, the physician who is at the first hour of their shift, might see 0 to 7 patients with their corresponding probabilities. We have the same case for all other available physicians and we need to iterate over all combinations of possible

PPH values for all available physicians. Due to the computationally intensive nature of iterating over all possible PPHs and patient counts for each hour, we employ a more efficient approach. Instead of recalculating probabilities for each possible PPH in every hour, we initially identify all potential total PPH values for a day, including the possible additional PPH from physicians on surge shifts. Subsequently, we construct a probability matrix based on these potential total PPH values, leading to expedited processing.

Given the anticipated total PPH and the expected number of arrivals at hour h , we can predict the number of patients who will be seen and discharged or added to the waiting room during that hour. Furthermore, when a surge call is initiated, we can assess the probability R of transitioning to the next state with $E^1 = 1$, and the probability of $1 - R$ transitioning to the next state with $E^1 = 0$. This estimation of patient flow enables us to determine the subsequent state for the following hour.

4.3.5 Costs

The immediate cost associated with each state is determined by comparing the expense of adding a surge call against the cost of one hour of patients' wait time. The aim is to uphold reasonable wait times or minimize the number of patients waiting while managing costs effectively. We express these immediate costs as:

$$c(\vec{s}) = W + \alpha E^2$$

It's important to highlight that the cost associated with each state is solely contingent upon the state itself, independent of any actions or decisions made within that state. This can be explained by the fact that the decision-making process does not inherently incur costs or alter the current PPH rate of the ED. Costs are only incurred and effectiveness is realized when a physician responds to the decision. The realization of these costs occurs two hours after the decision is made, denoted by E^2 , which equals 1 if a surge call issued two hours prior is answered by a physician, initiating the surge call at the current hour.

The objective is to minimize the above immediate cost function plus the cost to go function for the future state, which strikes a balance between the inconvenience to patients due to wait times and the financial impact on the ED of resorting to surge calls. The multiplier, α , represents the relative

cost of extending a shift compared to patient wait time. A lower α value prioritizes patient wait time, resulting in a policy that leans towards more frequent use of surge calls. Conversely, higher α values lead to an optimal policy with fewer surge calls and consequently longer wait times for patients. Our simulation model demonstrates that the average wait time and the frequency of surge calls per day are influenced by the α value corresponding to the optimal policy. Thus, the choice of α , and consequently the optimal policy, can be tailored based on the desired wait time for the system or the available budget for the average number of surge calls per day.

4.3.6 Optimality Equations and Solution Approaches

The optimal value function in our formulation, denoted by $v^*(\vec{s})$, corresponds to the minimum discounted cost over the infinite horizon for each state and satisfies the following optimality equations for all states:

$$v^*(\vec{s}) = \min_{a_{\vec{s}} \in A_{\vec{s}}} \{W + \alpha E^2 + \lambda \sum_{N-T=LB}^{UB} R * p(N-T|\vec{s}) v^*(h+1, W+N-T, a_{\vec{s}}, E^1, E^2, E^3, E^4, E^5, E^6, E^7) + (1-R) * p(N-T|\vec{s}) v^*(h+1, W+N-T, 0, E^1, E^2, E^3, E^4, E^5, E^6, E^7)\} \quad \forall \vec{s} \in S \quad (4.1)$$

where λ is the discount rate, $p(N-T|\vec{s})$ is the probability that the number of arrivals to the system minus the number of patients seen by the available physicians at hour h is equal to $N-T$ and R is the probability of a physician responding to a surge call. Based on the data, we can find a lower bound, LB , and an upper bound, UB , for the value of $N-T$. We solve this problem using the value iteration algorithm. In the subsequent section, we present the numerical results demonstrating that the optimal policy exhibits a monotone non-decreasing property.

4.4 Numerical Results

This section presents the numerical results of the MDP model presented in previous sections. To solve this problem and find the parameters for the distributions of number of arrivals to the ED and the PPH of physicians, we use patient level data from Foothills Medical Center (FMC) emergency department in Calgary, Canada. Our data covers approximately 100,000 visits to the ED for the period of Jan 2017 to Dec 2018. Using these parameters, we solve an optimization problem to find an

optimal schedule in the ED with 14 7-hour length shifts. In the optimization problem, the objective is to minimize the number of patients waiting while considering the stochasticity in arrivals and PPH of physicians. For more details on the optimization problem, please see the result and proposed staffing and scheduling optimization model in Chapter 2. Average number of arrivals follows a Poisson distribution with varying averages in different hours of the day. PPH of physicians is also stochastic with variable averages during a shift. Table 4.2 represents the resulting schedule.

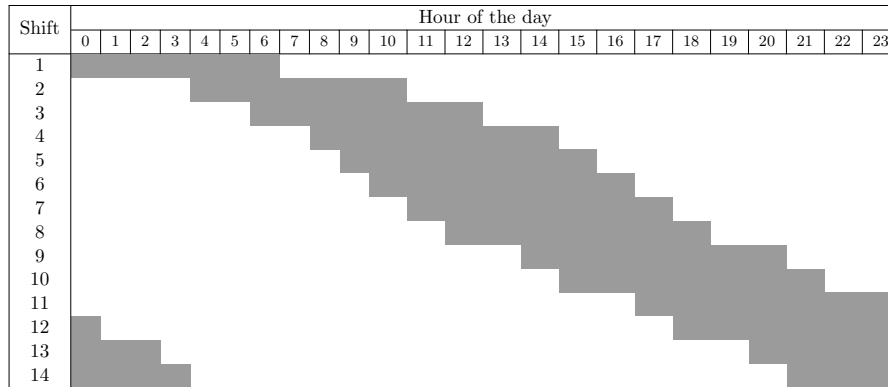


Table 4.2. The optimal fixed schedule with 14 shifts, each 7 hours length

The average hourly arrivals to the ED is presented in Figure 2.1. According to this graph, the peak number of arrivals occurs shortly before noon, with the early morning hours showing the lowest average arrivals.

As mentioned before, the number of new patients a physician can visit in each hour of the shift has a decreasing pattern. Based on the time stamp of initial assessment of each patient by each physician in our dataset, we can calculate the total number of patients seen in each shift. The total number of patients seen in each shift can be as low as 5 patients and as high as 19 patients and for more than half of the shifts, this number is between 9 and 12 patients. However, PPH is closely related to the hour of the shift; the average PPH in the first hour of the shift is 3.45 while for the last hour of the shift, this number decreases to 0.36. The probability distribution of PPH for different hours of the shift is presented in Figure 4.1.

Using the aforementioned parameters, we apply the discounted value iteration algorithm to explore various combinations of α and R . Figure 4.2a illustrates the optimal policy for hour 0 (midnight) scenario, where $\alpha = 30$ (Relative cost of issuing a surge call to costs associated with one hour waiting of one patient) and $R = 0.3$ (Response rate of physicians). Analyzing this graph reveals distinct decision-

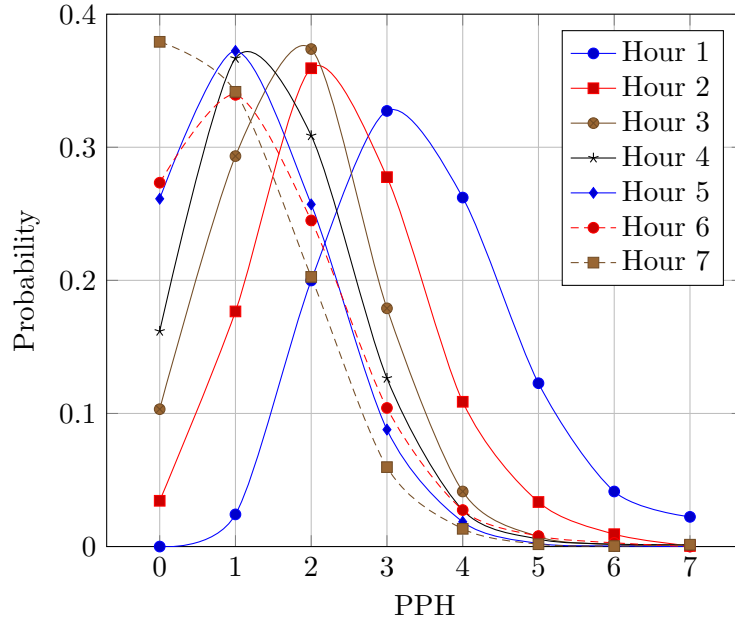
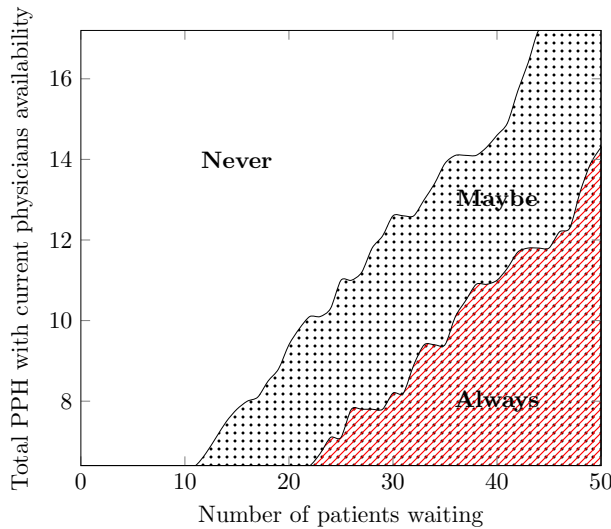


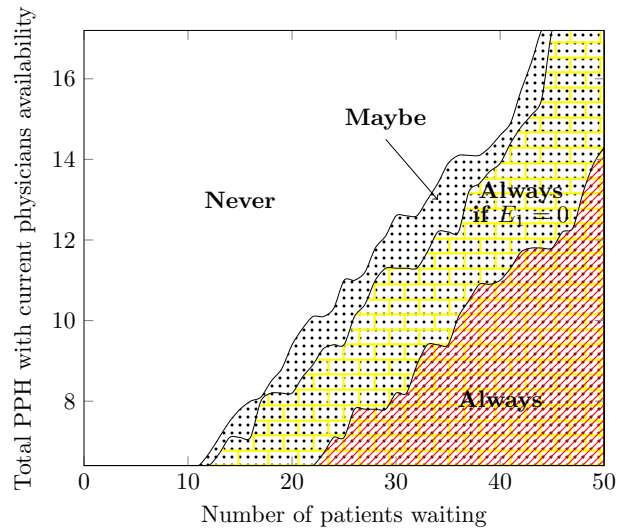
Figure 4.1. PPH distribution for different hours of the shift

making patterns: in the bottom right corner, where the number of patients waiting (D) is high and the total PPH is low, a surge call should always be initiated, irrespective of past decisions. Conversely, in the top left corner, characterized by low D and high PPH, surge calls are never warranted. However, the intermediate zones require consideration of past decisions over the previous 8 hours. Figure 4.2b demonstrates that decisions in the "Maybe" region primarily hinge on the decision made one hour prior. This graph shows the optimal policy for midnight, where $\alpha = 30$ and $R = 0.3$, same as Figure 4.2a. However, the difference is that we assume that the decision made one hour ago is to not to place a surge call, $E^1 = 0$. It can be observed that significant part of "Maybe" area in Figure 4.2a is now part of "Always" area in Figure 4.2b. This observation can be attributed to several factors. Firstly, decisions made between two to eight hours ago (E^2 to E^8) influence the current PPH, thereby impacting the current decision through the total available PPH. Conversely, the decision made in the preceding hour affects the total available PPH for the next 7 hours but does not directly influence the current PPH. Secondly, E^1 significantly influences the subsequent state two hours ahead, particularly if an additional physician is scheduled for a surge call, resulting in a substantial increase in total available PPH.

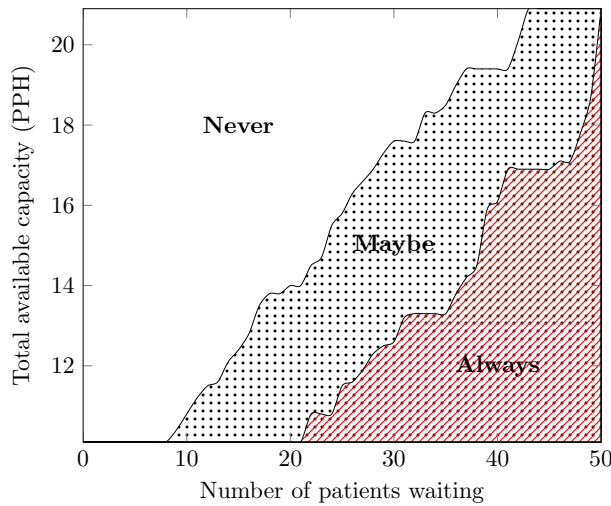
Hour of the day, the first element of the state space, determines the average number of arrivals to the ED and the available number of physicians according to the schedule. Therefore, we have different



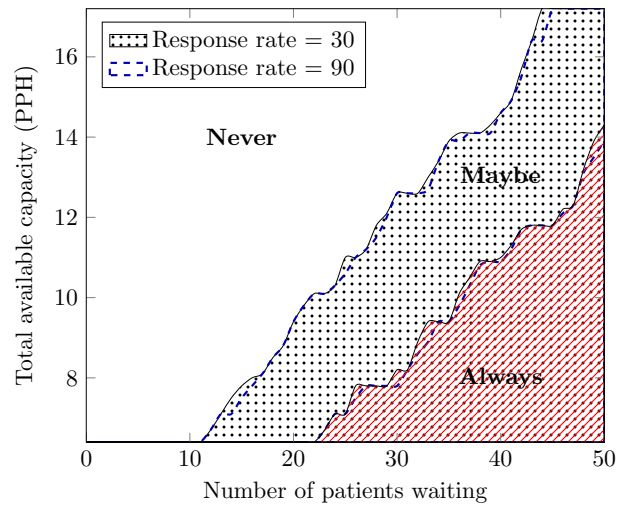
(a) Optimal strategy for issuing surge calls when Alpha=30, R=0.3, at h=0



(b) Optimal strategy for issuing surge calls when Alpha=30, R=0.3, at h=0, and $E^1 = 0$



(c) Optimal strategy for issuing surge calls when Alpha=30, R=0.3, at h=12



(d) Comparing the optimal strategy for issuing surge calls when R=0.9 vs R=0.3, and Alpha=30 at h=0

Figure 4.2. Optimal Policy for issuing a surge call based on the number of patients awaiting assessment and the total Patient Per Hour capacity of current physicians. The “Never” region indicates instances where a surge call should not be issued, while the “Always” region indicates when it should be. In the “Maybe” region, the decision to issue a surge call depends on the number of successful surge calls made within the past 8 hours.

policies for each hour of the day. The optimal policy for noon, as an example, is presented in Figure 4.2c to be compared with the optimal policy at midnight presented in Figure 4.2a . Total available capacity ranges from 6.4 to 17.2 for midnight and 10.1 to 20.9 for noon, as the expected number of arrivals is higher at noon and afternoon hours. Comparing the surge policies during peak (noon) and off-peak (midnight) hours, it becomes evident that the threshold for initiating a surge call is notably lower during peak times. For instance, when there are 30 patients waiting, the system won't trigger a surge call if the total available patients per hour (PPH) exceeds 12.6 at midnight. However, during peak hours around noon, a surge call may be warranted even with a PPH as high as 17.6. If we allow a backlog of patients to accumulate during peak arrival hours without addressing the overcrowding issue, the situation will rapidly deteriorate. Unattended demand will carry over to subsequent hours, exacerbating wait times as new patients continue to arrive. Conversely, during early hours when patient volume is anticipated to be lower, our threshold for extending shifts is higher. We presume that any excess demand can be managed in the subsequent hours, thereby justifying a more cautious approach to addressing overcrowding.

To conduct a sensitivity analysis and gauge the impact of the probability of responding to a surge call, denoted as R , we examine and contrast the optimal policy during midnight for two scenarios (both have an α of 30): when $R = 30\%$ and when $R = 90\%$ (referenced as Figure 4.2d). Through this comparative analysis, we observe slight adjustments in the policy. Specifically, with a higher R indicating a greater likelihood of physicians responding to the call, the region characterized by "Always" contracts marginally, while the "Never" region expands slightly. This suggests a reduction in surge calls compared to instances where R is lower. When response rate is low, in order to increase capacity we need to issue more surge calls. However, for higher values of response rate, when we issue a surge call it is more likely that we actually increase the capacity and as a result we need to issue fewer surge calls.

4.5 Simulation

Our discrete event simulation model is tailored to reflect the patient flow dynamics within an ED in Calgary, Canada. Aligning with the standard procedures observed in North American EDs, upon arrival, patients are promptly attended to by a triage nurse who conducts a brief assessment and

initiates an electronic record. Subsequently, patients await their turn in a waiting area, while their electronic records are systematically organized in a virtual queue, marking the commencement of their wait time. When a physician becomes available, they select the next patient to attend to from this virtual queue. The patient's initial consultation with the physician signifies the conclusion of their waiting period.

Our simulation model generates patient arrivals following a non-homogeneous Poisson process, with hourly rates derived from the average influx of patients observed during each hour throughout a comprehensive 2-year study period. Patients then enter a single-station queue, with the number of available physicians varying over time according to predetermined schedules. At any given hour, the staffing level and shift hours of physicians are predetermined. Once a physician concludes treatment for a patient, they are immediately available to attend to the next patient in line.

We simplify the model by assuming patients to be similar, thus omitting considerations for patient prioritization, as this does not affect our primary performance metric: the average waiting time for patients to consult with a physician in the ED. Consequently, patients are attended to on a First Come First Serve basis, with no distinction made based on urgency. We factor in the hourly variation in physician productivity, accounting for differences in patient throughput based on the hour of the shift and its duration.

To model the Inter Initial Assessment Time (IIAT) for patients under each physician's care during a shift, we employ a two-step process. Initially, we generate a random number based on empirical data, representing the total number of patients attended to during the shift. Subsequently, we fit separate Weibull distributions for the IIAT of consecutive patients seen within the same shift. This distribution is selected for its superior fit, determined through maximum likelihood estimation, among other theoretical distributions. Surge calls are initiated according to the policy derived from the MDP model, which is determined by the values of α and the response rate. Once a surge call is triggered, a random number is generated, with a probability equal to the response rate yielding 1 and 0 otherwise. If the generated number is 1, a new shift is scheduled to commence in 2 hours; otherwise, no additional shift is added.

The primary focus of our simulation model is to ascertain the average wait time experienced by all patients over a 10-week period of simulation. At the onset of hour zero, there are no patients

awaiting service. Therefore, we incorporate a 2-day warm-up period before each replication of the simulation. We conduct 100 replications and calculate the average waiting time across all replications as our principal output metric. The simulation model is implemented using the Stochastic Simulation in Java (SSJ) library (L'Ecuyer, 2016). For more details on the simulation model please see Section 2.5.2.

Moreover, within the ED, a subset of patients, typically those with non-urgent needs, may opt to depart without receiving medical attention if their wait surpasses their threshold of tolerance. The data lacks precise timestamps for their departure, with only the final point of contact noted, encompassing instances such as triage, self-declaration of departure, or missed appointments with physicians. Despite this, the rate of patients who Leave Without Being Seen (LWBS) can be approximated by comparing the number of patients triaged by nurses to those assessed by physicians.

Scholarly literature commonly models this patience threshold within queues using an exponential distribution, a methodology also adopted within this study. In our simulation model, mirroring the scheduling protocol of the FMC ED, we establish an average patience duration for patients, aligning the LWBS rate with observed trends in the FMC ED. This average patience duration serves as a benchmark for evaluating alternative scheduling strategies.

In our simulation, upon a patient's arrival at the ED, we simulate their patience threshold by drawing from an exponential distribution with an average equal to the established patient threshold. Subsequently, we schedule their departure based on this generated value. Should a patient remain unassessed by a physician by their scheduled departure time, they are recorded as having left without receiving medical attention. This process enables us to quantify the total number of LWBS cases over a simulated 10-week period for each scheduling strategy.

Utilizing our simulation model, we conduct a comparative analysis of patient waiting times, focusing on both average and 90th percentile metrics, across two scheduling scenarios. The first scenario excludes surge calls, while the second integrates fewer fixed shifts alongside surge calls. Additionally, we explore various policy configurations, varying the parameters α and response rate.

Figure 4.3 illustrates the simulation results across different response rates (10, 20, 30, 50, and 100), all with 14 fixed shifts. Depending on the policy, surge calls are strategically placed, augmenting the effective number of shifts per day. Each plotted line corresponds to a distinct response rate, with

the effective number of shifts per day increasing as α decreases, reflecting the diminishing cost of surge calls relative to patient wait times.

As α increases to 400, the cost of surge calls escalates, rendering them impractical under the optimal policy. Consequently, all scenarios result in a configuration without any surge calls, maintaining a fixed daily shift count of 14. Notably, at an effective number of shifts per day of 15, the 90th percentile wait time of all policies with different response rates outperform the 90th percentile wait time of having no surge calls. This underscores the efficacy of introducing flexibility into resource-constrained systems, notably improving the 90th percentile waiting time.

However, as α decreases further, signifying cheaper and more abundant resources, the utility of surge calls diminishes. Optimal scheduling tends to prioritize additional fixed shifts over surge calls, particularly when response rates fall below 30%. At response rates of 20% and 10%, optimal scheduling favors the inclusion of another fixed shift rather than resorting to surge calls. Eventually, as resources become increasingly abundant, solely augmenting the shift schedule proves superior to implementing surge call policies. This can be explained by the fact that while optimization models yield schedules tailored to stochastic parameters over an extended period, surge call policies primarily address short-term overcrowding issues and do not inherently align with the pursuit of daily schedule optimality.

It's worth noting that α can only decrease to 0. Consequently, in Figure 4.3, the line representing a Response rate of 10 concludes at an effective number of shifts of 16.4. When α reaches 0, surge calls incur no cost for the ED, prompting the issuance of a surge call every hour. However, given the 10% response rate, only an average of 2.4 of these surge calls will be answered, leading to an effective number of shifts of 16.4.

4.6 Conclusion

In the dynamic environment of EDs, characterized by the stochastic fluctuations in both patient demand and available capacity, optimal resource utilization poses a significant challenge. Despite meticulous scheduling efforts, instances inevitably arise where demand exceeds expectations, leading to increased patient waiting times and potentially compromising care quality. When faced with over-

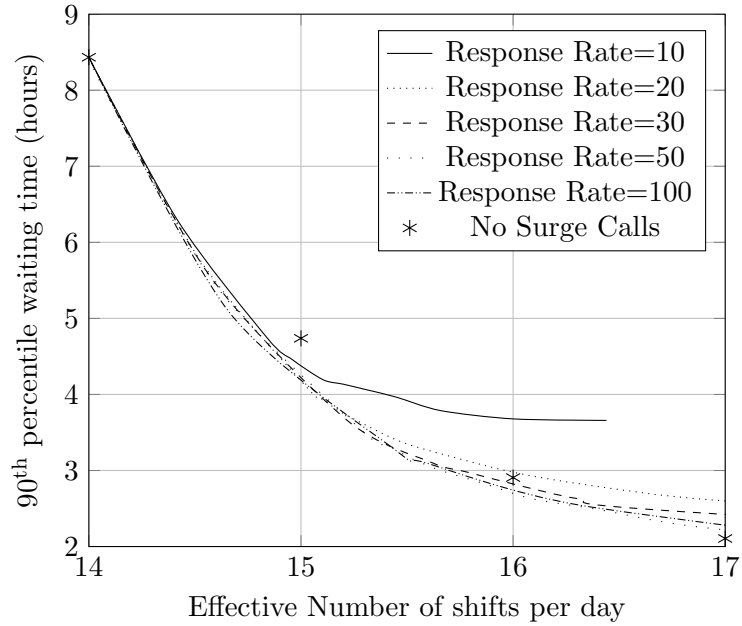


Figure 4.3. 90th percentile Wait time comparison for different surge call policies

crowding, a common recourse is to initiate surge calls, asking physicians to take on an extra shift in the ED when they are not originally scheduled a shift on that day. However, despite offering compensatory incentives, the response rate among physicians to these calls often falls short of expectations.

This study aims to address the inefficiencies in surge call implementations in the ED by formulating the problem as a Markov Decision Process (MDP) and finding an optimal policy for their issuance. Key considerations include the time of day, current ED scheduling, response rates of physicians, and the relative costs associated with implementing surge calls versus the expenses incurred by prolonging patients wait times. More particularly, based on the hour of the day and surge call decisions in previous hours, we provide an optimal policy which is a control limit type and depends on the number of patients waiting to be seen by a physician.

Central to this investigation is the question of response rate: what should be the minimum response rate to make this strategy effective? we answer this question by using a simulation model and show that depending on level of scarcity of resources (number of fixed shifts in a day in ED) the minimum required response rate would change. For example, when there are 16 shifts scheduled in a day, as long as the response rate is more than 20%, we can see some improvements in the average and 90th percentile wait time of patients. However, when there are only 15 fixed shifts scheduled in a day, i.e. resources are more scarce, the flexibility resulting from surge call will result in improvement in

wait time of patients even when response rate is as low as 10%. On the other hand, when there are 17 fixed shifts in the schedule, then surge calls will not be effective even for higher response rates. This is due the fact that, surge calls are responses to overcrowding in ED and is not an optimal schedule. Therefore, when we have enough fixed shifts in a day, it is better to schedule them using an optimization model as there is less Chance that overcrowding happens in ED. In sum, simulation findings indicate that under conditions of limited daily shifts and heightened resource scarcity, the implementation of surge call protocols can yield significant improvements in patient outcomes. For instance, in scenarios where only a modest number of shifts are scheduled per day, specifically 15 shifts, and resources are at a premium, the adoption of surge calls has been observed to reduce the 90th percentile waiting time by a noteworthy 32 minutes compared to static scheduling models. However, in instances where the ED is configured with a greater number of daily shifts, such as 17 shifts, the introduction of surge calls may exacerbate uncertainty and inadvertently prolong patient waiting times, thus yielding counterproductive outcomes. So, overall surge calls will be more effective when response rate is high and we have scarce resources.

By shedding light on the intricate relationship between surge call implementation, scheduling dynamics, and resource availability, this study seeks to equip ED administrators and policymakers with practical insights to streamline patient flow and improve care delivery in the face of the unpredictable challenges inherent in emergency healthcare provision.

We conclude with some possible extensions and discussing the limitations of this study. for future studies, we can relax the assumption of having a fixed schedule and instead find a fixed schedule with considering the option of surge calls. Also, for now, we are assuming the same response rate for all hours of the day. However, it is possible that the response rate is different at nights or on weekends. It is alos possible to use a joint strategy where we have an on call physicians on hours that the probability of issuing surge call is higher and secure a response rate of 100% for those hours and for other hours, rely on surge call and avoid paying physicians to be on call.

Bibliography

- Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L. and Aratow, M., 2016. Accurate emergency department wait time prediction, *Manufacturing & Service Operations Management* **18**(1): 141–156.
- Ferrand, Y. B., Magazine, M. J., Rao, U. S. and Glass, T. F., 2018. Managing responsiveness in the emergency department: Comparing dynamic priority queue with fast track, *Journal of Operations Management* **58**: 15–26.
- Grant, B., Gurvich, I., Mutharasan, R. K. and Van Mieghem, J. A., 2022. Optimal dynamic appointment scheduling of base and surge capacity, *Manufacturing & Service Operations Management* **24**(1): 59–76.
- Hu, Y., Chan, C. W. and Dong, J., 2021. Prediction-driven surge planning with application in the emergency department, *Submitted to Management Science* .
- Huang, L., Song, J.-S. and Tong, J., 2016. Supply chain planning for random demand surges: Reactive capacity and safety stock, *Manufacturing & Service Operations Management* **18**(4): 509–524.
- Konrad, R., DeSotto, K., Grocela, A., McAuley, P., Wang, J., Lyons, J. and Bruin, M., 2013. Modeling the impact of changing patient flow processes in an emergency department: Insights from a computer simulation study, *Operations Research for Health Care* **2**(4): 66–74.
- L'Ecuyer, P., 2016. SSJ: Stochastic simulation in Java, software library. <http://simul.iro.umontreal.ca/ssj/>.
- Mills, A. F., Helm, J. E. and Wang, Y., 2021. Surge capacity deployment in hospitals: effectiveness of response and mitigation strategies, *Manufacturing & Service Operations Management* **23**(2): 367–387.
- Park, E., Ouyang, H., Wang, J., Savin, S., Leung, S. C. and Rainer, T. H., 2023. Patient sensitivity to emergency department waiting time announcements, *Manufacturing & Service Operations Management* .
- Patrick, J., Puterman, M. L. and Queyranne, M., 2008. Dynamic multipriority patient scheduling for a diagnostic resource, *Operations research* **56**(6): 1507–1525.

Rath, S. and Rajaram, K., 2022. Staff planning for hospitals with implicit cost estimation and stochastic optimization, *Production and Operations Management* **31**(3): 1271–1289.

Somanchi, S., Adjerid, I. and Gross, R., 2022. To predict or not to predict: The case of the emergency department, *Production and Operations Management* **31**(2): 799–818.

Xu, K. and Chan, C. W., 2016. Using future information to reduce waiting times in the emergency department via diversion, *Manufacturing & Service Operations Management* **18**(3): 314–331.

Chapter 5

Conclusion

The average wait time in Canadian EDs can be as high as nearly 4 hours, with approximately 10% of patients enduring wait times exceeding five hours before receiving initial physician care. Such prolonged delays can exacerbate patient conditions or lead to patients leaving the ED without receiving any medical attention. Patients are usually seen by a triage nurse immediately upon arrival and their urgency level is evaluated. Their initial triage is considered as the start of their wait time in ED. Then they wait in the waiting area while being in a virtual queue based on their CTAS level. Once a physician becomes available, they choose a patient from this virtual queue based on patients CTAS levels and their wait times. The initial assessment of the patient by a physicians will be the end of patients wait time. Therefore, physicians and their availability is one of the main determinants of patients wait time.

In this thesis, I approach this issue from two main perspectives. First, I propose a problem formulation to optimize physician scheduling in EDs that aims to minimize the mismatch between patient arrivals and available ED capacity throughout each hour of the day. A critical aspect of aligning demand and supply involves considering the variable productivity of physicians during shifts. Based on our data and according to the literature, physicians PPH decreases during a shift. More particularly, a physicians who has just started their shifts can see on average more than 3 new patients per hour, while a physician who is working at the last hour of their shift, can only see less than one new patient on average. In other words, in both cases we have one physician available, while in terms of productivity and capacity of ED, we have totally different scenarios. Considering this

variable productivity, we formulate the problem as a two-stage stochastic problem and solve it using Benders decomposition method. Through simulation modeling, I demonstrate that by accounting for this factor, we can reduce patient wait times simply by adjusting the start times of the existing ED schedule. Namely, without introducing new resources and with the same number of hours scheduled in a day, just by slightly changing the start time of shifts, we can have significant improvement in the average wait time of patients. Our proposed method integrates stochastic demand, physician productivity, and variable average Patients Per Hour (PPH) rates for physicians.

Due to the stochastic nature of EDs and the fact that they provide health care services in emergency situations, there will be times when the number of arrivals to the ED and/or the treatment time of patients exceeds expectations. Hence, the ED will experience overcrowding. Since resources are limited and expensive in EDs, it is not possible to increase resources at all times to be ready for these situations. Therefore, there is a need for a clear response strategy to address the challenging problem of overcrowding in EDs. EDs commonly employ strategies to address such situations by having flexible schedules: using overtime, having on-call physicians, or issuing surge calls. However, often there is no clear policy on how to implement these strategies, and the decision is made subjectively, which often fails to consider both current system status and future arrival expectations. Given the predictable daily pattern of ED arrivals, our decision-making on resource allocation should factor in both present and projected capacity needs.

A common approach to add flexibility to schedule is to use overtime or shift extensions. An advantage of shift extension is the fact that the physicians is already in the ED and no time will be wasted for commute. Therefore, we can make the decision closer to the time that we need the extra capacity for. Moreover, shifts extensions are usually shorter, for example two hours, and therefore, less costly compared to using on call physicians or surge calls where we add one whole shift. In the third chapter of this thesis, we formulate the problem as a Markov Decision Process with the main goal of finding an optimal policy in when to extend a physician's shift. The policy depends on the relative cost of shift extension to costs incurred by one more hour waiting time of a patient. Based on the hour of the day and expected number of arrivals to the ED and number of available physicians in the current and next few hours, the optimal policy is a control limit type for the number of patients waiting to be seen at the beginning of the hour. More particularly, at the beginning of any hour, based on the current status of the system we can decide whether we need to extend a shift. The Policy is very easy

to implement and can be very helpful for managers and schedulers of the ED to make a decision that considers the current and expected status of the ED.

However, the drawback of the shift extension approach is that the additional time is added at the end of the shift when the PPH (patients per hour) is typically not very high. Data shows that the average PPH of physicians, or the number of new patients they can see per hour, is less than one patient in the final hours of their shift. Therefore, with shift extensions, the ED must pay for two hours of the physician's time, even though their PPH will be below average, making this an overpriced resource. Additionally, shift extensions are only feasible if a shift is nearing its end. In other words, there will be times during the day when, despite needing more capacity, there is no shift close to ending, making it impossible to use shift extensions to increase capacity.

Another common ED strategy involves issuing surge calls to off-duty physicians. In this approach, whenever the ED is facing overcrowding, a surge call is issued to physicians notifying them about the surge in demand and asking them to join the ED if possible. However, when physicians are not scheduled for the day, they usually have other plans and consequently the response rate to the surge calls is lower than desired. Moreover, the lack of a clear optimal policy for when to issue a surge call, makes this approach ineffective. In the fourth chapter of this thesis, we frame the surge call issue as an MDP, and develop an optimal policy based on the current and expected future state of the ED. We also identify response rate thresholds necessary for strategy effectiveness. We show that depending on the number of shifts in a day and how scarce are the resources, we will need different response rates to make the strategy effective. The resulting optimal policy depends on the hour of the day, the current ED schedule, response rate of physicians, and the comparative costs of surge call deployment versus prolonged patients wait times. Results show that as resources (or physicians) become more scarce, the implementation of surge call policies can yield significant improvements in patient wait times. Also, when resources are very scarce, even response rate of 10% can promise improvement in average wait time while when resources are less scarce, we need to have a response rate of at least 30% to have an effective implementation of surge call policy.

In summary, this thesis provides ED managers and schedulers with valuable tools, policies, and guidelines for optimizing physician schedules and reducing average wait times in EDs. By implementing the proposed schedule outlined in Chapter 2, the average wait time for patients can be significantly

improved through better alignment of supply and demand. Additionally, in instances of overcrowding, the policies outlined in Chapters 3 and 4 offer helpful guidance for ED managers in addressing this challenging issue.

The main limitation of this thesis lies in its focus solely on the ED without considering its interconnected dynamics with the hospital as a whole. Specifically, one of the primary reasons for prolonged wait times in EDs is the insufficient availability of beds, as patients admitted to the hospital from the ED often need to wait for transfer. Improving physician schedules and increasing initial patient assessments may inadvertently exacerbate this issue by increasing the number of patients awaiting beds in the intake area. Essentially, this may shift the bottleneck from physician availability to bed availability. In future studies, we aim to investigate this by analyzing the length of stay distribution using our data and integrating it into our simulation model and to track the number of patients in the intake area in each hour of the day.

Another limitation is that in the third and fourth chapters we first use the optimization model from Chapter 2 to find a fixed schedule and then find optimal policies on when to issue a surge call or extend a shift. However, the ED wait time can be further improved if we solve an integrated problem where we find the schedule considering that we will extend the shift or issue a surge call if the number of patients waiting to be seen is more than a threshold.

The next potential avenue for future study could involve integrating both shift extensions and surge calls as potential strategies to address a surge in demand and determining an optimal policy for their use. Each strategy carries its own set of advantages and disadvantages. With shift extensions, we can reliably increase the capacity when needed, since the physician is already present in the department. However, the additional capacity is limited to an extra 2 hours for each scheduled shift per day. Conversely, surge calls offer less constrained additional capacity, as we can issue a surge call every hour of the day and each additional successful surge call will result in seven hours of extra capacity. However, the response rate is not guaranteed to be 100%, and there is a risk of no physicians responding to the call even when more are needed in the ED. Integrating both strategies could enhance flexibility and provide ED managers with more effective tools for managing overcrowding in EDs.