UNIVERSITY OF OTTAWA

TELFER SCHOOL OF MANAGEMENT

M.SC. IN HEALTH SYSTEMS THESIS

A Stochastic Optimization Approach for Staff Scheduling Decisions at Inpatient Clinics

---

Author:
Sajjad DEHNOEI

Supervisors:
Dr. Antoine SAURÉ
Dr. Onur OZTURK

Abstract

Staff scheduling is one of the most important challenges that every health care organization faces. Long wait times due to the lack of care providers, high salary costs, rigorous work regulations, decreasing workforce availability, and other similar difficulties make it necessary for healthcare decision-makers to pay special attention to this crucial part of their management activities. Staff scheduling decisions can be very difficult. At inpatient clinics, there is not always a good estimate of the demand for services and patients can be discharged at any given time, consequently affecting staff requirements. Moreover, there are many other unpredictable factors affecting the decision process. For example, various seasonal patterns or possible staff leaves due to sickness, vacations, etc.

This research describes a solution approach for staff scheduling problems at inpatient clinics where demand for services and patient discharges are considered to be stochastic. The approach is comprehensive enough to be generalizable to a wide range of different inpatient settings with different staff requirements, patient types, and workplace regulations.

We first classify patients into a number of patient groups with known care-provider requirements and then develop a predictive model that captures patients' flow and arrivals for each patient category in the inpatient clinic. This model provides a prediction of the number of patients of each type on each specific day of the planning horizon. Our predictive modelling methodology is based on a Discrete Time Markov model with the number of patients of different types as the state of the system. The predictive model generates a potentially large set of possible scenarios for the system utilization over the planning horizon. We use Monte Carlo Simulation to generate samples of these scenarios and a well known Stochastic Optimization algorithm, called the Sample Average Approximation (SAA) to find a robust solution for the problem across all possible scenarios. The algorithm is linked with a Mixed-Integer Programming (MIP) model which seeks to find the optimal staff schedule over the planning horizon, while ensuring maximum demand coverage and cost efficiency are achieved. To check the validity of the proposed approach, we simulated a number of scenarios for different inpatient clinics and evaluated the model's performance for each of them.

Contents

List of Figures

List of Tables

List of Abbreviations

LOS	**L**ength **O**f **S**tay
RN	**R**egistered **N**urse
MIP	**M**ixed **I**nteger **P**rogramming
DTMC	**D**iscrete **T**ime **M**arkov **C**hain
ER	**E**mergency **R**oom
PRISM	**P**ediatric **RI**sk (of) **M**ortality
ARIMA	**A**uto **R**egressiv **I**ntegrated **M**oving **A**verage
SAA	**S**ample **A**verage **A**pproximation
SVM	**S**upport **V**ector **M**achine
IP	**I**nteger **P**rogramming
CTAS	**C**anadian **T**riage (and) **A**cuity **S**cale

List of Symbols

Predictive Model:

i	Index for a patient type
j	Index for a day
k	Index for a shift
I	Number of patient types
J	The last day of the planning horizon
C	Number of beds(i.e., capacity)
S	Number of staff
d	Index for Care-provider type
X_{ijk}	Number of patients of type i on shift k of day j
P	Transition probability matrix
P^i_{mn}	The probability of observing n type i patients in the next shift after observing m patients type i now
λ_{ijk}	Rate of arrival of patients of type i on shift k of day j
A_{ijk}	Number of arrivals of type i on shift k of day j
R_{ijk}	Number of patients of type i on shift k from day $j-1$
WL_{ijk}	Length of the wait list of type i patients on shift k of day j
X_{jk}	Total number of patients on shift k of day j
R_{jk}	Total number of remaining patients on shift k of day j
WL_{jk}	Total Length of the wait list on shift k of day j
LoE_i	Length of stay for patient type i
$ELOS_i$	Expected length of stay added and updated by physicians
$Poi(\lambda)$	Poisson distribution with parameter λ
$exp(\mu)$	Exponential distribution with parameter μ
$bin(n, p)$	Binomial distribution with parameters n and p
E	The probability event of an observed census

in the clinic

IP Scheduling Model:

i	Index for an illness/patient type
j	Index for a day
k	Index for a shift
s	Index for a staff type
S	Number of staff
I	Number of patient types
ξ	Index for a scenario
$S_{\xi\, sjk}$	Assignment of staff s to shift k of day j for scenario ξ
$X_{\xi\, ijk}$	Number of patients of type i on shift k of day j for scenario ξ
SR_{dik}	Required ratio of staff d for shift k of patient type i
$\delta^{-}_{\xi\, djk}$	Over-staffing of care-provider d on shift k of day j for scenario ξ
$\delta^{+}_{\xi\, djk}$	Under-staffing of care-provider d on shift k of day j for scenario ξ
$\Pi^{-}_{\xi\, s}$	Number of shifts over the weekly target for care-provider s in scenario ξ
$\Pi^{+}_{\xi\, s}$	Number of shifts under the weekly target for care-provider s in scenario ξ
Λ^{-}_{ξ}	Patients of the wait-list above its target in scenario ξ
Λ^{+}_{ξ}	Patients the wait-list below its target in scenario ξ
$\kappa^{-}_{\xi\, sj}$	Number of overtime daily shifts for staff s on day j for scenario ξ
$\kappa^{+}_{\xi\, sj}$	Number of daily shifts under target for staff s on day j for scenario ξ
C^{-}_{djk}	Cost associated with over-staffing of care-provider d on shift k of day j
C^{+}_{djk}	Cost associated with under-staffing of care-provider d on shift k of day j
C^{-}_{s}	Cost of a shift covered after fulfilling the target for shifts, for staff s
C^{+}_{s}	Cost of a shift covered before fulfilling the target for shifts, for staff s
C^{-}_{sj}	Cost of a care-provider

	having a number of shifts higher than the daily target
C_{sj}^{+}	Cost of a care-provider having a number of shifts less than the daily target
C_{WL}^{-}	Penalty for when the length of the wait-list is more than its target

Stochastic Optimization:

N	Number of required samples from the SAA method
N_0	Initial number of samples in the SAA
N_1	Number of samples in the evaluation step of the SAA method
M	Number of replications in the SAA method
ϵ	Optimality tolerance for the SAA method
α	Stopping criterion for the SAA method
AOI	Approximate Optimality Index

Chapter 1

Introduction

What is the best staff schedule for a health care facility? This has always been an important question for health care decision-makers at different managerial levels. Poor scheduling decisions could cost money to a health care organization, increase wait times for patients, cause personnel troubles, and, decrease the overall quality of care. Many efforts have been undertaken in both academia and practice to tackle this problem. However, there is still a lot to do. Rigid workplace regulations and the unpredictable nature of patients' arrivals and lengths of stay make staff scheduling decisions at inpatient clinics challenging. It is difficult to make a decision on the number and type of staff for future days based on today's limited information. This research addresses this challenge with a focus on inpatient health care facilities.
We focus mostly on dealing with uncertainty in demand for care and patient discharges. Uncertainty can make the task of determining the right mix of staff on any given shift extremely difficult. Ignoring uncertainty or treating it poorly can cause significant over-staffing or under-staffing. Obviously both of these may adversely affect the quality of care and increase costs. There are of course other goals for staff schedules such as ensuring a fair distribution of shifts among staff, flexibility over changes to staff availabilities, etc.
Here the challenge is to determine the optimal number of staff with different specializations with regard to demand uncertainty to ensure patients quality of care, personnel satisfaction, demand coverage while avoiding overtime/idle time and calling-ins as much as possible.
This research uses a range of mathematical techniques to find a solution to this problem. The details of our modelling and solution approach are presented in chapters 4 and 5. Here we briefly summarize our approach to prepare the reader for the literature review section.

1. **Modelling Arrivals, Discharges, Care-providers, Patients, and Treatments**
We consider an inpatient clinic which is staffed by multiple care providers such as nurses with different skills, physicians, etc. For each care provider we consider a set of skills, availability, and costs of regular and over-time shifts. We also take into account some of the most common schedule rules such as maximum number of weekly

shifts as well and maximum number of daily shifts. We have also considered a set of complementary constraints in A.

These staff provide a set of treatments to patients hospitalized with different illnesses. Patients are categorized into different groups based on a set of characteristics including their disease type, required intensity of care, the time they have already spent waiting to get admitted, etc. Patient groups can be created using a well-known Clustering technique called K-means. Each patient group requires different resources and has a different priority in regards to admission. The model captures admissions from referrals and from the ED department and allows discharges during any of three shifts daily. When a patient cannot enter the clinic, based on the clinic's utilization, the model assigns the patient to a wait-list or simply rejects him. Patients in the wait-list have priority over new arrivals. Amongst those patients in the wait-list, further priority is given to each category. Patients are not admitted to the system unless all higher priority patients are admitted.

2. **Solution Approach-MIP Model and SAA Algorithm**

We consider a predictive model that simulates patient arrivals and discharges based on different scenarios of clinic census. Then, a Mixed-Integer Programming model uses the forecasts with the goal of optimizing over and under staffing levels with respect to the constraints. However, because of the stochastic nature of the problem and the so-called curse of dimensionality, there exists an enormous number of possible scenarios. This makes the task of creating a good schedule extremely difficult.

In order to address this challenge, we combine the results of the predictive model and the MIP model using a Stochastic Optimization technique called Sample Average Approximation. This algorithm uses Monte Carlo Simulation to and provide a robust and reliable solution across all these scenarios created by the predictive model.

Finally, we present some test datasets corresponding to different inpatient clinic configurations with different arrival and discharge rates, patients categories, and care-providers to test the solution approach. Chapter 6 presents this information and Chapter 7 discusses the results of running the algorithm on these datasets.

Chapter 2

Background and Literature Review

Johann Wolfgang von Goethe, famous German statesman, said:

"*A person who does not know the history of the last 3,000 years wanders in the darkness of ignorance, unable to make sense of the reality around him.*"

Before anything else, it is necessary to review the literature to get a better understanding of the areas of interest related to this research.
Here, we first present an overall review of the literature on scheduling in healthcare which we then categorize into 4 different sections. The first section, explores the literature on *Predictive Models* in healthcare. It describes strategies that have been previously used to model a healthcare setting and forecast the demand for its resources. The sub-categories of this section are *Markov Models*, *Simulation Methods*, *Queuing Theory*, *Machine Learning Algorithms*, and a final sub-section called *Other Models*, including *Time-series* forecasting models and other approaches. Next, a section is devoted to *Scheduling Models* previously developed for similar settings. How have researchers modelled an inpatient clinic with the goal of optimizing care providers' schedule? The focus is mostly on Mixed Integer Programming (MIP) models as it is one of the methodologies used in our work as well. Further justification and explanation for its suitability is provided later in this chapter. After these two sections, another section is dedicated to *Stochastic Optimization*. As briefly mentioned earlier, after the predictive modelling piece is used, we will have a large set of scenarios. For this reason, we need a model that is capable of producing robust solutions across all these scenarios. Thus, the question in this thesis is what methodologies and techniques have been applied by researchers to solve Stochastic Optimization problems similar to the one we have here. Finally, a section is dedicated to notions and ideas related to our research that are inspired by other studies but do not fit under any of the previous categories. Qualitative and quantitative works were considered and the more related ones are discussed here.

2.1 An Introduction to the Literature Review

Throughout the years, the ever-increasing costs of healthcare organizations have brought attention to the potential of a number of mathematical techniques to optimize all kinds of operational, administrative, and managerial activities in healthcare and settings (Medicine, 2001; Squires and Anderson, 2015; W. Raghupathi and V. Raghupathi, 2014). Operations Research in Healthcare has received considerable attention for more than three decades (see Rais and Viana, 2011 for a comprehensive review). Staff Scheduling and Demand Forecasting are amongst the most important problems in this area (Chaari et al., 2014; Rais and Viana, 2011). Forecasting and scheduling have also been explored and applied to settings other than healthcare, in the next chapters, we cite some of those studies as well.

2.2 Predictive Models

Predicting the number of patients of different types on any given day of the planning horizon is equivalent to predicting arrivals, discharges, and lengths of stay. Essentially, we have some patients currently in the clinic, a group of these patients can be discharged at some point in the future and some new patients can be admitted. In the literature, there is no unique approach toward identifying some of these variables and predicting the others. For example, one can predict LOS directly from historical data and then consider a rate of arrival to the clinic, or one can simply consider the number of patients in the clinic and use time-series to find an estimate of the future census. For this reason, below we group papers based on their predictive modelling strategy rather than the variables they aim to predict.
As mentioned above, this section is divided into five different sub-sections as follows:

1. **Markov Chains**:
In this section, we mention studies considering the Markov property, or more specifically Markov Chains, whether in discrete or continuous time, to model a healthcare system and find possible projections for the future number of patients and their lengths of stay based on their arrival distribution and/or discharge rates. The main common feature of these studies is the use of the Markov Property[1] in some part of the modelling process.
In our work, we use a Discrete-Time Markov model to capture arrivals, discharges, and generate possible future states of the system.

[1] A stochastic process has the Markov property if the conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, not on the sequence of events that preceded it.

2. **Simulation Methods**:
In this section, we mention studies that apply simulation modelling to capture the system's behaviour in order to find estimations for the amount of resources such as care-providers, beds, etc. There is a link between Simulation and other methodologies here.
In our work, we simulate the number, and consequently, the mix of patients in the clinic in the future using a Monte Carlo Simulation approach. These simulated trajectories of the system play a key role in the formulation of the Stochastic Optimization model that we eventually solve.

3. **Queuing Theory**:
Here we consider studies involving Queuing Theory applications to find predictions for upcoming possible stages of the system. It should be noted that there is an intersection between Queuing models and Markov chain models. Both of these methodologies model a stochastic process using the same approach (Ross, 2014). In this literature review, these two concepts have been distinguished based on the focus of the study. Papers with a focus on finding queue length as a proxy for waiting times are considered in the third group (queuing) while studies predicting ward utilization via the application of Markov Chains are categorized in the first group.
In this research, we use some elements of Queuing Theory to first evaluate possible states of the system and second, model and evaluate the number of people waiting for services and the cost associated with the length of the wait list.

4. **Machine Learning Algorithms**:
Machine Learning algorithms and models are very popular for finding patterns in large datasets in a fast manner with an efficient effort (Lodhia, Rasool, and Hajela, 2017). More specifically, Machine Learning algorithms, and Data Mining and Data Clustering techniques, form a notable part of the literature on the application of quantitative techniques in healthcare (Jothi, Husain, et al., 2015; Ahmad, Qamar, and Rizvi, 2015; Callahan and Shah, 2017a).
Various Machine Learning algorithms have been used to develop predictive models for the utilization of resources in healthcare facilities (Tanuja, D. U. Acharya, and Shailesh, 2011; Liu et al., 2006). Here we briefly discuss some of the most relevant studies.
In our research, we have evaluated these techniques and the possibility of applying some of them to forecast demand for care services. We also propose a clustering technique to categorize patients with similar resource requirements into meaningful categories.

5. **Other Models**:
Some other quantitative techniques have also been applied in healthcare settings to

predict demand. These are more traditional mathematical approaches that cannot be grouped under any specific category. A short summary of some papers using these methodologies is presented as well. One could confidently label these studies as less difficult methods that provide relatively robust results with less efforts (Hyndman and Athanasopoulos, 2018).

2.2.1 Markov Chains

A Markov Chain is a stochastic model describing a process in which the probability of visiting each state of the system depends only on the state attained in the previous event (Ross, 2014). Markov Chains have been used to model various aspects of healthcare systems. Sato and Zouain, 2010, discuss a variety of these applications. Here we consider articles that more specifically model arrivals and discharges in order to capture uncertainty around hospital occupancy (i.e, the number of patients of different types at the healthcare facility on any day of the planning horizon). The focus is on inpatient clinics.

A number of researchers have defined a patient's level of care as the states of the system. For example, Gordon Taylor, Mclean, and Peter Millard, 1996, considered Acute Care, Rehabilitative Care, Community, and Dead as states. Obviously Dead is an absorbing state (Ross, 2014) and the limiting probabilities were calculated to find an estimation for the time a patient stays in a Long-term Care Facility until death. The same researchers also found that the number of patient could be modelled as a function of the length of stay and assumed that such a function follows a three-term exponential distribution (Gordon Taylor, Mclean, and Peter Millard, 1996). After a few years, they elaborated more on their previous study by applying it to a 16-year census dataset and proposed a four-compartment mixed-exponential model, which is a function of patient's length of stay and predicts the number of geriatric patients in St. George's Hospital in London (GJ Taylor, S. McClean, and Millard, 2000).

The usefulness of LOS prediction models that use exponential terms for geriatric care has been shown by other researchers too (Marshall, Sally I McClean, and Peter H Millard, 2004). There is even a decision-support system that tries to find the best mixed-exponential model for patients' length of stay (Mackay, 2001). In addition, GJ Taylor, S. McClean, and Millard, 2000, assigned costs to states and found the best spending strategy for the hospital they studied. This is a constructive example of how Mathematical Modelling can help reduce healthcare expenditures and how it has actually been applied to a healthcare setting for forecasting purposes.

Another comprehensive example of the use of Markov Chains as a predictive modelling technique, this time with a focus on an emergency department, is described in

Zhang et al., 2013. Discrete Time Markov Chains have been applied to model the patient census in inpatient clinics. Broyles and Cochran, 2009, present a Discrete Time Markov Chain methodology based solely on historical local patient census. They considered the number of patients in the hospital as the states of the system and used the Markov model to predict length of stay in the hospital. The limitations of their work are, first, the number of patient types they consider(they only consider one patient type) and, second, the lack of patients' arrival modelling as a separate part of the model. We will consider these two factors in this thesis.

Kapadia et al., 2000, define three states for a Discrete Time Markov Chain model: low, medium, and high. These states represent the situation of each patient based on a well-defined medical term called "Paediatric Risk of Mortality (PRISM)". The Markov model is used to predict the number of patients from each type in the paediatric department of the Texas Medical Center in Houston (Kapadia et al., 2000). This work is an example of the application of Markov Models in paediatric care settings. Kapadia et al., 2000, along with Gordon Taylor, Mclean, and Peter Millard, 1996, show the comprehensiveness of Markov Chain models for forecasting demand for settings with different LOS.
A considerable number of researchers have used Semi-Markov models [2] to estimate patients' length of stay. For example, Jain, 1989, defines a continuous-time Markov model [3] with states representing the various stages of a patient's care level and then uses the model to predict patients length of stay. Vasilakis and Marshall, 2005, also applied Semi-Markov modelling but using a phase-type probability distribution[4] for the length of stay instead of an exponential distribution.
Continuous-time Markov Chains have been applied to psychiatric care settings. For example, Eaton and Whitmore, 1977, use a continuous-time Markov model to predict length of stay for hospitalization of schizophrenia patients. Squires and Anderson, 2015, define a Continuous-time Markov Chain in a community care environment with states representing patients' place of care (Residential Home Care, Nursing Home care, and Discharge) and estimate the survival time of elderly patients community care services.

As another example of how Markov models can cope with uncertainty in healthcare settings, one can consider Patrick, 2012a's work. Chanchaichujit et al., 2019, also mention Markov Chains as important and robust stochastic modelling tools for

[2] A semi-Markov process is equivalent to a Markov renewal process in many aspects, except that a state is defined for every given time in the semi-Markov process, not just at the jump times. Therefore, the semi-Markov process is an actual stochastic process that evolves over time.

[3] A continuous-time Markov chain is a continuous stochastic process in which, for each state, the process will change state according to an exponential random variable and then move to a different state as specified by the probabilities of a stochastic matrix.

[4] A phase-type distribution is a probability distribution constructed by a mixture of exponential distributions.

healthcare problems.

Another study that is very helpful to understand how Stochastic Modelling can be applied to inpatient units and how it can be used together with Simulation to forecast demand is Broyles, Cochran, and Montgomery, 2010's use of Discrete-Time Markov Chains (DTMC) as a predictive modelling tool for forecasting inventory levels (i.e., number of patients) at an inpatient hospital. After developing a DTMC and applying it to an inpatient clinic they compared it with an Auto Regressive Integrated Moving Average (ARIMA) model and showed significant improvement in terms of forecast accuracy.

Altogether, Markov Chains, especially when combined with Simulation have proven to often offer an appropriate approach to deal with the uncertainty inherent in healthcare problems (Capan et al., 2017).

2.2.2 Simulation Methods

Simulation has been used implicitly in some of the studies already discussed above specifically to find transition probabilities in Markov models (Gordon Taylor, Mclean, and Peter Millard, 1996; Broyles, Cochran, and Montgomery, 2010). It has also been used in some of the following studies to simulate the length of a queue as a proxy for wait times in a queuing system (e.g., Bidhandi et al., 2019). In this section, we review studies with focus on Simulation as the main methodology. It is important to note that almost all of the simulation models in the relevant literature predict steady state probabilities and do not provide predictions for transient periods (Broyles, Cochran, and Montgomery, 2010).

Nguyen et al., 2007, assume that patients' lengths of stay follow a normal distribution and then use the set of equations in their model to run a simulation and find the number of beds required in the future by considering how requirements fluctuate over time. Using a Normal distribution for LOS and creating random samples based on that is possible in our approach. In a different study, but using the same methodology, the same researchers compared their approach with classic bed occupancy prediction techniques (such as simple mean, moving average, etc.) and showed better results. Gallivan et al., 2002, developed a mathematical model and used it to predict lengths of stay with the ultimate goal of predicting the number of occupied beds. They also showed that real length of stay data tends to follow exponential probability distributions (Gallivan et al., 2002).

A large number of studies have applied various simulation techniques such as Monte Carlo Simulation, Discrete-Event Simulation, System Dynamics, and Agent-based Simulation to problems relating to healthcare (Medicine, 2001; Katsaliaki and Mustafee,

2011). Mustafee, Katsaliaki, and S. J. Taylor, 2010, provide a comprehensive survey of applications of these methods in healthcare. Simulation is also a popular method for evaluating a system's performance under a number of possible future scenarios (Patrick, 2012b; Bidhandi et al., 2019). We will use Monte Carlo Simulation to evaluate the performance of the proposed method for an inpatient clinic under multiple possible realizations of arrivals, discharges, and care-provider requirements. As briefly explained in 1, samples of upcoming demand for care at the clinic will eventually be a part of a Stochastic Optimization algorithm called *Sample Average Approximation*. The algorithm solves the expected value problem by repetitively generating Monte Carlo samples. Complete details of this approach are presented in Chapter 4. It is important to assure that a similar approach has been applied to an inpatient setting by Bagheri, Devin, and Izanloo, 2016. The differences between their work and what we do is first, the fact that they only considered one care-provider type (only nurses) while we consider different types of nurses along with other specialities. Second, while they only consider one patient type, our model is capable of considering various patient types. Finally, unlike their work, our model considers wait-lists as well. More discussion on these aspects of our model will be presented in Chapter 4.

2.2.3 Queuing Theory

Queuing Theory might be one of the most frequently used approaches by quantitative healthcare researchers. There is an extensive literature on applications of Queuing Theory in healthcare (Lakshmi and Iyer, 2013; Fomundam and Herrmann, 2007; Mehandiratta, 2011). These review papers of course include applications regarding patient flow modelling (Lakshmi and Iyer, 2013). Specifically, Queuing models are useful when minimizing average wait time (Ernst et al., 2004a). In this subsection, a summary of the most relevant studies is presented.

Green et al., 2006, evaluated some critical health-service quality factors such as the number of patients admitted and the length of stay before and after the implementation of an $M/M/1$ queuing model to predict care provider needs of patients arriving to an emergency department. Results showed outstanding improvements after model implementation.

Cochran and Roche, 2008, used a multi-server queuing model to represent bed availability in an inpatient setting in Phoenix, Arizona. After various levels of data analysis, they found that the arrival rate of patients to the clinic (coming from 4 different sources) followed a Poisson distribution allowing a better forecast of upcoming arrivals.

De Bruin et al., 2007, developed an M/M/∞ queue model to capture patient arrivals and lengths of stay at an emergency cardiac inpatient setting. They modelled the arrivals as a Poisson process and service times (i.e., LOS) with an exponential distribution. They also managed to show that the Poisson and exponential assumptions properly match the distributions of historical data. These assumptions correspond to some of the scenarios we consider in the analysis of the performance of our model.

Cooper and Corcoran, 1974, presented an estimation for the number of beds needed for acute Myocardial Infarction patients based on a multiple-server Queuing model. Their model replaced an old traditional arithmetic model which was prevailing in the healthcare system by that time.

Some studies like that of Yankovic and Green, 2011, use an M/M/C model to find an optimal nursing level. Their work assume Poisson arrivals and exponential service times similar to Cooper and Corcoran, 1974.

Lastly, El-Darzi, Vasilakis, et al., 1998, applied a basic Queuing model with a three-compartment exponential length of stay approximation. A three-compartment Exponential distribution consists of the summation of three exponential terms instead of a single term. Assuming a Poisson arrival process, they found the average length of stay for each patient group (Acute, Rehabilitation, and Long Stay) in a geriatric clinic (El-Darzi, Vasilakis, et al., 1998).
In staff scheduling, a Queuing model by used in Agnihothri and P. F. Taylor, 1991, to determine the staffing levels needed to handle call arrivals for hospital appointments.
As another example, Ernst et al., 2004a mention that Queuing models are elegant and may provide analytical results about the length of a queue and the congestion level, but in general many unrealistic simplifications/assumptions need to be made. Of course, the same as for Markov Chains, Queuing models are usually combined with simulation.

2.2.4 Machine Learning Algorithms

In today's world, after the emergence of various information gathering techniques and technologies, an immense and precious amount of data is available to researchers. In healthcare, Machine Learning algorithms and Data Mining and Data Clustering techniques have been constantly applied to achieve various goals whether in improving quality of care or for financial reasons. Koh, Tan, et al., 2011; Jothi, Husain, et al., 2015; Shafqat et al., 2018 provide reviews of applications of Data Mining in healthcare, and Callahan and Shah, 2017b; S. Dua, U. R. Acharya, and P. Dua, 2014,

on the applications of Machine Learning in healthcare.

Here we analyse the literature from two different points of view. First, as previously mentioned, we categorize patients into a number of groups to be able to better predict the clinic's occupancy level. Second, we determine their care-providers requirements more accurately. There are two possible approaches for doing this: defining patient categories based solely on clinical information such as disease type, predicted LOS, etc. or considering a clustering algorithm for patients based on additional characteristics. Consequently, it is important to study and understand these algorithms. Another point of view is to consider Machine Learning algorithms and how they have been applied in healthcare settings to predict demand.
Isken and Rajagopalan, 2002's work is a good starting point. In this paper, the authors study similarities between different patient groups by the use of a K-means data clustering algorithm. They showed how K-means Clustering can improve the results from the application of Simulation and other analytical methodologies in hospitals.
Another group of studies that are of interest are those applying Machine Learning and Data Mining techniques to predict lengths of stay. For example, Hachesu et al., 2013, applied three different data classification algorithms to predict cardiac patients' lengths of stay (Decision Trees, Support Vector Machine (SVM), and Artificial Neural Networks). They found that SVM was more accurate in comparison with the other techniques. The same approach was applied in a different setting by Liu et al., 2006. They applied Decision Trees and Naive Bayesian classifiers to predict geriatric patients' length of stay. A comprehensive study of applications of Machine Learning to demand forecasting in inpatient clinics and ED departments is described in Ibrahim, 2019. In this paper, various algorithms are considered and tested, with the Gradient Boosting Classifier showing considerably better accuracy for inpatient facilities,. More straightforward methods have also been used to predict length of stay, and hospital's occupancy. For example, Don Liew, Danny Liew, and Kennedy, 2003, applied Logistic Regression to a geriatric setting to predict whether a patient is a long or short stay one. Finally, Tanuja, D. U. Acharya, and Shailesh, 2011, studied the performance of four different algorithms for predicting length of stay (Neural Networks, Naive Bayes Classifier, K-means Clustering, and Decision Trees). After examining their performances on a large amount of data, they concluded that a model based on a Neural Network algorithm produced the best results.

2.2.5 Other Models

In addition to the previously mentioned methodologies, there exist other methodologies and approaches for demand prediction in a staff scheduling context. As a well known example of these methodologies one can consider time-series methods

and their important role in predicting demand in healthcare. Lapierre et al., 1999; Kao and Pokladnik, 1978 provide some examples of this approach.
Some researchers used average length of stay as a tool for predicting up-coming discharge rates (Farmer and Emami, 1990). Although these models have some strengths; there are at least three reasons why the average length of stay should be avoided when developing bed utilization models (Mackay and Lee, 2005). First, the length of stay is not Normally distributed, in fact its distribution typically has a high degree of skewness (Mackay and Lee, 2005). Secondly, the length of stay distribution is complex (Mackay and Lee, 2005). Typically, the term "inpatient" represents a heterogeneous group of people who have been admitted to a hospital for a variety of reasons and of course there are various factors affecting their lengths of stay. In fact, the main endeavour of the current research is to manage such complexity. Thirdly, arrival rates, discharge rates, and availability of resources (care-providers in our case) are strongly affected by seasonality.
Mackay and Lee, 2005, show that to some extent, there is an inverse relationship between model complexity and its efficiency in predicting length of stay. We will try to find a reasonable balance between model complexity and its ability to predict inpatient bed occupancy. Although we did not directly used any of these techniques for prediction, these techniques could be useful for determining rates and/or benchmarking the results we get from the proposed predictive model.

2.3 Mixed Integer Programming (MIP) Staff Scheduling Models

This section is on Staff Scheduling models developed using Mixed Integer programming (MIP). MIP models have been widely used to address personnel scheduling problems in practice (Van den Bergh et al., 2013). There is a variety of staff scheduling studied using MIP models in healthcare (Van den Bergh et al., 2013; Trilling, Guinet, and Le Magny, 2006; Ernst et al., 2004b). We will elaborate more on the studies that are most relevant to ours.

Staffing problems occur in many different healthcare settings (Van den Bergh et al., 2013; Pardalos et al., 2013). Ernst et al., 2004b, present a throughout list of staff scheduling problems in healthcare. These studies provide a general description of staff scheduling problems and how they should be approached and modelled in general. The article provides a useful framework for the development of a comprehensive staff scheduling model. Sitompul and Randhawa, 1990 present another (slightly older) review on Staff Scheduling.
As a specific example of staff scheduling in healthcare, Brunner, Bard, and Kolisch,

2009, developed a Mixed Integer programming model with the objective of assigning physicians to some rotations subject to standard real-world constraints. The same group of researchers extended their work by adding flexible physician shifts. They solved the model using a Branch and Price method (Brunner, Bard, and Kolisch, 2010). Eveborn, Flisberg, and Rönnqvist, 2006, defined and solved a pretty similar staff scheduling problem by using the Repeated Matching Algorithm, which was part of a decision support system meant to help home care planners.
It is important to note that most MIP models for this type of problem consider deterministic patient arrivals and/or resource requirements which makes the research in this thesis a reasonable contribution to the field. Pardalos et al., 2013, consider the development of dynamic models which incorporate the stochastic nature of hospital data as an open challenge.

We were also inspired by the idea mentioned in Trilling, Guinet, and Le Magny, 2006, to improve the quality of the resulting schedules from our work. This is discussed in detail in Chapter 4. We consider a two-stage MIP model with binary decision variables assigning care-providers to different shifts on different days in the first stage, and the second-stage stage penalizes over and under staffing. This methodological approach is similar to the one in Bagheri, Devin, and Izanloo, 2016.

2.4 Stochastic Optimization

Making decisions under uncertainty is obviously a difficult task. In our problem setting, staff scheduling decisions are supposed to be made while complete information of the future demand for care is not available. Although a predictive model can narrow down the number of possible demand scenarios and simulation can help us project current states of the system, there is still the need to have an on-line, efficient strategy to link these forecasts and possible scenarios to the actual scheduling decisions. In more precise words, we need a strategy to find an optimal mix of care-providers based on the scenarios created by the predictive model. We address this issue by adopting a Stochastic Optimization strategy which helps us deal with an enormous number of scenarios generated via simulation in order to find a unique staff schedule.
In the rest of this chapter, we first provide a brief definition of Stochastic Optimization, then a summary of some of the methodologies used in Stochastic Optimization, and lastly, we elaborate more on the technique of choice which is the Sample Average Approximation.

2.4.1 A Brief Introduction to Stochastic Optimization

In recent decades, since the development of digital computers, many researchers have studied the problem of numerically optimizing an objective function (Fouskakis

and Draper, 2002). One of the resulting approaches, in which the search for the optimal solution incorporates randomness in some constructive way, is called Stochastic Optimization (Fouskakis and Draper, 2002). In another definition, Homem-de-Mello and Bayraksan, 2014, define Stochastic Optimization as "a combination of modelling and methodology for optimizing the performance of systems while taking into account the uncertainty in some of the problem's parameters explicitly".
The same as for other types of optimization, Stochastic Optimization models consist of a set of parameters, some decision variables, an objective function, and a set of constraints. The difference with respect to deterministic models is that the goal usually is to optimize the expected value of the objective function and/or satisfy a set of constraints in expectation. Typically some part of the problem is not deterministically known and instead only a probability distribution of it is known.
Equation 2.1 shows a generic representation of a Stochastic Optimization model.

$$min_{x \in X}\{g_0(x) := E[G_0(x,\xi)] | E[G_k(x,\xi)] \leq 0, k = 1,2,...,K\} \tag{2.1}$$

where G_k, $k = 0,1,...,K$, are real-valued functions with inputs being the decision vector x and a random vector ξ. A variety of Stochastic Optimization models can be formulated based on different forms of Equation 2.1. Fouskakis and Draper, 2002, present a comprehensive list of all possible forms of a Stochastic Optimization model. A well known form of Equation 2.1 is a two-stage Stochastic Optimization model when $k = 1$ and $E[G_k(x,\xi)]$ is in form of optimizing second-stage recourse variables (Homem-de-Mello and Bayraksan, 2014). In our study, we consider a specific case, where the first-stage variables assign care-providers to shifts and the second-stage variables are the penalties associated with over or under staffing based on the realization of the demand for care. More details on this are presented in Chapter 4.

In general terms, Stochastic Optimization can be seen as a suitable solution approach when modelling uncertainty as a set of random variables and their relationships is possible (Bianchi et al., 2009). Otherwise, other techniques may be more appropriate. For example, when only information on the bounds of the variables is available, Robust Optimization might be a good choice (Bianchi et al., 2009). In fact, Stochastic Optimization is the closest to Deterministic Optimization when handling uncertainty (Bianchi et al., 2009).
Since there exist different categories of Stochastic Optimization models, the methodology for handling uncertainty vary widely. For small cases, when the integral of the function $G_0(x,\xi)$ can be calculated easily, optimizing a close-form expression of expected value function is the best approach. On the other hand, when it is not easy to handle $G_0(x,\xi)$ and/or the number of realizations of ξ is too large, other approaches have been proposed (Homem-de-Mello and Bayraksan, 2014).

One set of approaches for managing a large number of possible scenarios is Heuristics and Meta-heuristics[5]. These methods have been frequently applied to Stochastic Optimization problems (Fouskakis and Draper, 2002). Among the most important and most practical ones in the literature are Simulated Annealing (SA), Genetic Algorithms (GA), Ant Colony Optimization (ACO), and Tabu Search (TS)(Bianchi et al., 2009). These algorithms have been widely used in Deterministic Combinatorial Optimization as well, particularly when solution/execution times are significant and/or the number of possible solutions is beyond enumeration. For a comprehensive review of the application of these methods in Stochastic Optimization and a guideline for their use, see (Bianchi et al., 2009). We did not employ any of these Heuristics for two reasons. First, because unlike the need for the use of algorithms, the cost of creating solutions in our case (which is the execution time of the model) is not considerable and, second, finding the exact value of the objective function for a solution is not time-consuming. Based on Bianchi et al., 2009, the Sample Average Approximation method is a better choice for a problem with these features.
Sample average techniques are a practical approach to use for large problems (Bianchi et al., 2009). These techniques often provide good solutions for problems with a very large number of scenarios using a small number of samples in the calculation of the expected value function. For example, Linderoth, A. Shapiro, and Wright, 2006, evaluate the quality of the solutions to a two-stage Stochastic Optimization model found through Monte Carlo Simulation Sample Approximation. They consider various evaluation techniques such as enumerating the optimality gap or statistical KKT tests. Kleywegt, A. Shapiro, and Homem-de-Mello, 2002, also present a comprehensive mathematical proof of how Monte Carlo Simulation-based approximation algorithms converge to optimal solutions with probability 1 as the sample size increases. Not only is their work not limited to the evaluation of the convergence rate, but they also present a statistical proof of the convergence of the approximate solution found through simulation. Bianchi et al., 2009, provide a comprehensive summary of the literature on applications of Sample Approximation Techniques to various problem settings.
We have built our uncertainty handling strategy based on the Sample Approximation approach. Next section presents more literature on the SAA and Chapter 4 explains how we use this technique in our solution approach.

2.4.2 Sample Average Approximation (SAA)

The concept of Sample Average Approximation appears to have been used for the first time in Linderoth, A. Shapiro, and Wright, 2006. The basic idea behind the

[5] A heuristic technique, or simply a heuristic, is any approach to problem solving that employs a practical method that is not guaranteed to be optimal, perfect, or rational, but is nevertheless sufficient for reaching an immediate, short-term goal.

Sample Average Approximation is simple and reasonable. It is that the optimal solution to an Expected Value Stochastic model can be approximated by the solution that considers a large enough random sample of possible scenarios. More precisely, a random sample of scenarios is generated and then the expected value function of the original problem one wants to solve is approximated by the corresponding sample average function. The resulting sample average optimization model is solved and the procedure is repeated several times until a stopping criterion (which is usually a pre-specified optimality gap) is satisfied (Linderoth, A. Shapiro, and Wright, 2006). This process can be described as follows (Linderoth, A. Shapiro, and Wright, 2006):

SAA Algorithm for Stochastic Discrete Optimization problems.

1. Choose initial sample sizes N and N', a decision rule for determining the number M of SAA replications , a decision rule for increasing the sample sizes N and N' if needed, and tolerance ϵ.

2. For $m = 1, ..., M$, do steps 2.1 through 2.3.
   2.1 Generate N samples and solve the SAA model in 2.1 obtaining the objective value $\hat{v_m}$ and the optimal solution $\hat{x_m}$ associated with it.
   2.2 Estimate the optimality gap by creating N' scenarios and finding the expected value objective function associated with them using the solution $\hat{x^m}$ and find the variance of the gap estimator.
   2.3 If the optimality gap (and/or the variance of the gap estimator) is sufficiently small, go to step 4.

3. If the optimality gap (and/or the variance of the gap estimator) is too large, increase the sample sizes N and/or N', and return to step 2.

4. Choose the best solution $\hat{x}$ among all candidate solutions $\hat{x_m}$ produced in M iterations, using a screening and selection procedure.
   Stop.

The above algorithm is in its general form. This means that based on different values for the parameters N, N', M, ϵ, and stopping criterion, different versions of the Sample Average Approximation method can be obtained.
The algorithm should start with some values for these parameters. Generally speaking, although there are several guidelines for initializing an SAA (Kim, Pasupathy, and Henderson, 2015; Linderoth, A. Shapiro, and Wright, 2006), finding the best mix of these values for starting the algorithm can be done using a trial-and-error

approach (Linderoth, A. Shapiro, and Wright, 2006). In our problem, it is particularly important to find a balance between large initial values of N, N', M, and the algorithm's run time. Starting the algorithm with large values of N and M causes the solution time to increase drastically, while with lower values of N and M the algorithm has to do more iterations to satisfy the stopping criterion and may end up with running out of RAM memory. We found these parameters through trial-and-error. More details on this would be presented in Chapter 4.

A final note on the SAA algorithm is about variance reduction for the objective function value across the M replications. Besides all benefits of the SAA method, there could still exist problems with very slow rates of convergence (Homem-de-Mello and Bayraksan, 2014). In other words, the solution can take a very long time to converge. We initially faced such a problem when working with a large set of variables. Monte Carlo sampling-based approximations and algorithms can also be significantly improve by reducing the variability of the estimates they generate (Homem-de-Mello and Bayraksan, 2014). There are ways of reducing the variance and thus improving the performance of the solution approach. These techniques are known as Scenario Reduction techniques (Homem-de-Mello and Bayraksan, 2014). Homem-de-Mello and Bayraksan, 2014, mention some of these techniques and how one can use them. Here is a brief summary of these methods.

1. The Antithetic Variates (AV) method aims to reduce variance by inducing correlations between solutions.

2. The Latin hypercube sampling method is based on the idea of partitioning the sample space and fixing the number of samples coming from each component of the partition proportional to the probability of that component. This is done to ensure that the number of sampled points from each region will be approximately equal to the expected number of points to fall in that region.

3. The Quasi-Monte Carlo method removes the randomness from a Monte Carlo Simulation.

In numerical analysis, the quasi-Monte Carlo method is a sampling technique which chooses samples on a non-random based. This is in contrast to the regular Monte Carlo Simulation, which is based on sequences of Pseudorandom numbers .

While we did not use any of these techniques, knowing them might be useful for future applications of the proposed model.

Chapter 3

Research Contributions

This chapter situates our research in the literature. We categorize the relevant literature into three areas: Predictive Models, Staff Scheduling IP Models, and Stochastic Optimization. Stochastic optimization, which is a relatively new area of research (Keller and Bayraksan, 2009), forms a less extensive part of the literature compared to the other areas. The literature on almost all aspects of the problem that we tackle is pretty vast. Various settings similar to inpatient clinics have been studied and multiple methodologies have been applied with a variety of results. Now, the question that remains to be answered is: What is the contribution of this research? First of all, it is important to note that our research strategy is part of a wider problem solving approach which seeks to find a solution to a specific problem. Thus, finding research gaps and working towards addressing them is a secondary priority. However, like for any other research work, new areas of interest could be better recognized in the light of the proposed solution approach. Here is a summary of our contribution to the literature on different areas:

1. **Stochastic Optimization Application in Care-Providers Scheduling**
Despite the long tradition of using IP and MIP models to address scheduling problems, Stochastic Optimization approaches have been significantly less applied to this type of problems (Leeftink, Vliegen, and Hans, 2019). To the best of our knowledge, while Mathematical Programming has been widely applied to Nurse Scheduling problems (Sitompul and Randhawa, 1990), applications of Stochastic Optimization to nurse scheduling (and care-providers scheduling) are much less common. Even fewer studies have applied Sample Average Approximation algorithms to Nurse Scheduling problems in inpatient settings. Among those, to our knowledge, only a few have applied SAA to an inpatient facility (Legrain, Omer, and Rosat, 2018; Bagheri, Devin, and Izanloo, 2016).

2. **Multiple Patient Types**
Considering more than one patient type is relatively new. Although this leads to

complex calculations and model dynamics, it helps us capture the number of patients and their requirements more accurately because it allows us to model arrivals and discharges for each patient group separately. It also allows us to simulate much more realistic scenarios.
This aspect of our model also helps us match more accurately patients to care-providers. Broyles, Cochran, and Montgomery, 2010, list a unique patient type as one of the main limitations of previous studies.

3. **Bayesian Probability Updating**
The second part of Section 7 discusses how the discharge probabilities in our model can be updated in the light of new information to enhance its accuracy. This is mathematically straightforward and important because it keeps one of the most important parameters of the model (the discharge probability for each patient group) up to date. Doing so we enhance the model's accuracy over time without increasing its complexity and number of calculations, unlike in many previous studies in which complexity is high (Littig and Isken, 2007) and (Broyles and Cochran, 2009).

All being said, the main contribution of this study relies on combining some methodologies to solve a specific group of scheduling problems at inpatient clinics. Our study stands in the intersection of these realms, namely optimization, predictive modelling, and simulation. Figure 3.1 shows these relationships.

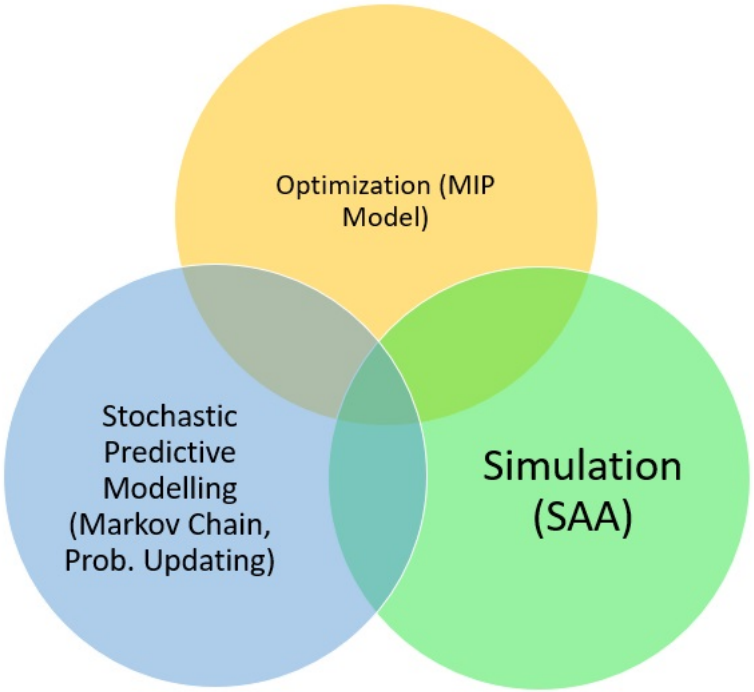

FIGURE 3.1: A big picture of the methodological approach used in this research

Chapter 4

Methodological Framework

This chapter describes the steps we took to develop a solution approach for the problem under study. First, a more detailed description of the problem and how the system works is presented in Section 4.1. Then, in Section 4.2, we describe how we have modelled the system and how the model is capable of capturing the parameters that represent the behaviour of the system in reality. The next section explains our solution strategy and finally, the last section of this chapter is a description of different ways in which the solution that we have proposed can be validated.

4.1 Problem Description

The problem we address is a typical Staff Scheduling problem at an inpatient clinic. For a better understanding of how a regular Scheduling process in a healthcare facility works and what we mean by that, one can look at Cheang et al., 2003.
We consider an inpatient clinic with a limited and constant number of beds. The capacity of the clinic does not change in the planning horizon and those patients who are blocked from being admitted remain in a wait list until a bed is available for them. In the case when there is no wait list (i.e., when the wait list capacity is 0), blocked patients seek help somewhere else and exit the system.

4.1.1 Patient and Care-Provider Groups

A number of care providers serve patients at the clinic. We consider multiple types of patients as well as different specialities for care providers. Grouping care providers can be done based on their specialities or the type of care they provide. However, defining a grouping system for patients is a more difficult task.
As previously mentioned, we need to cluster patients into meaningful groups to better map their characteristic to their care requirements. There are multiple options to define groups and assign patients to those groups. For example, one could simply categorize patients based on the intensity of care they need. As an example of this method, one can consider the Canadian Triage and Acuity Scale (CTAS) that is currently used to triage patients at Canadian emergency departments (J Murray, 2003).

The triage system categorizes patients by both possible injuries and/or physiological findings and ranks them into 5 groups (Emergency, Urgent, Less Urgent, and Non Urgent) with severity levels from 1 to 5 (1 being highest)(J Murray, 2003). More complex classification systems have also been applied. For example, El-Darzi, Abbi, et al., 2009, present a comprehensive study in which patient groupings are based on their LOS, and that could be used as a proxy for measuring the amount of resources patients consume.

While the patient groups created by any approach can be incorporated directly into our model, we also propose a structured way of clustering patients based on K-means Clustering if patient data is available.

K-means clustering is a clustering method that aims to partition n observations into k clusters or groups in which each observation belongs to the cluster with the nearest mean, serving as a representative of the cluster (Likas, Vlassis, and Verbeek, 2003). K-means clustering is an unsupervised partitioning technique. It tries to cluster vectors into a meaningful number of clusters based on minimizing the total distance between each point and its centroid. Any clinical information that is in the form of quantitative values or can be converted to quantitative values can be used as a part of the algorithm. Using this algorithm, when new dimensions are added to the data is not time-consuming and is a straightforward task. Algorithm 1 shows the K-means clustering algorithm we propose.

Algorithm 1: K-means Clustering Algorithm for Defining Patient Groups

input : A set of quantitative patient data (e.g., CAT score, age, expected LOS, wait-time, etc.), a set of k initial group centres

output: Clusters means, standard deviations, and bounds, a cluster assigned to each patient ID

1. Step 0. Start with the initial values for cluster centres.
2. Step 1. For each patient ID, find the closest cluster centre.
3. Step 2. Replace each centre point with the average over the data points in each cluster.
4. Step 3. Iterate 1,2 until convergence of the algorithm (i.e., until cluster centres do not change).
5. Step 4. Report clusters means, standard deviations, and a cluster assigned to each patient ID.

We also propose 4 features that can be a basis for grouping patients and can be used in the clustering algorithm.

1. Risk Assessment: different patients may have different diseases and, based on their symptoms and physical/mental injuries, can be grouped into some meaningful clusters. The amount of risk to a patient's health can be quantified by applying a risk assessment system such as the CTAS (J Murray, 2003). It can also be adapted

to each specific setting as well. For example, Balaratnasingam et al., 2011 present a system for mental health risk assessment. Similar approaches for other settings can be developed as well.

2. Care-Provider Needs: patient needs can be quantified in terms of care provider-to-patient ratio. There are scheduling and staffing standards as well as workplace regulations that can be used to find the specific care-provider needs of patients. One can look at Spetz et al., 2008 for a study of these standards. In addition, collective agreements between care-providers unions and hospitals/clinics can be used as another source of these information.

3. Wait Time: the amount of time a patient waits before being admitted to the clinic. Ideally, with all the other criteria being the same, the more a patient waits he/she has a higher priority of admission to the clinic. Wait time data can be easily extracted from clinical data and act as an input for the clustering algorithm.
Figure 4.1 depicts the factors mentioned above and their relationships.

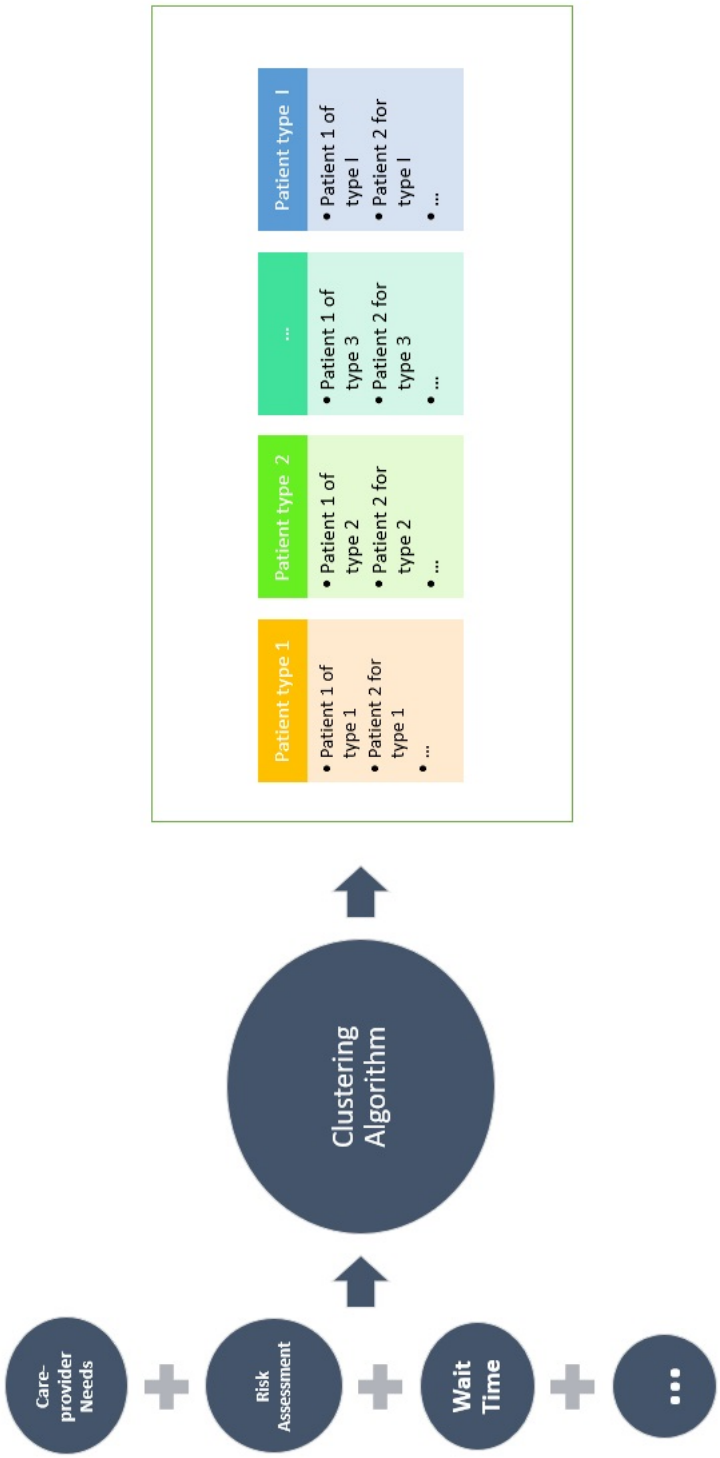

FIGURE 4.1: The proposed clustering algorithm for grouping patients into patient groups

One can consider Xu, Wong, and Chin, 2014 as an example of using K-means clustering to create patient groups. In this paper, the authors use several quantitative and qualitative methods to find patients with similar characteristics. Their results

are valuable for two reasons; not only do they show that clustering patients with similar characteristics ends up in groups with similar medical procedures, but their work also demonstrates the superiority of K-means over other clustering techniques.

4.1.2 Arrivals and Discharges

We assume two main arrival types to the inpatient clinic:
1. Arrivals through the Emergency Department
2. Arrivals via referrals

In our model, X_{ijk} represents the number of patients of type i who are in the clinic on shift k of day j.

We assume a single arrival rate which combines these two arrival types into one single arrival process. This is a reasonable assumption knowing that these are the only ways of being admitted to the inpatient clinic. If arrivals to a clinic are from only one of these two demand streams, one can simply consider a zero rate for the other arrival type.

We assume the arrival process for patients of type i follows a Stochastic Process with a specific probability distribution such as Poisson, Uniform, or Negative Binomial. This is based on the specific distribution the arrivals follow, and the simulation accordingly. For now, and for the sake of time and space, we only show the case where arrivals of patients of type i in shift k of day j follow a Poisson distribution with rate λ_{ijk}. One could also model a constant arrival rate for each patient type over the planning horizon. However, modelling arrivals for each shift separately lets us consider a different arrival process for each shift. This is helpful since, for example, arrivals during night shifts are undoubtedly less than arrivals during the day, and/or arrivals over weekends are typically fewer than during weekdays.

After patients arrive to the clinic, they receive treatment. Each treatment requires different number and mix of care providers. We have already explained how these treatments are modelled, as care-provider needs, in subsection 4.1.1. After completing their treatment, patients are discharged from the clinic. Modelling discharges is equivalent to modelling LOS. We assume that for each patient type i the length of stay follows a Geometric probability distribution with parameter p_{ijk}. Parameter p_{ijk} represents the probability that a patient of type i remains in the clinic after shift k of day j. Another way of looking at this is through a Binomial probability distribution. Assuming a Geometric LOS implies that the number of patients of type i who remain in the clinic after shift k of day j follows a Binomial distribution with parameters $n = X_{ijk}$ and $p = p_{ijk}$. We denote this number by R. For example, R_{ijk} represents the number of patients of type i who remain in the clinic after shift k of day j.

When a patient arrives at the clinic but the system is full he/she needs to wait for admission. We represent the length of the wait list by WL. WL_{ijk} represents the number of patients of type i who are in the wait list at the beginning of shift k of day

j.
We also assume a constant capacity (C) for the clinic.
As mentioned before, an inpatient clinic can be staffed with care providers of different specialities. We consider a different mix and number of care providers for different scenarios. Modifying the model for the case of an inpatient clinic with a different mix of care providers does not change the generality of the equations and can be easily done.
Figure 4.2 depicts a representation of how the inpatient clinic works at a high level.

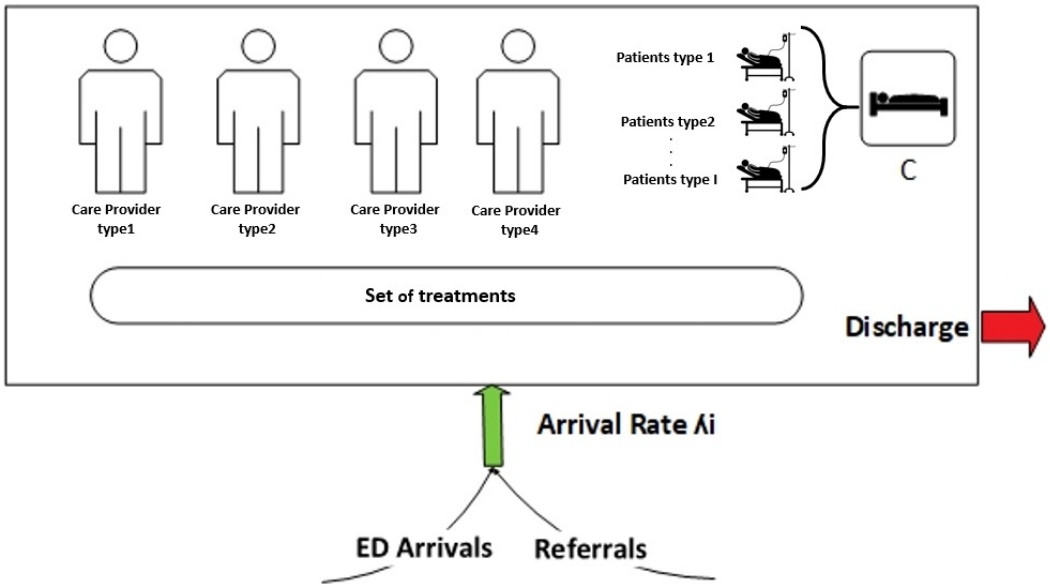

FIGURE 4.2: Arrivals, discharges, treatments, and care providers in an inpatient clinic at a general level

4.1.3 Planning Horizon

Scheduling decisions must be made ahead of time before actual arrivals and discharges happen. Decisions are made over a "Planning Horizon" of length J days. If, for example, $J = 14$ it essentially means that the model needs to generate demand scenarios for the next 14 days and based on those scenarios a schedule of 14 days must be determined.

4.2 Modelling Clinic's Utilization, Arrivals, and Discharges

This section discusses how we have developed a mathematical representation of the clinic in order to be able to capture the essential characteristics of its dynamics including patient flow through the system, arrivals, discharges, LOS, care providers requirements, and clinic capacity in order to determine the required number of staff. These parameters are embodied in a predictive model that assumes that the number

of patients on each shift can be modelled and predicted by knowing the number of patients on the most recent shift. Thus, the number of patients on the evening and night shifts are predicted from the number of patients on the morning and evening shifts, respectively, and the number of patients on the morning shifts are predicted from the number of patients on the night shift. The output of this model is a set of scenarios representing the possible future incoming demand over the planning horizon. Each scenario represents a mix of patients in the clinic on each shift of each day of the planning horizon.
The rest of this section discusses the mathematical details of this predictive model. Note that a complete mathematical notation is presented in the "Abbreviations" and "Symbols" section at the beginning of the thesis.

The following indices are defined:
The index i represents patient types. We assume that there are I patient types.

The index j represents a day of the planning horizon. The planning horizon consists of J days.

The index k represents a shift on a day. Shifts are being modelled as follow:
$k = 1$ for the Day Shift, $k = 2$ for the Evening Shift, and $k = 3$ for the Night Shift.

The index s represents a care provider. We assume there are a total of S care providers in the clinic. So, $s \in \{1, 2, ..., S\}$. In the case of different care-provider types, (e.g., different specialities) one can order care providers by group. For example, consider the case where the inpatient clinic is staffed with two types of staff, say 10 Nurses and 5 Social Workers. We can let $s \in \{1, 2, 3, ..., 10\}$ represent the Nurses and $s \in \{11, 12, ..., 15\}$ represent the Social Workers. Modifying the staff types does not change the generality of the formulation.
We assume discharges can only happen at the end of each shift. If, a specific clinic does not allow discharges during some shifts, the model can be modified without loss of generality.
Thus, the total number of patients on shift k of day j is equal to:

$$X_{jk} = \sum_{i=1}^{I} X_{ijk} \quad \forall j, k \tag{4.1}$$

This is the total occupancy of the clinic which must always be less than or equal to C, the number of beds in the clinic.
Second, we assume that both types of arrivals (ED and referrals) constitute a single arrival process following a Poisson distribution. We also assume that admissions happen at the beginning of each shift and patients cannot leave the clinic on the

same shift that they have been admitted, which is a reasonable assumption in inpatient hospital settings (Broyles, Cochran, and Montgomery, 2010). A_{ijk} represents the number of arrivals of type i on shift k of day j. If things are different in a specific clinic, for example if no arrival happens during the night shift, one can simply change the arrival rate for that shift. Thus, we have:

$$A_{ijk} \backsim Poisson(\lambda_{ijk}) \tag{4.2}$$

Third, we need to characterize patients more specifically. We assume that for any patient of type i who is in the clinic during shift k of day j there is a probability p_{ijk} that he/she will remain in the clinic for the next shift. The p_{ijk} probabilities can potentially come from historical data. Alternatively, one can use predictions from subject matter experts (potentially physicians responsible for patients). In both cases, historical data and predictions, can be updated regularly to enhance accuracy. We elaborate more on this later in Chapter 7.
We denote the number of patients of type i who remain in the clinic for shift k of day j by R_{ijk}. We have:

$$R_{ijk} \backsim Bin(X_{ijk}, P_{ijk}) \tag{4.3}$$

Also, we define R_{jk} as follow:

$$R_{jk} = \sum_{i=1}^{I} R_{ijk} \quad \forall j, k \tag{4.4}$$

4. The next step is to find the relationship between the number of patients in the next shift and the number of patients in the current shift. As mentioned earlier, we have considered I different types of patients. We described patient groups in section 4.1.1. We assume that there is a descending priority for these groups. Ideally, there are some clinical methods that can be used to prioritize patients. For example, one can use the ICD-10 coding system to find patient groups (Organization), 1993). Patient groups can also be modified easily for different settings.
If we consider a higher priority for arrivals of type i over type i', we admit patients of type i' only after assigning enough capacity to type i arrivals. After observing the number of patients remaining in the clinic for the next shift, we have a grasp of the number of beds available to be assigned to the new arrivals. Next, we start allocating empty beds to patients in a descending priority.
Assigning capacity to new arrivals only occurs when all patients in the wait list have

been assigned to beds. In other words, within a specific priority group, patients who have waited for services are assigned to the beds before those just arriving to the clinic. Obviously, those new arrivals blocked from entering the ward will enter the wait list and will stay there waiting until a bed becomes available.
For the length of the wait list on shift k of day j we have:

$$WL_{jk} = \sum_{i=1}^{I} WL_{ijk} \quad \forall j,k \tag{4.5}$$

The relationship between arrivals, discharges, and the wait list from one shift to the next is modelled below. In order to reduce complexity, we divide the formulation into two parts. Equation 4.6 shows how the model predicts X_{ijk} for the evening and night shifts (i.e., when $k = 2,3$) from the morning and evening shifts of the *same* day (i.e., when $k = 1,2$), respectively. Equation 4.7, on the other hand, shows how the value of X_{ijk} for the morning shift (i.e., $k = 1$) is predicted based on the value from the night shift of the previous day. Thus, every time a simulation is run, after all stochastic values are generated, the model predicts the census of the clinic for the next shifts, one after another, using the following two equations.

$$\begin{aligned}
&When\ k\ \leq 2\\
&if(C - R_{j(k+1)} - min_{j(k+1)} \leq WL_{ijk})\{\\
&X_{ij(k+1)} = R_{ij(k+1)} + (C - R_{ij(k+1)} - min_{j(k+1)})\\
&WL_{ij(k+1)} = WL_{ijk} + A_{ij(k+1)} - (C - R_{j(k+1)} - min_{j(k+1)})\\
&\}\\
&else\ if(C - R_{j(k+1)} - min_{j(k+1)} - WL_{ijk} \leq A_{ij(k+1)})\{\\
&X_{ij(k+1)} = R_{ij(k+1)} + (C - R_{j(k+1)} - min_{j(k+1)} - WL_{ijk})\\
&WL_{ij(k+1)} = A_{ij(k+1)} - (C - R_{j(k+1)} - min_{j(k+1)} - WL_{ijk})\\
&\}\\
&else\ \{\\
&X_{ij(k+1)} = R_{ij(k+1)} + A_{ij(k+1)} + WL_{ijk}\\
&WL_{ij(k+1)} = 0\\
&\}\\
&min_{j(k+1)} = min_{j(k+1)} + X_{ij(k+1)} - R_{ij(k+1)}
\end{aligned} \tag{4.6}$$

Before explaining equations 4.6 and 4.7, one should note that min_{jk} in these equations dynamically flags the changes in the number of patients of each type during the most recent shift, whether the difference comes from admitted arrivals, patients from the wait list being admitted to the clinic, or a combination of these or none. This term basically keeps track of the capacity allocated to patients of different types

so that no patient of lower priority gets admitted before patients of higher priorities. Also, note that these equations run for all patient groups (is) in the order of their priorities. So, each time capacity is being allocated, the min_{jk}s for each shift starts from 0 for the first priority group. Then the number of patients of each group that are admitted (by arrivals, wait list, or a combination of both) is added to min_{jk} every time one of the conditional formulas in equations 4.6 or 4.7 takes place. In this way, the capacity that has been allocated to other patients will not be re-assigned to patients with lower priority.

Equation 4.6 decides for the next evening or night shift. The first part of expression 4.6 determines the number of patients on the next evening or night shift when the available capacity ($C - R_{j(k+1)} - min_{j(k+1)}$) is less than the number of patients in the wait list (WL_{ijk}). In this case, available capacity will be allocated to those in the wait list.

The second part of expression 4.6 determines the number of patients in the clinic when there is some free capacity in the clinic and this capacity needs to be allocated between the patients on the wait list and new arrivals (i.e., $A_{ij(k+1)}$). Observe that priority has been given to the patients on the wait list.

The last piece in expression 4.6 determines the number of patients on the next shift when, after allocating capacity to remaining patients and patients in the wait list, there is still some room for arrivals. Thus, the wait list becomes empty.

$$
\begin{aligned}
&When\ k = 3 \\
&if(C - R_{(j+1)1} - min_{(j+1)1} \leq WL_{ijk})\{ \\
&X_{i(j+1)1} = R_{i(j+1)1} + (C - R_{i(j+1)1} - min_{(j+1)1}) \\
&WL_{i(j+1)1} = WL_{ijk} + A_{i(j+1)k} - (C - R_{i(j+1)1} - min_{(j+1)1}) \\
&\} \\
&else\ if(C - R_{(j+1)1} - min_{(j+1)1} - WL_{ijk} \leq A_{i(j+1)k})\{ \\
&X_{i(j+1)1} = R_{i(j+1)1} + (C - R_{(j+1)1} - min_{(j+1)1} - WL_{ijk}) \\
&WL_{ij(k+1)} = A_{ij(k+1)} - (C - R_{j(k+1)} - min_{j(k+1)} - WL_{ijk}) \\
&\} \\
&else\ \{ \\
&X_{i(j+1)1} = R_{i(j+1)1} + A_{i(j+1)1} + WL_{ijk} \\
&WL_{i(j+1)1} = 0 \\
&\} \\
&min_{j(k+1)} = min_{j(k+1)} + X_{ij(k+1)} - R_{ij(k+1)}
\end{aligned}
\tag{4.7}
$$

Equation 4.7 follows the exact same logic as Equation 4.6. The only difference here is that in Equation 4.7, instead of $k = 1, 2$, we have $k = 1$ which represents the next

morning shift and instead of j we consider the next day $j + 1$.
Figure B.1 shows a graphical representation of the flow conservation equations 4.6 and 4.7.

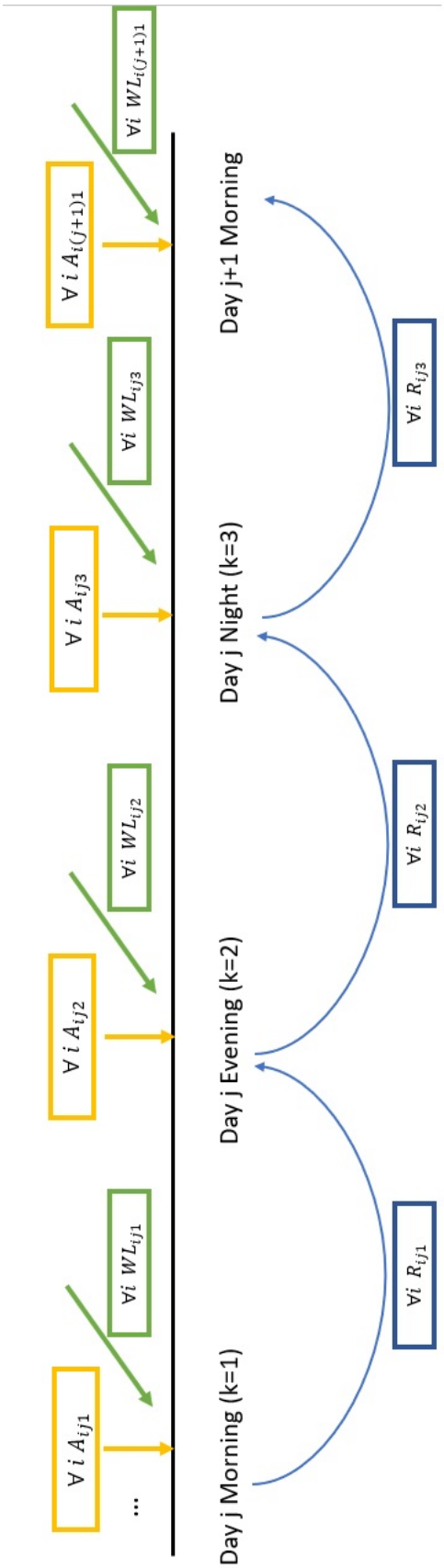

FIGURE 4.3: A timeline of arrivals, discharges, wait list admissions, and remaining patients in the clinic (patient flow conservation equations)

4.3 States of the System and Transition Probabilities

As mentioned earlier, we consider a Markov model where the number of patients of each type defines the state of the system and the transition probabilities depending on the most recent state (Figure B.1). For each patient type i one can find the transition probability from census m to census n as follows:

$$\begin{aligned} P^i_{mn} =& P\{X_{ij(k+1)} = m | X_{ijk} = n\} \\ =& P\{X_{ij(k+1)} = m | X_{ijk} = n, \sum_i^I X_{ijk} = X_{jk}, A_{i,j(k+1)} \backsim Poi(\lambda_i), R_{ij(k+1)} \backsim Bin(X_{ijk}, p_{ijk})\} \end{aligned} \tag{4.8}$$

Note that Equation 4.8 only works when $k = 1, 2$. When $k = 3$, one can easily rewrite Equation 4.8 with the exact same logic.
In order to fully characterize Equation 4.8, we assume that for each patient type arrivals follow a Poisson distribution and the number of remaining patients in the clinic follows a binomial distribution (i.e., we assume Geometric length of stay). Equation 4.9 describes the probability in 4.8 for the number of patients type i:

$$
P^{i}_{nm} = \begin{cases}
if\ C - R_{(j+1)i} - min_{(j+1)i} \leq WL_{ijk} \rightarrow \\
\\
P\{R_{ij(k+1)} = m\} = \binom{n}{n-m} p_{ijk}^{n-m} (1 - p_{ijk})^{n-(n-m)} = \\
\binom{n}{m} p_{ijk}^{n-m} (1 - p_{ijk})^{m} \\
\\
\\
if\ C - R_{(j+1)i)} - min_{(j+1)i} - WL_{ijk} \leq A_{i(j+1)k} \rightarrow \\
\\
P\{R_{ij(k+1)} = m - (C - R_{jk})\} = \binom{n}{m-(C-R_{jk})} p_{ijk}^{m-(C-R_{jk})} (1 - p_i)^{n-m+(C-R_{jk})} \\
\\
\\
if\ A_{ij(k+1)} \leq C - R_{(j+1)i)} - min_{(j+1)i} - WL_{ijk} \rightarrow \\
\\
P\{R_{j(k+1)} + A_{ij(k+1)} = m\} = \sum_{\alpha=0}^{C} P\{R_{j(k+1)} = m - \alpha | A_{ij(k+1)} = \alpha\} \times P(A_{ij(k+1)} = \alpha) \\
\\
\\
= \sum_{\alpha=0}^{C} P\{R_{j(k+1)} = m - \alpha\} \times P\{A_{ij(k+1)} = \alpha\} \\
\\
\\
= \sum_{\alpha=0}^{C} \binom{C}{m-\alpha} p_{ijk}^{m-\alpha} (1 - p_{ijk})^{C-(m-\alpha)} \times \frac{e^{\lambda_{ijk}} \lambda_{ijk}^{\alpha}}{\alpha!}
\end{cases}
\tag{4.9}
$$

To explain Equation 4.9, please note that these are evaluated in the order of priorities. The first condition in expression 4.9 occurs when the number of patients of type i in the wait list is higher than the available capacity (which is determined after allocating capacity to remaining patients and patients waiting of higher priority). In this case, the new capacity is allocated to patients of type i waiting.
The second part of Equation 4.9 models the situation where beds are allocated to all patients of higher priority, all remaining patients, and the patients of the same priority in the wait list. Thus, the remaining capacity is assigned to new arrivals.
The last part in Equation 4.9 occurs when we have allocated capacity to all higher priorities and the remaining capacity is less than all arrivals but more than all remaining patients. Thus, this remaining capacity needs to be allocated to both remaining patients and new arrivals.
One can use these equations with a set of random numbers to find the corresponding probabilities for each transition. However, for most scenarios (i.e., for most possible combinations of different arrivals and discharge distributions), these probabilities

are very small. Thus, they are not useful for enumeration techniques unless for low capacity clinics (C less than 2). We can simulate patient flow equations for some scenarios of arrivals and discharges but there are still too many possible future states for each current state of the system that no decision could be made based on them. On the one hand, the higher the capacity of the clinic, the arrival volume, or the rate of the LOS, the more the number of possible states would be. This causes the probability matrix to be intractable. Additionally, the volume of the calculations could cause intractability as well. The main reason we have presented these equations here is first, the readers can observe their complexity and the fact that they are practically very small and second, because in some small cases enumeration techniques can be used and simulated based on these equations.
To address these problems and to find good staff schedules, we apply a Stochastic Optimization technique called Sample Average Approximation(SAA)and combined the Simulation and the MIP model to find the optimal mix of staff.
As mentioned in Chapter 2, when the number of possible scenarios for a Stochastic Optimization problem is very large, or even infinite, it becomes convenient to use Monte Carlo Simulation to find a good approximation of the optimal mix of staff. More details on SAA are presented in 4.5.
The Monte Carlo Simulation was implemented in *Java*. Each time the algorithm runs, it generates discrete random numbers of the arrivals and discharges in each shift based on the distributions of arrivals and discharges used in equations 4.6 and 4.7.

4.4 MIP Staff Scheduling Model

In this section, a Mixed Integer Programming model is introduced to complement the predictive model. The complete list of decision variables and parameters for this MIP model can be found in the corresponding section at the beginning of this document. Here we describe the model by explaining its decision variables, objective function, and constraints. Please note that we may only considered a general model without considering specific constraints that vary based on each inpatient clinic. We discussed a handful of potential additional constraints in Appendix A.

4.4.1 Decision Variables

$$S_{sjk} = \begin{cases} 1, & \text{if care provider s is assigned to shift k of day j} \\ 0, & \text{otherwise} \end{cases} \tag{4.10}$$

Please note that for simplicity of notation, one can group care-providers of a similar type. Here we assume a total of S care providers. Thus, $s \in \{1, 2, ..., S\}$ and possibly $s \in \{1, ..., s'\}$ are of speciality 1, and $s \in \{s' + 1, ..., S\}$ are care providers of specialty 2 (or any other combination of number and type of care-providers). We also assign index d to refer to care provider types. We assume there are D care-provider specialities in the clinic. Thus, $d \in \{1, ..., D\}$.

Next, we present a set of slack variables that are used for transforming inequalities to equalities in the model. These variables represent possible values of under and over specific targets. A description of these targets is provided later on this chapter.

$$\delta^{-}_{\xi\, djk} = \text{overstaffing of care-provider type d on shift k of day j for scenario } \xi \quad (4.11)$$

$$\delta^{+}_{\xi\, djk} = \text{understaffing of care-provider type d on shift k of day j for scenario } \xi \quad (4.12)$$

The variables in equations 4.11 and 4.12 are associated with the mismatch between resources and demand for care. Ideally, both of these values should be equal to zero but in practice, for most scenarios, there are shifts for which the number of staff allocated to ($\sum_s S_{sjk}$) will be less than or equal to the amount needed to cover the demand. $\delta^{-}_{\xi\, djk}$ and $\delta^{+}_{\xi\, djk}$ are penalized in the objective function.

$$\Pi^{-}_{\xi\, s} = \text{Number of shifts more than the weekly target for care provider s in scenario } \xi \quad (4.13)$$

$$\Pi^{+}_{\xi\, s} = \text{Number of shifts less than weekly target for care provider s in scenario } \xi \quad (4.14)$$

Typically, there is a limit on the number of weekly shifts for each care-provider. If this number is exceeded, it is either against union rules or the clinic has to pay extra to care providers who work a number of shifts above the target value. Variables 4.13 and 4.14 determine these two values. These two variables are minimized in the objective function with respective weights.

$$\Lambda_{\xi}^{-} = \text{Length of the wait list more than its target in scenario } \xi \tag{4.15}$$

$$\Lambda_{\xi}^{+} = \text{Length of the wait list less than its target in scenario } \xi \tag{4.16}$$

The model assumes a target value for the ideal wait list size. Each time that the length of the wait list, which comes from the predictive model, is more than its target a penalty is considered in the objective function (obviously no cost is associated in the objective function with a lower than target wait list). Variables 4.15 and 4.16 represent this penalty. Note that these are nominal constraints and in fact no decision is made based on the value of 4.15 and 4.16.

$$\kappa_{\xi\, sj}^{-} = \text{Number of overtime daily shifts for staff s on day j of scenario } \xi \tag{4.17}$$

$$\kappa_{\xi\, sj}^{-} = \text{Number of under target daily shifts for staff s on day j of scenario } \xi \tag{4.18}$$

Variables 4.17 and 4.18 model the number of shifts covered by each care provider s, which could be more or less than the daily target. These number are also penalized in the objective function.

4.4.2 Sets and Parameters

The index s represents a care providers.

$$s = \begin{cases} 1 & \text{Care provider 1} \\ 2 & \text{Care provider 2} \\ \dots & \\ S & \text{Care provider S} \end{cases} \tag{4.19}$$

As mentioned before, index j represents a day of the planning horizon.

$$j = \begin{cases} 1 & \text{Day 1} \\ 2 & \text{Day 2} \\ 3 & \text{Day 3} \\ \dots & \\ J & \text{Day J} \end{cases} \tag{4.20}$$

And, index k is one of the following shifts.

$$k = \begin{cases} 1 & \text{Day Shift} \\ 2 & \text{Evening Shift} \\ 3 & \text{Night Shift} \end{cases} \tag{4.21}$$

$X_{\xi\ ijk}$ denotes the number of patients of type i on shift k of day j or scenario ξ. These values are obtained from the predictive model.

SR_{dik} is the ratio of care-provider type d to the number of patients type i on shift k. As mentioned before, shifts are necessary because typically, care-provider requirements differ during the day with possibly more care required during the morning shift and less care required during the evening and night shifts.

$Target_{weekly}$ is the weekly goal for the number of shifts covered by each care provider. This can be found from work regulations currently in place or from union agreements. This is a parameter so the manager/decision-maker for the clinic can set it based on their special needs.

$Target_{daily}$ is the daily goal for the number of shifts covered by each care provider. This can also be found from work regulations and/or collective agreements.

$Target_{waitlist}$ is the desired length of the wait list for the clinic. Its value can come from standards or guidelines in place at each facility. In Ontario, these goals are set by the Ministry of Health and Long-term Care for some care settings.

The following sets of parameters define the cost terms or penalties in the objective function. These could be in dollars or ideally in terms of ratios (i.e. relative values) . For example, one could assume a ratio of 1.4 for a regular shift with respect to an overstaffed shift. In this way, the model will minimize over staffing more than regular shifts.

C^{-}_{djk} = Cost associated with overstaffing care provider type d on shift k of day j (4.22)

C^{+}_{djk} = Cost associated with understaffing care provider type d on shift k of day j (4.23)

Parameters 4.22 and 4.23 represent the costs associated with over/under staffing levels.

C^{-}_{s} = Cost of an additional shift covered after fulfilling the weekly target value of shifts, for staff s (4.24)

C^{+}_{s} = Cost of an additional shift covered before fulfilling the weekly target value of shifts, staff s (4.25)

Parameters in 4.24 and 4.25 represent costs associated with the number of shifts fulfilled less or more than the weekly target shifts.

C^{-}_{sj} = Cost of an additional shift covered after fulfilling the daily target value of shifts, for staff s (4.26)

C^{+}_{sj} = Cost of an additional shift covered before fulfilling the daily target value of shifts, for staff s (4.27)

Parameters 4.26 and 4.27 represent the costs associated with the number of shifts extra or short from fulfilled the daily target values. Ideally, the penalty for extra shifts should be considered higher due to the fact that it is usually against collective agreements for care providers to work more than one consecutive shift. The under daily target the under daily target penalty can be thought as a way to avoid extreme minimization of the number of staff. However, when the latter is not a goal, one can consider C^{+}_{sj} equal to 0.

$$C^{-}_{WL} = \text{Penalty for when the length of the wait list is more than its target} \quad (4.28)$$

Finally, the parameter in 4.28 is the penalty for exceeding target wait list size. Note that unlike the other decision variables here, no cost is associated with a wait list of less than target. In other words, a wait list of less than target is not penalized. Please note that this is not added to the model with a negative cost which means that if the length of the wait list is less than what has been set for the target, there is no extra cost in the model.

4.4.3 Objective Function

Equation 4.29 shows the objective function for the MIP model.

$$\begin{aligned} Min \;\; Z = \mathbb{E}_{\tilde{\xi}}[& \sum_{djk} C^{-}_{djk}\delta^{-}_{\tilde{\xi}\,djk} + \sum_{djk} C^{+}_{djk}\delta^{+}_{\tilde{\xi}\,djk} + \sum_{s} C^{-}_{s}\Pi^{-}_{\tilde{\xi}\,s} + \sum_{s} C^{+}_{s}\Pi^{+}_{\tilde{\xi}\,s} + \\ & \sum_{sj} C^{-}_{sj}\kappa^{-}_{\tilde{\xi}\,sj} + \sum_{sj} C^{+}_{sj}\kappa^{+}_{\tilde{\xi}\,sj} + C^{-}_{WL}\Lambda^{-}_{\tilde{\xi}}] \end{aligned} \quad (4.29)$$

The objective function seeks to minimize the expected over and under staffing costs to the schedule. One can look at Equation 4.29 as a Two-stage Stochastic Programming model, where the first stage decisions are the schedule (found based on the problem constraints and not directly considered in the objective function) and the second stage seeks to minimize the expected cost of the schedule over all possible scenarios.
The first two terms inside the expectation in Equation 4.29 are associated with possible under/over staffing levels for each scenario. The next two terms are minimizing corresponding to the costs associated with the weekly targets for the number of shifts. The fifth and sixth terms correspond to the costs of violating the daily targets for the number of shifts. Finally, the last term is associated with penalties for penalties for exceeding the wait list target and penalties.

4.4.4 Constraints

We have considered four groups of constraints for the MIP model. Each one of these constraints represents a specific part of the problem.
1. *Demand Coverage*:
This set of constraints ensures the required level of care for each group of patients based on the demand.

$$\sum_{s} S_{sjk} + \delta^{+}_{\tilde{\xi}\,djk} - \delta^{-}_{\tilde{\xi}\,djk} = \sum_{i} SR_{dik} \times X_{\tilde{\xi}\,ijk} \quad \forall \tilde{\xi}\,, k, j, d \quad (4.30)$$

The right-hand side of Equation 4.30 is the total care required from all admitted patients during shift k of day j for scenario ξ. The left-hand side is a combination of the number of care providers during shift k of day j and the corresponding second stage costs of the decision S_{sjk}.

2. *Weekly workload*:
The second group of constraints seeks to meet the weekly targets for the number of shifts the care providers should cover.

$$\sum_{j,k} S_{sjk} + \Pi^{+}_{\xi\, s} - \Pi^{-}_{\xi\, s} = Target_{weekly} \quad \forall \xi\ , s \tag{4.31}$$

The left-hand side of this constraint considers the weekly number of shifts covered by each care provider based on decision variables S_{sjk} and the corresponding slack variables associated with it. The right-hand side is the target for the weekly number of shifts for each care-provider.

3. *Daily workload*:
The set of constraints 4.32 ensures that the daily target for the number of shifts is met.

$$\sum_{k} S_{sjk} + \kappa^{+}_{\xi\, sj} - \kappa^{-}_{\xi\, sj} = Target_{daily} \quad \forall \xi\ , s, j \tag{4.32}$$

The left-hand side involves the total number of shifts on a day for a specific care provider and the corresponding slack variables associated with it.

5. *Wait list target*:
The final set of constraints determine the cost associated with exceeding the target considered for the length of the wait list.

$$WL_{\xi\, J3} + \Lambda^{+}_{\xi} - \Lambda^{-}_{\xi} = Target_{wait\ list} \quad \forall \xi \tag{4.33}$$

The left hand side consists of the number of patients in the wait list on the last shift ($k = 3$) of the last day ($j = J$) of scenario ξ and the associated second-stage costs for a wait list of length above the target and the slack variable Λ^{+}_{ξ}.

4.5 Sample Average Approximation (SAA)

A detailed description of the general form of the SAA algorithm is presented in Chapter 2 (2.4.2). In summary, the of Sample Average Approximation is based on

the idea that the solution to an Expected Value Stochastic Problem can be approximated by solving the problem for a large enough random sample of possible scenarios. The SAA algorithm provides a robust platform to make sure that solving the expected value problem for a limited number of scenarios (instead of all possible realizations of the random parameters) provides sufficient accuracy for the original problem which consists of all the realizations of vector ξ. The algorithm eventually provides this sample size and one can solve a stochastic MIP problem based on this number.

4.5.1 Algorithm Description

We use one of the most common versions of the SAA algorithm which is based on the vanilla SAA algorithm mentioned and explained in Ahmed, A. Shapiro, and E. Shapiro, 2002. the Algorithm 2 below provides details of the algorithm we have developed for the scheduling problem described above (in 4.10 to 4.33).

Algorithm 2: Monte Carlo Optimization Algorithm based on an SAA

input : N_0 as the initial sample size, M as the number of replications in the algorithm, N_1 as the number of scenarios in the evaluation step, ϵ as the optimality gap tolerance, and α as the stopping criteria

output: N as the required sample size, $\overline{v}_{N_0}$, $\overline{v}_{N_1}$ as the lower and upper bounds on the optimal value of the equation 4.29, and AOI_N as Approximation Optimality Index

initialization Let $N = N_0$;

Step 1. Simulation Optimization

for $m \leftarrow 1$ **to** M **do**

  Step 1.1 Scenario Generation

  Generate N independent and identically distributed (*i.i.d.*) scenarios of $X_{\xi\ ijk}$; number of patients of type $i = 1, ..., I$ on shifts $k = 1, 2, 3$ of days $j = 1, ..., J$.

  Step 1.2 Solving the Expected Value problem

  Determine the expected objective function value (Equation 4.29) with the scenarios generated in Step 1.1 (consider equal probability for each scenario) and record the corresponding optimal objective value v_N^m and the optimal schedule S_{sjk}^m.

  Step 1.3 Evaluating the solution (schedule) via Monte Carlo Simulation

  Generate N_1 independent and identically distributed (*i.i.d.*) scenarios of the number of $X_{\xi\ ijk}$; patients of type $i = 1, ..., I$ on shifts $k = 1, 2, 3$ of days $j = 1, ..., J$.

  Use the schedule S_{sjk}^m and the scenarios from Step 1.3 to find an estimate of the second-stage costs associated with S_{sjk}^m in Equation 4.29 and let that be $v_{N_1}^m$.

Step 2. Compute the average of v_N^m and $v_{N_1}^m$ over all ms as follows:

$\overline{v}_N = \frac{1}{M} \sum_m v_N^m$ $\qquad\qquad$ $\overline{v}_{N_1} = \frac{1}{M} \sum_m v_{N_1}^m$

Step 3. Compute the Approximate Optimality Index as follow:

$AOI_N = \frac{\overline{v}_{N_1} - \overline{v}_N}{\overline{v}_{N_1}}$

Step 4. If AOI_N satisfies α (i.e. if AOI is less than ϵ) terminate and output the schedule S_{sjk},N, and AOI. Otherwise $N \leftarrow 2N$ and go to Step 1.

At each iteration, the Algorithm 2 first starts by letting the number of scenarios that are going to be determined by the algorithm (N) be equal to N_0. Next, it generates N random scenarios (samples) of the number of patients of each type on each shift and day in the planning horizon ($X_{\xi\ ijk}$). Then, in Step 1.2, the Sample Average Problem is solved with the following objective function:

$$Min \ \ \nu_N^m = \frac{1}{N}[\sum_{\xi\ djk} C^-_{djk}\delta^-_{\xi\ djk} + \sum_{\xi\ djk} C^+_{djk}\delta^+_{\xi\ djk} + \sum_{\xi\ s} C^-_s \Pi^-_{\xi\ s} + \sum_{\xi\ s} C^+_s \Pi^+_{\xi\ s} + \sum_{\xi\ sj} C^-_{sj}\kappa^-_{\xi\ sj} + \sum_{\xi\ sj} C^+_{sj}\kappa^+_{\xi\ sj} + C^-_{WL}\Lambda^-_\xi] \tag{4.34}$$

subject to constraints 4.30 to 4.33 for $\xi = 1, ..., N$. The algorithm records the corresponding objective function value ν_N^m and the optimal schedule S^m_{sjk}. In Step 1.3 the algorithm first generates N_1 scenarios of $X_{\xi\ ijk}$ and computes the average objective function for the N_1 scenarios. Equation 4.35 evaluates the schedule S^m_{sjk} obtained in step 1.2 of the algorithm.

$$\nu_{N_1}^m = \frac{1}{N_1}[\sum_{\xi\ djk} C^-_{djk}\delta^-_{\xi\ djk} + \sum_{\xi\ djk} C^+_{djk}\delta^+_{\xi\ djk} + \sum_{\xi\ s} C^-_s \Pi^-_{\xi\ s} + \sum_{\xi\ s} C^+_s \Pi^+_{\xi\ s} + \sum_{\xi\ sj} C^-_{sj}\kappa^-_{\xi\ sj} + \sum_{\xi\ sj} C^+_{sj}\kappa^+_{\xi\ sj} + \sum_{\xi} C^-_{WL}\Lambda^-_\xi] \tag{4.35}$$

After m iterations, the average of ν_N^m and $\nu_{N_1}^m$ is calculated in Step 2. In Step 3, the approximate optimality index which is an unbiased statistical estimator of the optimality gap for equation 4.29 is computed as follow:

$$AOI_N = \frac{\overline{\nu}_{N_1} - \overline{\nu}_N}{\overline{\nu}_{N_1}} \tag{4.36}$$

In Step 4, the algorithm checks the condition α which can be in form of the AOI being less than ϵ or it could be any other condition (for example, stop when AOI does not change in three iterations). If it is, then the algorithm terminates with N as the required number of samples for a robust answer to 4.29.
In terms of finding the actual schedule associated with the algorithm, there are two approaches. First, after determining N, one can solve the expected value problem once again with this N and find the schedule. Or, one can use one of the screening techniques in Kleywegt, A. Shapiro, and Homem-de-Mello, 2002 to find the best schedule among the M solutions. Choosing the answer associated with the best solution is a commonly used method (Kleywegt, A. Shapiro, and Homem-de-Mello, 2002).
[1] A flow chart of the SAA algorithm is presented in Appendix B.

[1] According to the experience I had with solving this problem, when the sample size meets the stopping criteria the standard deviation of the answers associated with M scenarios is very small that in fact it does not make much different which solution we choose.

4.5.2 Parameters

Based on the values of N_0, N_1, M, and ϵ there are different ways of initializing the algorithm. Ideally, one can try a combination of these numbers to find a good starting point. We tried different values and chose the best combination for each setting. The results for different trials is presented in the next Chapter.
There are two main observations to make about the parameters. First, usually N_1 is chosen to be higher than N_0. This is because a possibly lower sample size N_0 may be able to provide a good enough solution for a larger number of scenarios. Second, another helpful way of starting the algorithm is to find a good value of N_0 through a series of single runs with $M = 1$. In this way, the number of times samples are generated and the SAA formulation is solved is tremendously decreased.

Chapter 5

Test Data

Before describing the test data we used to evaluate the performance of the proposed approach, we need to explain why we are using artificial test data instead of real data. During the last few months, the world faced a universal pandemic caused by the outbreak of Coronovirus 19 (COVID-19). This pandemic caused many healthcare facilities including hospitals to prioritize their resources to essential activities . As a result, we did not have access to any real data to test the approach. Consequently, we created a test data set based on the problem faced by CHEO's inpatient mental heath wards and using it to evaluate the approach.

In this chapter, we define a set of parameters and datasets in order to be able to evaluate the proposed solution approach. In the subsequent chapter, we use the models developed in Chapter 4 to evaluate the suggested staffing decisions for the different problem settings defined in this chapter. The main goal of this chapter is to describe a set of realistic, comprehensive, and yet practical scenarios.
This chapter consists of five sections. The first section defines some clinical settings and four scenarios for arrivals for each setting. The second and third sections describe the characteristics of the care providers we consider for each model. The third section describes the patient groups we considered. The fourth section, elaborates the time horizon and shift we considered and finally, the last section is about the parameters we used for the SAA algorithm.

5.1 Clinical Settings

The goal was to consider a comprehensive range of settings and scenarios to can make sure that the model had a good performance under different conditions, and to compare the model's performance for different scenarios. We considered three different inpatient clinics. These clinics differ in terms of capacity (i.e., number of beds), mix of care providers, number of care providers, presence (or absence) of a wait list, and the required ratio of care-provider to patients.
We consider four arrival distributions for each clinic. For each setting, the first two

scenarios simulate a Uniform Arrival Process and the third and fourth scenarios follow Poisson Arrival Processes. The first and third arrival distributions have lower rates than the second and the fourth distributions. This pattern is repeated for all three settings generating a total of 12 scenarios.

5.1.1 Setting 1

As one can see in Table 5.2, the first four scenarios are associated with the first setting which is the smallest care unit that we have considered to evaluate the performance of the solution approach. This setting only has five beds. There is no wait list for patients who are not admitted to this clinic immediately. In other words, patients are either admitted to the clinic or receive care somewhere else. This could be considered a model for critical care services (Marino and Sutin, 1998).
There is only one type of care provider in this clinic, and 15 individuals of this type are providing services to patients. The required ratio of care providers to patients is 1 to 2 for all shifts and all patient types. This ratio is reasonable in an Intensive Care Unit (Marino and Sutin, 1998) [1]. Scenarios 1 to 4 are based on this setting.

5.1.2 Setting 2

The second setting has a higher capacity with 10 beds. Patients who are not admitted to the clinic immediately join a wait list. Patients who have waited the longest have priority over the others to be admitted to the clinic. The MIP model considers a wait list target and second stage costs with respect to the wait list target.[2]

Considering a wait-list target corresponds to the negative outcomes of people waiting long times for care. Fomundam and Herrmann, 2007, mention some of these possible consequences. This setting also has 1 type of care-provider with 15 individuals.
We consider 2 patient groups in this setting, patient group one has a care-provider to patient ratio of 1 to 2 for morning and evening shifts and a ratio of 1 to 4 for night shifts. Patients in group two have a ratio of 1 to 4 in the morning and evening shifts but they do not need any care-providers of type 2 and 3 during the night.

[1] For a more technical justification on why this ratio makes sense and could be appropriate see Amaravadi et al., 2000

[2] These costs were explained in Chapter 4. One can consider this cost as associated with the consequences of having a long wait list.

TABLE 5.1: List of required resources for different patient types

Patient Type	Care-provider 1			Care-provider 2			Care-provider 3		
	Morning	Evening	Night	Morning	Evening	Night	Morning	Evening	Night
Type 1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
Type 2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{1}{16}$	$\frac{1}{16}$	0
Type 3	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{1}{16}$	$\frac{1}{16}$	0

5.1.3 Setting 3

The third setting is the largest one with 20 beds. There is a wait list for blocked arrivals. This setting is staffed with three types of care providers: type 1, type 2, and type 3. There are 15 individuals of type 1, 10 of type 2, and 5 of type 3.
We consider three patient types. Table 5.1 shows the required resources for these patient types.

For patient type 1, the ratio of care-provider type 1 and 2 to patient is considered to be 1:2 across all shifts except for the one associated with care-providers of type 2 and the night shift. Based on the 2016 Annual Report of the Auditor General of Ontario, the average ratio of nurse to patients in Ontario is around 1:4, with hospitals with ratios as low as 1:2 (Intensive Care Units) and as high as 1:6 or 1:8 (Community Hospitals) (Auditor-General, 2016). Scenarios 9 to 12 are associated with this setting.

Table 5.2 provides a description of the characteristics of the 12 scenarios considered in this analysis.

TABLE 5.2: List of scenarios and their parameters

Scenario		Number of beds (C)	Wait-list (WL)	Staff Types (d)	Number of Staff (S)	Staff to Patient Ratio (SR_{ik})	Arrivals (A_{ijk})
1	Setting 1	5	No	1	15 type 1	$\frac{1}{2} \forall i,k$	Uniform([1,3],[1,2],[1,2],[1,2]) $\forall j,k$
2	Setting 1	5	No	1	15 type 1	$\frac{1}{2} \forall i,k$	Uniform([5,7],[3,5],[3,5],[3,5]) $\forall j,k$
3	Setting 1	5	No	1	15 type 1	$\frac{1}{2} \forall i,k$	Poisson(2,1,1,1) $\forall j,k$
4	Setting 1	5	No	1	15 type 1	$\frac{1}{2} \forall i,k$	Poisson(5,4,4,4) $\forall j,k$
5	Setting 2	10	Yes	1	15 type 1	$\frac{1}{2} \forall i = 1, k = 1,2$ and $\frac{1}{4} \forall i = 1, k = 3$ $\frac{1}{4} \forall i = 2, k = 1,2$ and $0 \forall i = 2, k = 3$	Uniform([1,3],[1,2],[1,2],[1,2]) $\forall j,k$ $\forall j,k$
6	Setting 2	10	Yes	1	15 type 1	$\frac{1}{2} \forall i,k = 1,2$ and $\frac{1}{4} \forall i,k = 3$ $\frac{1}{4} \forall i = 2, k = 1,2$ and $0 \forall i = 2, k = 3$	Uniform([5,7],[3,5],[3,5],[3,5]) $\forall j,k$ $\forall j,k$
7	Setting 2	10	Yes	1	15 type 1	$\frac{1}{2} \forall i,k = 1,2$ and $\frac{1}{4} \forall i,k = 3$ $\frac{1}{4} \forall i = 2, k = 1,2$ and $0 \forall i = 2, k = 3$	Poisson(2,1,1,1) $\forall j,k$ $\forall j,k$
8	Setting 2	10	Yes	1	15 type 1	$\frac{1}{2} \forall i,k = 1,2$ and $\frac{1}{4} \forall i,k = 3$ $\frac{1}{4} \forall i = 2, k = 1,2$ and $0 \forall i = 2, k = 3$	Poisson(5,4,4,4) $\forall j,k$ $\forall j,k$
9	Setting 3	20	Yes	3	15 type 1, 10 type 2, and 5 type 3	$\frac{1}{6}$ of type 1, $\frac{1}{8}$ of type 2, and $\frac{1}{8}$ of type 3 $\forall i,k$	Uniform([1,3],[1,2],[1,2],[1,2]) $\forall j,k$
10	Setting 3	20	Yes	3	15 type 1, 10 type 2, and 5 type 3	$\frac{1}{6}$ of type 1, $\frac{1}{8}$ of type 2, and $\frac{1}{8}$ of type 3 $\forall i,k$	Uniform([5,7],[3,5],[3,5],[3,5]) $\forall j,k$
11	Setting 3	20	Yes	3	15 type 1, 10 type 2, and 5 type 3	$\frac{1}{6}$ of type 1, $\frac{1}{8}$ of type 2, and $\frac{1}{8}$ of type 3 $\forall i,k$	Poisson(2,1,1,1) $\forall j,k$
12	Setting 3	20	Yes	3	15 type 1, 10 type 2, and 5 type 3	$\frac{1}{6}$ of type 1, $\frac{1}{8}$ of type 2, and $\frac{1}{8}$ of type 3 $\forall i,k$	Poisson(5,4,4,4) $\forall j,k$

5.2 Care Providers

This section describes how we have modelled the parameters associated with the recourse in the model. The recourse actions are those we have to make in order to mitigate the consequences of the mismatch between the scheduled resources and the demand for care. For example, calling someone in on a short notice can be considered as recourse action associated with having less than the required number of care providers. This section describes the corresponding costs. For example, the costs associated with an over-target wait list.

5.2.1 Costs of Shifts

In practice, each care-provider type is modelled using a specific cost per shift. These costs are incorporated into the model as part of the second-stage recourse. For example, consider a case when the number of care providers determined by the model (based on the forecasts) is less than the required level found based on the actual (simulated) arrivals for a specific shift. Here, the clinic will need to call in someone on a short notice and/or ask other care providers who have already covered their weekly/daily targets to stay for an extra shift. In such a situation, the cost of an extra shift is more than a normal one. This is what we call "understaffing". Conversely, the required level of care for a certain shift can be less than what has been anticipated by the model and, as a result, staff will be idle. In such situation, the clinic may be basically assigning resources to the wrong shift which can cause less quality of care on other shifts. The model penalizes this situation by a second-stage recourse as well. This situation is called "overstaffing".
Please note that when there is more than one type of care provider, the second-stage costs associated with over-staffed or under-staffed shifts can be different for each care-provider group.
Table 5.3 shows how we have madethese second-stage recourse values proportional to the cost of a regular shift for each care provider group. The following points about Table 5.3 are important to mention:

1. We have considered similar recourse values across all shifts. This implies that the cost of calling in someone on a short notice and the cost of idle time are the same for all shifts. Please note that in practice, this can be easily modified for different settings. This is a reasonable assumption based on what I have found in the "Ontario Nurses' Association Collective Agreement" (see NURSES'ASSOCIATION, 2013).

2. We have considered that the cost of an over-target shift (when understaffing happens) is 40% more than that of a regular shift. This was also obtained from

TABLE 5.3: List of recourse action costs associated with different care-providers and shifts

Care provider type	Morning Shift		Evening Shift		Night Shift	
	Overstaffed	Understaffed	Overstaffed	Understaffed	Overstaffed	Understaffed
Type 1	1.40	1.20	1.40	1.20	1.40	1.20
Type 2	1.68	1.44	1.68	1.44	1.68	1.44
Type 3	2.59	2.40	2.59	2.40	2.59	2.40

TABLE 5.4: List of model parameters associated with the different recourse action costs

Care provider type	Morning Shift		Evening Shift		Night Shift	
	Overstaffed	Understaffed	Overstaffed	Understaffed	Overstaffed	Understaffed
Type 1	C^-_{1j1}	C^+_{1j1}	C^-_{1j2}	C^+_{1j2}	C^-_{1j3}	C^+_{1j3}
Type 2	C^-_{2j1}	C^+_{2j1}	C^-_{2j2}	C^+_{2j2}	C^-_{2j3}	C^+_{2j3}
Type 3	C^-_{3j1}	C^+_{3j1}	C^-_{3j2}	C^+_{3j2}	C^-_{3j3}	C^+_{3j3}

the ONA collective agreement (NURSES'ASSOCIATION, 2013).

3. We consider the cost of a care-provider idle for an entire shift to be 20% more than that of a regular shift. This is a reasonable assumption for nurse scheduling problems in the literature (Bagheri, Devin, and Izanloo, 2016).

4. Finally, the recourse costs associated with the different care-provider types increase from type 1 to type 3. The specific values used are based on an hour of work for a Nurse, a Social Worker, and a General Practitioner obtained from the Ontario Public Sector Salary Calculator.

Table 5.4 shows how the costs mentioned in Table 5.3 are mapped with the cost parameters in the objective function of the model used by the SAA algorithm (see Equation 4.29). One can compare these two tables to get a better understanding of how the recourse values are chosen in the model.

5.2.2 Targets and Penalties Violating Them

As mentioned before, we have considered three targets in the model.
First, the daily number of shifts for each care provider, which corresponds to the maximum number of shifts (out of three) that an individual can cover during a 24-hour period. The target for this constraint is denoted in the model by $Target_{daily}$. We have set this value to 1 across all scenarios. This is a reasonable assumption based on what is discussed in Bagheri, Devin, and Izanloo, 2016 and NURSES'ASSOCIATION, 2013. We have denoted the costs of going over or under this constraint by C^-_{sj} and C^+_{sj} respectively. We have set these costs at 2 times the cost of a regular shift for each type of care provider to make sure that this constraint is not violated unless it is an

TABLE 5.5: List of cost parameters associated with the different targets

Target	Cost of Going	
	Over	Under
Weekly	C_s^-	C_s^+
Daily	C_{sj}^-	C_{sj}^+
Wait-list	C_{WL}^-	NoCost

TABLE 5.6: Costs of violating the different targets in the model

Target	Cost of Going	
	Over	Under
Weekly	1.4	1.4
Daily	2	2
Wait-list	1	NoCost

absolute emergency.
The second target in the model is the weekly target for the number of shifts for each individual. This target is denoted by $Target_{weekly}$ and is set at 5. This is also a reasonable assumption for a scheduling problem (NURSES'ASSOCIATION, 2013). The parameters C_s^- and C_s^+ denote the costs for going over or under this target, respectively. We have set each one of these at 1.4 times the cost of a regular shift.
The third target is for the length of the wait-list. We consider a weekly target of $Target_{waitlist}$. We have set this target equal to the capacity of the hospital for each scenario (i.e., 5, 10, and 20). The cost for exceeding the target value is considered equal to the cost of a regular shift for the smallest cost ratio (i.e., 1). This is a reasonable assumption in the healthcare facilities as well (Van den Bergh et al., 2013).

Table 5.5 summarizes the notation used to represent the costs associated with violating different targets.

Table 5.6 shows the values we have considered for the targets and the costs associated with their violation.

5.3 Patient Groups

We consider that these clinics face arrivals of four types. As mentioned before, grouping patients into categories can be done by Subject Matter Experts or on the

TABLE 5.7: List of shift and their start and end hours

Shift	Start	End
Morning	7:30	15:30
Evening	15:30	23:30
Night	23:30	7:30

basis of output of a clustering method. Here we assume that all of the four inpatient clinics are similar in terms of the number of patient types and their associated discharge probabilities. Obviously, this is just an assumption to be able to evaluate the model's performance. A more detailed characterization of the parameters would require actual data.
In accordance with subsection 4.1.1, we have considered four patient types with the following discharge probabilities.

$$\begin{aligned} p_{1jk} &= 0.2 \\ p_{2jk} &= 0.3 \\ p_{3jk} &= 0.5 \\ p_{4jk} &= 0.7 \end{aligned} \tag{5.1}$$

All scenarios (1 to 12) use these probabilities to determine discharges.

5.4 Time Horizon and Shifts

We assumed that the model is used to develop schedules for a 7-day period.

$$J = 7 \tag{5.2}$$

Also we consider three shifts for each day with the start and end hours mentioned in Table 5.7.

5.5 Parameters to Start the Sample Average Approximation

The last set of parameters we need to define are those required to start the SAA algorithm (see algorithm 4.5.1 for a detailed description of the SAA algorithm). As mentioned before, the initial values of these parameters are mostly found by a trial and error approach. This section provides a description of how we have chosen these

values: First of all, it should be noted that instead of starting the algorithm with a low value of N_0 (such as 1) and then running the algorithm for a large amount of time until a stopping criteria is met, we have decided to change N_0 gradually and report the values of $\overline{\nu}_N$ and $\overline{\nu}_{N_1}$ every time N changes. This way, we make sure that the algorithm does not cause the system to run out of memory without meeting a stopping criterion that maybe is too strict for that problem. We also check the values of $\overline{\nu}_N$, $\overline{\nu}_{N_1}$, and the run time for each N_0. In addition, this helps to record $\nu^m_{N_1}$ for all ms and check their standard deviation to make sure the robustness of the answer is sufficient. We have set the initial value of N_0 as 1 and then gradually increased it with the following values:

$$N_0 \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 500, 1000, 1500, 2000, 3000\} \tag{5.3}$$

For all scenarios, we have considered N_1 equal to 5000 and M equal to 10. While we did not use a stopping criteria in the form of the gap being less than ϵ, we have reported the number of sample scenarios required to get either a same gap for three consecutive values of N or the number of sample scenarios required to get the best gap.

Chapter 6

Results and Discussion

In this chapter, we present the results obtained using the model described in Chapter 4 on the data-set defined in Chapter 5. Here, we first report the results for all 12 scenarios and highlight some important characteristics of the solutions and then, we compare the results of two of the scenarios in order to show meaningful insights about the solution's behaviour.

6.1 Results

Table 6.1 shows the results for the 12 scenarios discussed in the first section of Chapter 5. This table displays the arrival distributions and the scenario number from the previous chapter along with some information on the solution obtained.
The first column is the setting which was discussed in Section 5.1. The second column describes arrivals in terms of their distributions and the corresponding parameters. The third column is the scenario number which comes from the previous section (for a complete description of the scenarios see Table 5.2). The fourth column is the required sample size. As described in Section 5.5, this column shows the number of sample scenarios which will either give three consecutive equal $\overline{v}_{N_1}$ values or is equal to the maximum number of scenarios N (i.e., 3000). The latter happens when the observed solution value keeps improving until N reaches the maximum number of scenarios.
The fifth column of Table 6.1 is the 95% confidence interval for $\overline{v}_N$ across all replications (ms). Since here we have consider less scenarios in our sample (N_0) compared to the main problem which considers all possible cases in an expected value form, this can be a lower bound for the main problem. After a solution is found using the N_0 scenarios, this solution/schedule is evaluated via N_1 additional scenarios in the second step of the SAA algorithm. Thus, $\overline{v}_{N_1}$ can be seen as an upper bound for the objective function of the main expected value problem. The sixth column in the table is the 95% confidence interval for $\overline{v}_{N_1}$ calculated across all ms. The last column in Table 6.1 shows the corresponding Average Optimality Gap (AOI) values which is a way of capturing the optimality gap between the lower and the upper bounds.

Note that these confidence intervals have been calculated under the assumption of a Normal distribution for $\overline{v}_N$ and $\overline{v}_{N_1}$. For a justification of why these two follow a Normal distribution, one can see Kleywegt, A. Shapiro, and Homem-de-Mello, 2002.
The AOI values are calculated as follows:
$AOI = \frac{\overline{v}_{N_1} - \overline{v}_N}{\overline{v}_{N_1}}$
All models have been run and evaluated on a PC with the following specifications:

$$\begin{aligned} Processor &: Intel(R)\ Core(TM)\ i5-6200U\ CPU\ 2.30GHz,\ 2400\ Mhz,\ 2\ Core(s), \\ &\quad 4\ Logical\ Processor(s) \\ SystemType &: 64-bit\ Operating\ System,\ x64-based\ Processor \\ Made &: BaseBoard\ Product\ K401UB \\ Memory &: DDR3L\ 1600\ MHz\ SDRAM\ ,\ OnBoard\ Memory\ 4\ GB \end{aligned} \tag{6.1}$$

Table 6.1 presents the results of the model:

TABLE 6.1: Results obtained using the SAA algorithm for the 12 scenarios from Chapter 5

Scenario (ξ)	Arrivals (A_{ijk})	Setting	Required Sample Szie (N)	95% CI Lower Bound	95% CI Upper Bound	AOI
1	Uniform([1,3],[1,2],[1,2],[1,2]) $\forall j,k$	Setting 1	1000	[49.03, 49.09]	[49.05, 49.10]	<0.001
2	Uniform([5,7],[3,5],[3,5],[3,5]) $\forall j,k$	Setting 1	1000	[33.73, 33.73]	[33.73, 33.74]	<0.001
3	Poisson(2,1,1,1) $\forall j,k$	Setting 1	1000	[53.85, 53.85]	[53.86, 53.86]	<0.001
4	Poisson(5,4,4,4) $\forall j,k$	Setting 1	1500	[34.24, 34.24]	[34.25, 34.26]	0.002
5	Uniform([1,3],[1,2],[1,2],[1,2]) $\forall j,k$	Setting 2	3000	[57.51, 57.68]	[57.55, 58.01]	0.04
6	Uniform([5,7],[3,5],[3,5],[3,5]) $\forall j,k$	Setting 2	2000	[28.00, 28.06]	[28.06, 28.11]	<0.001
7	Poisson(2,1,1,1) $\forall j,k$	Setting 2	2000	[24.02, 24.09]	[24.45, 24.79]	<0.001
8	Poisson(5,4,4,4) $\forall j,k$	Setting 2	3000	[26.47, 26.55]	[26.54, 26.62]	<0.001
9	Uniform([1,3],[1,2],[1,2],[1,2]) $\forall j,k$	Setting 3	3000	[102.45, 102.48]	[107.25, 107.35]	0.04
10	Uniform([5,7],[3,5],[3,5],[3,5]) $\forall j,k$	Setting 3	3000	[167.45,167.55]	[170.62, 170.69]	0.01
11	Poisson(2,1,1,1) $\forall j,k$	Setting 3	1500	[122.74,122.74]	[122.96, 122.99]	<0.001
12	Poisson(5,4,4,4) $\forall j,k$	Setting 3	3000	[157.41, 157.41]	[153.98, 154.00]	0.02

A few general points can be made considering the results:
1. First, as one can see in Table 6.1, as the scenarios get bigger (i.e., arrival rates increase, number of staff types increases, capacity of the clinic increases, the ratio of care-providers to patients increases, etc.), the width of the 95% interval of both $\overline{v}_N$ and $\overline{v}_{N_1}$ increases. This comes in part from the fact that as more variables are added to the model, the set of possible scenarios expands. This results in less accurate predictions for the number of patients of each type and generally wider lower bound and upper bound. These wider lower and upper bounds translate into higher optimality gaps and AOI values.
In practice, this means that as the problem increases in size, more samples sizes are required to obtain a lower optimality gap. This is of course associated with higher costs, both in modelling (by adding more variables and constraints) and solution time (by consuming more memory). For any particular case, it is important to find a balance between a desired optimality gap and the cost of solving the model. This signals the importance of choosing a wise optimality gap tolerance and stopping criterion. One way of finding a good optimality gap and a stopping criterion is to check the model's performance for another method, ideally, with a more rudimentary one or one that is simpler than the SAA approach. Comparison of the SAA with the manual methods have also been done before (Kleywegt, A. Shapiro, and Homem-de-Mello, 2002; Bagheri, Devin, and Izanloo, 2016).

2. We observed during this study that the larger the scenarios, the more likely the model is to show very small changes after increasing the sample size (even for large sample sizes). As an example, one can see that all scenarios in Setting 3 ended up not improving the AOI even after increasing the sample size from 1,000 to 3,000. This happened while considerable changes we observed in the values of $\overline{v}_{N_1}$ in sample sizes 1 to 1000. While this result cannot be generalized to other problems (Linderoth, A. Shapiro, and Wright, 2006), it suggests that for very large settings, it is better to set a reasonable tolerance for AOI instead of increasing the sample size indifferently until satisfying the stopping criteria. This can in particular prevent the model from consuming valuable resources without improving the results. Linderoth, A. Shapiro, and Wright, 2006, prove that in some cases, specially with large problems, the SAA algorithm will not converge to low ϵ values unless an extremely large sample size is used.
As an example of what was discussed, consider Figure 6.1. This figure depicts the behaviour of the evaluation step in Scenario 12 (the largest scenario). We can see in Figure 6.1 that increasing sample size has decreased the upper bound ($\overline{v}_{N_1}$, i.e., the evaluation objective function in the SAA) in the range of 1 to 1000. However, after a sample size of 1000 ($N_1 = 1000$), we did not observe any significant changes in the upper-bound.

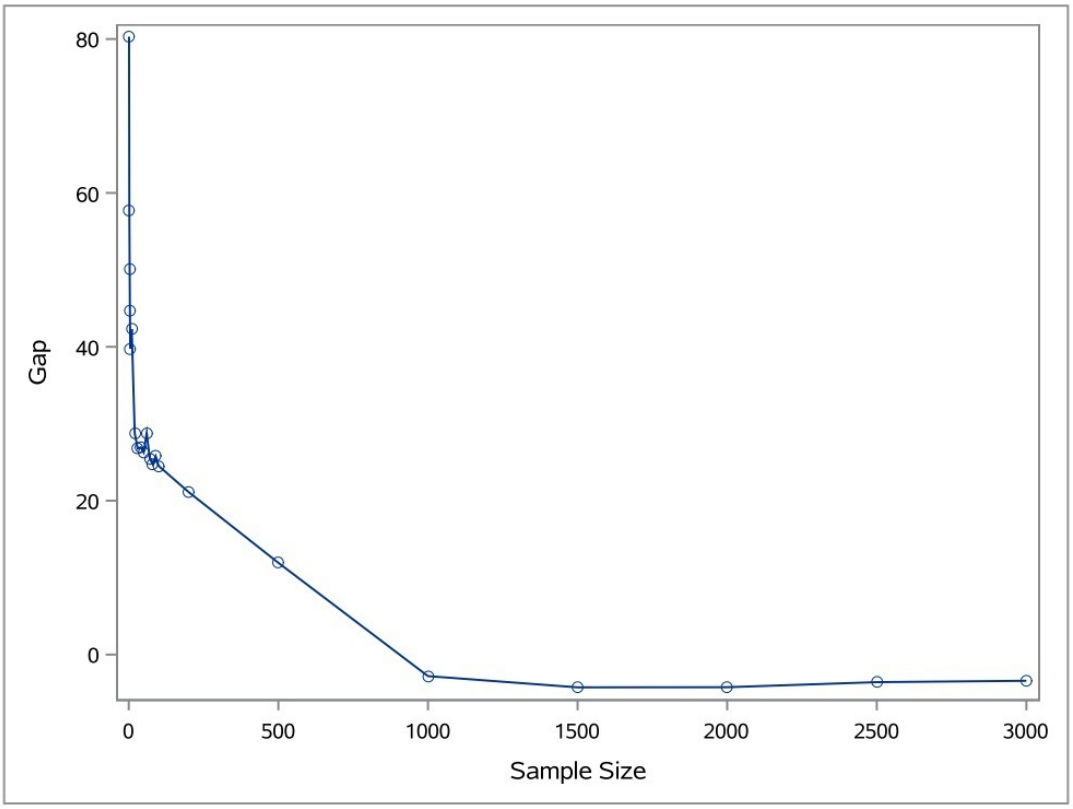

FIGURE 6.1: Changes in the values of $\overline{v}_{N_1}$ as the sample size increases for Scenario 12

Figure 6.2 shows the CPU time for this scenario. We can observe that while increasing the sample size does not have any significant impact on the value of $\overline{v}_{N_1}$, it significantly increases the run time.
Figure 6.2 shows the changes in the run time captured in seconds.

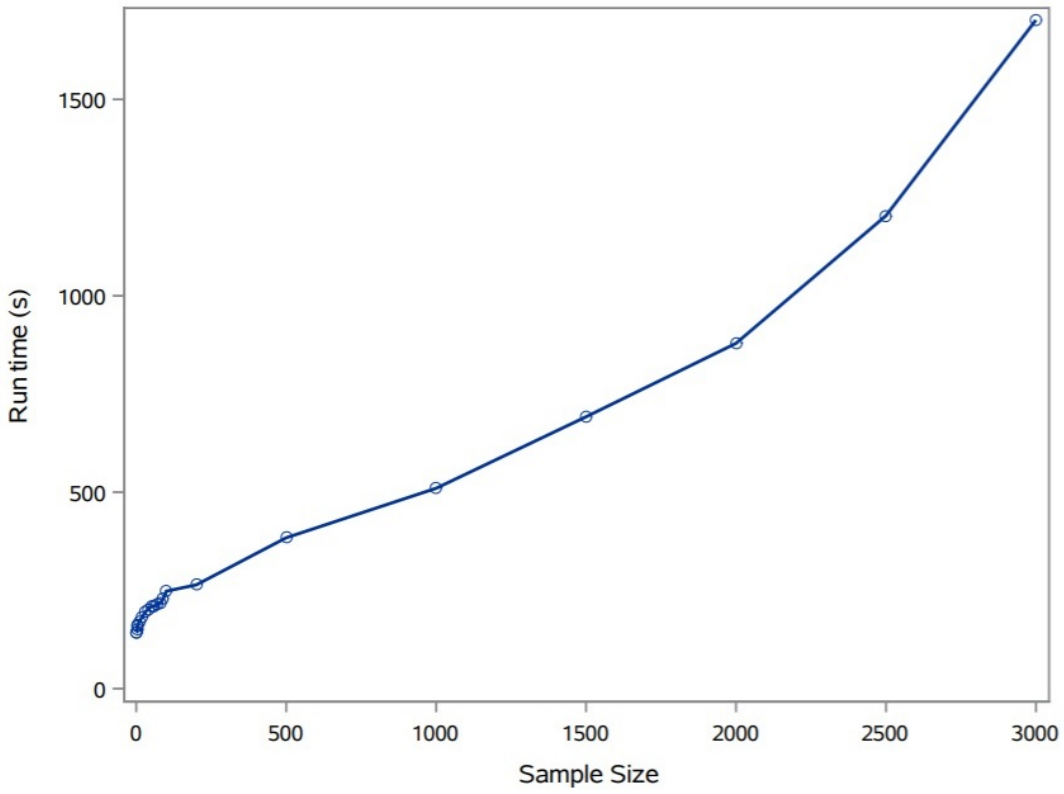

FIGURE 6.2: Changes in the CPU time as the sample increases size for Scenario 12

Choosing a moderate tolerance value for this problem is especially reasonable because the difference between a schedule found by SAA and that from the other methods, even with high tolerance values is significant (Bagheri, Devin, and Izanloo, 2016; Linderoth, A. Shapiro, and Wright, 2006; Ahmed, A. Shapiro, and E. Shapiro, 2002). The next chapter elaborates more on this.

3. The three groups of scenarios show different behaviours in their optimality gap, AOI, and the lower and upper bounds.
In the first four scenarios, we observe that for the smaller scenarios the algorithm reaches relatively lower optimality gaps within an equal range of sample sizes. The gaps are lower for smaller scenarios and higher for larger ones in each distribution. In the second setting, the two scenarios with Uniform distributions result in lower AOI values than the scenarios with Poisson distributions. This could be in part a result of lower variability in the arrivals. Finally, the last group of scenarios all result in a higher AOI values, which is the same for all arrivals.

4. We can observe that across all scenarios that the difference between two scenarios with Uniform arrivals is more than two scenarios with Poisson arrivals. This is true

for comparison between each of the following pairs.

$$\begin{aligned} &(1,2)\ v\ (3,4)\\ &(5,6)\ v\ (7,8)\\ &(9,10)\ v\ (11,12) \end{aligned} \tag{6.2}$$

6.2 Benchmarking the Model

In this section, we present a way of evaluating the results obtained via the proposed solution approach and checking their quality in terms of the total cost associated with the second stage variables. Our framework for this section is based on solving the scheduling problem as a simple deterministic problem via the method explained in algorithm 3 and then use the schedule from that problem in the evaluation step of the SAA algorithm along with SAA itself to check which answer provides better results. Algorithm 3 describes this process.

Algorithm 3: Expected Value Evaluation Problem for evaluating the SAA

input : Arrival distributions and corresponding parameters (A_{ijk} as the Poisson rate, and Min_{ijk} and Max_{ijk} for the the Uniform maximum and minimum arrivals)

output: Schedule S_{EVP}

1 Step 0. Ignore distribution type and store arrival rates (λ_{ijk} for a Poisson arrival and Max_{ijk} and Min_{ijk} for a Unifrom arrival).
2 Step 1. Calculate the average arrival rates for each shift as follows.
3 for a Poisson distribution:
4 $X_{ijk} = \sum_i A_{ijk}$
5 for a Uniform distribution.
6 $X_{ijk} = \frac{(Max_{ijk}+Min_{ijk})}{2}$
7 Step 2. Solve a deterministic version of problem 4.29 to 4.33 letting all number of patients equal to X_{ijk} found in Step 1 and Record the optimal solution S_{EVP} for all j, k .
8 Step 3. Solve the SAA algorithm objective function (equation 4.34) for N_1 using the schedule S_{EVP}. Report the answer Objective Function Ojb_{EVP}

While the process of finding Obj_{EVP} might not be a well-known algorithm we have considered it as an algorithm in its essential sense here. Because first of all the definition of an algorithm applies to it, [1] and second, this method has been previously used in the literature as a benchmarking strategy to evaluate the performance of SAA algorithm (Bagheri, Devin, and Izanloo, 2016; Ernst et al., 2004a).
Next, we compare the 95th percentile of the $\overline{v}_{N_1}$'s distribution with the EVP objective

[1] An algorithm is a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation

TABLE 6.2: A comparison between the results found wiht the SAA approach and the average case (EVP approach)

Scenario	Objective Value	
	SAA	EVP
1	49.08	301.40
2	33.73	280.63
3	53.75	165.00
4	34.24	323.14
5	57.56	129.35
6	28.07	333.64
7	24.62	316.53
8	26.62	317.45
9	104.89	225.19
10	169.08	253.04
11	122.99	253.04
12	155.70	221.28

value. Table 6.2 presents these values and the corresponding EVP objectives.

As we can see from Table 6.2, the EVP schedule has resulted in worse objective values for all scenarios. This shows the better performance of the SAA algorithm. In fact, considering the stochastic arrivals and discharges, and the use of simulation, significantly affects the cost of the second-stage recourse for all scenarios. There are two reasons for this improvement in the cost of the problem. First of all, the EVP algorithm assumes the same resource utilization for all similar shifts in the schedule which is not a realistic assumption. In our model, the evolution of the Markov predictive model provides more realistic scenarios for the future census of the clinic. Second, the EVP algorithm does not consider the distribution of arrivals. However, our approach takes into account the probability distributions of the arrival and discharges. This helps the model make better predictions of the future census of the clinic.

In many cases, even for a very small sample size (less than 10), the proposed approach can produce a better schedule than the one found with the EVP approach. Note that, in general, the average approach is better than most of the traditional

methods people use in practice.

In addition to the overall better performance, we can also observe that for the larger scenarios the difference between the two approaches is more than that for the smaller scenarios. This, together with the results presented in the previous section, can explain why using the SAA algorithm for large settings might not result in smaller gaps. Nevertheless, it can still be used as a robust and efficient method for staff scheduling. Figure 6.4 shows the objective values of the SAA and EVP approaches for the 12 scenarios.

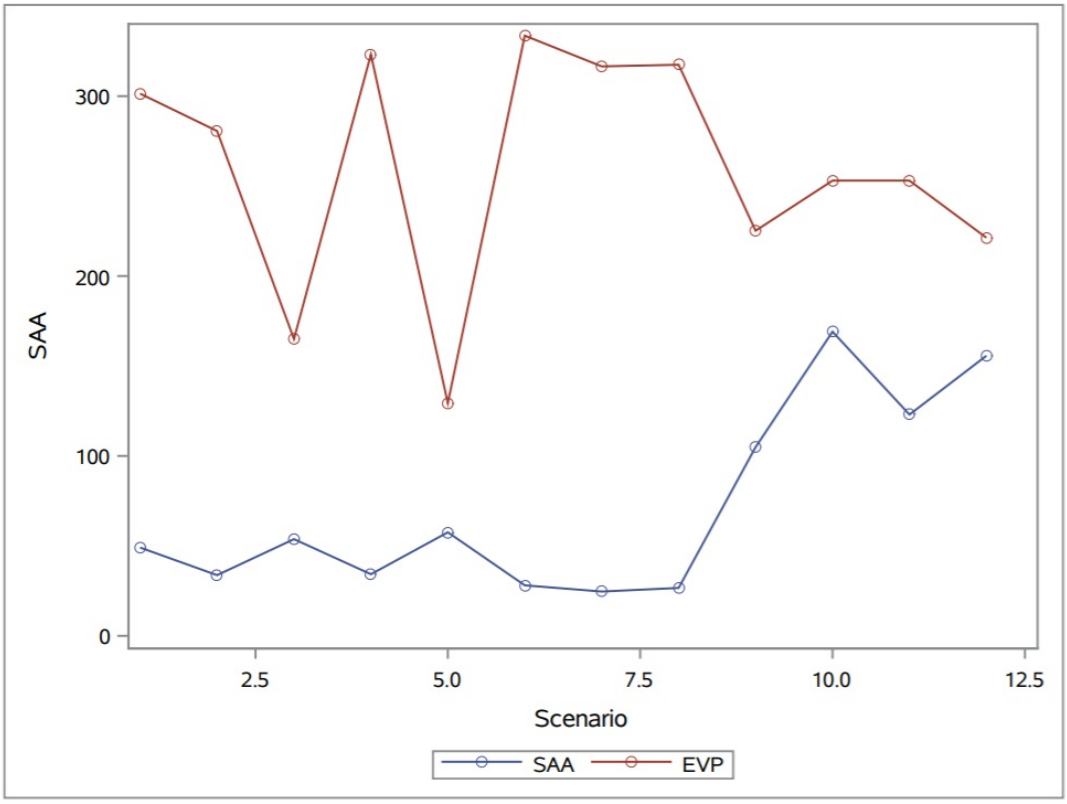

FIGURE 6.3: Changes in the performance of the SAA and EVP approaches for the different scenarios

One can also understands these differences by looking at the following bar chart:

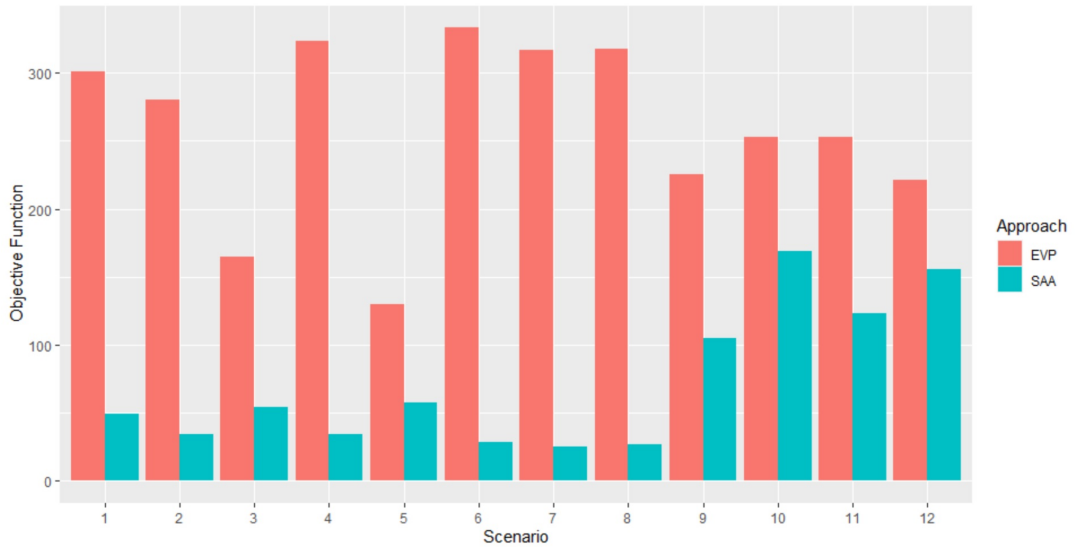

FIGURE 6.4: Comparison of the objective function value of the EVP and SAA approaches

6.3 Comparison

In this section, we provide a detailed description of the behaviour of the SAA algorithm across all sample sizes for the smallest scenario, scenario 1, and the largest scenario, scenario 12. We present the results in three parts.
First, the values of $\overline{v}_{N_1}$, $\overline{v}_N$, and Obj_{EVP} are depicted across different sample sizes. One can observe how these three values change as the sample size increases.
Second, the optimality gap ($\overline{v}_{N_1}$ - $\overline{v}_N$) is shown across all sample sizes.
Third, we provide the CPU times for these two scenarios.

6.3.1 $\overline{v}_{N_1}$, $\overline{v}_N$, and Obj_{EVP}

Figure 6.5 shows the objective value of the average problem ($\overline{v}_N$ in Equation 4.35), the values coming from the evaluation step($\overline{v}_{N_1}$ in Equation 4.36), and the expected value problem (Obj_{EVP} found in algorithm 3) for the first scenario.
In this figure, we can observe how as sample size increases $\overline{v}_{N_1}$ decreases and $\overline{v}_N$ increases. We can also compare both of these values with respect to the EVP problem.

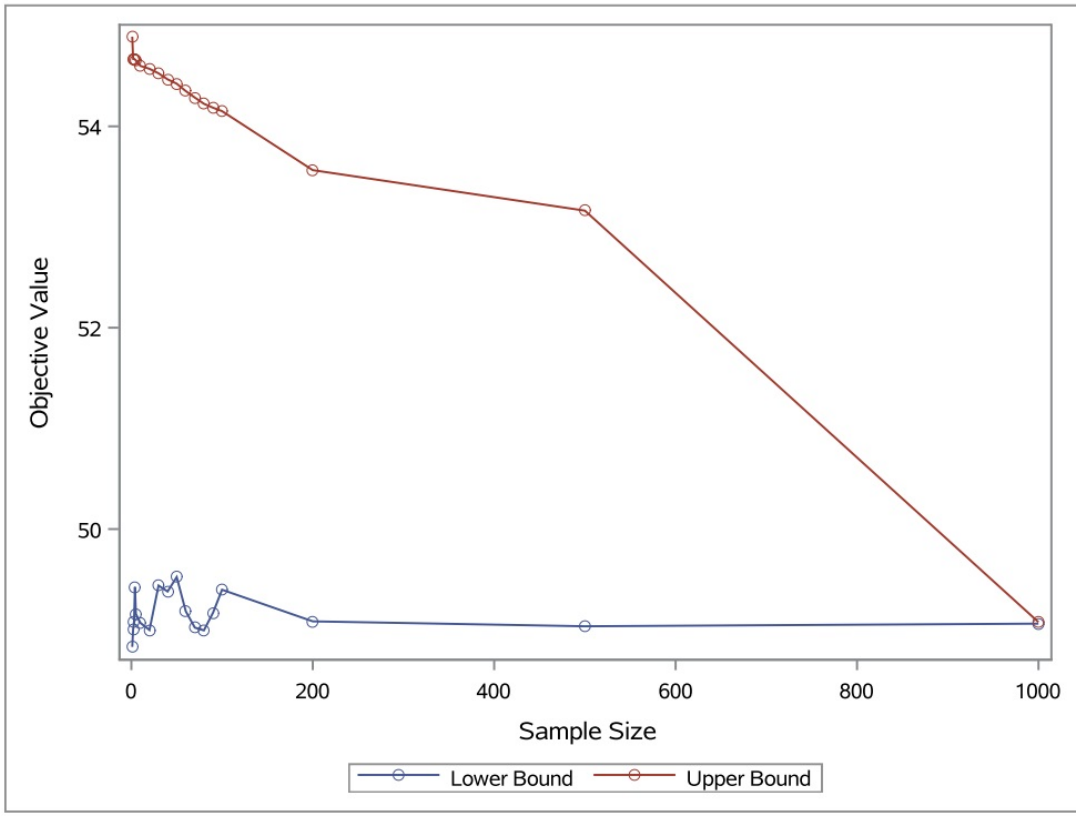

FIGURE 6.5: The values of $\overline{v}_{N_1}$ and $\overline{v}_N$ for different sample sizes for Scenario 1

Figure 6.6 has the same information but for scenario 12. Here, we can observe that the changes in the values of $\overline{v}_N$ are smaller and values of $\overline{v}_{N_1}$ do not show any considerable change for sample sizes more than 1000.

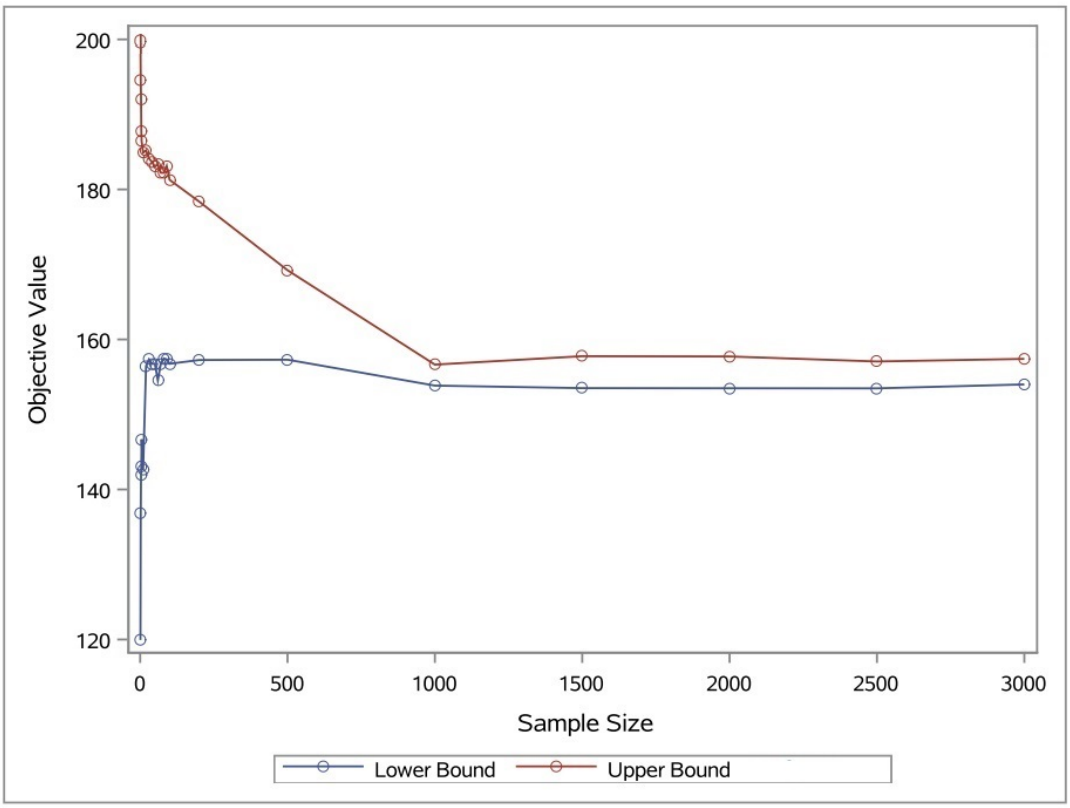

FIGURE 6.6: The values of $\overline{v}_{N_1}$ and $\overline{v}_N$ for different sample sizes for Scenario 12

6.3.2 Optimality Gap

Figures 6.7 and 6.8 present the different optimality gaps for scenarios 1 and 12. We can observe how the gap decreases for Scenario 1 and remains constant for Scenario 12.

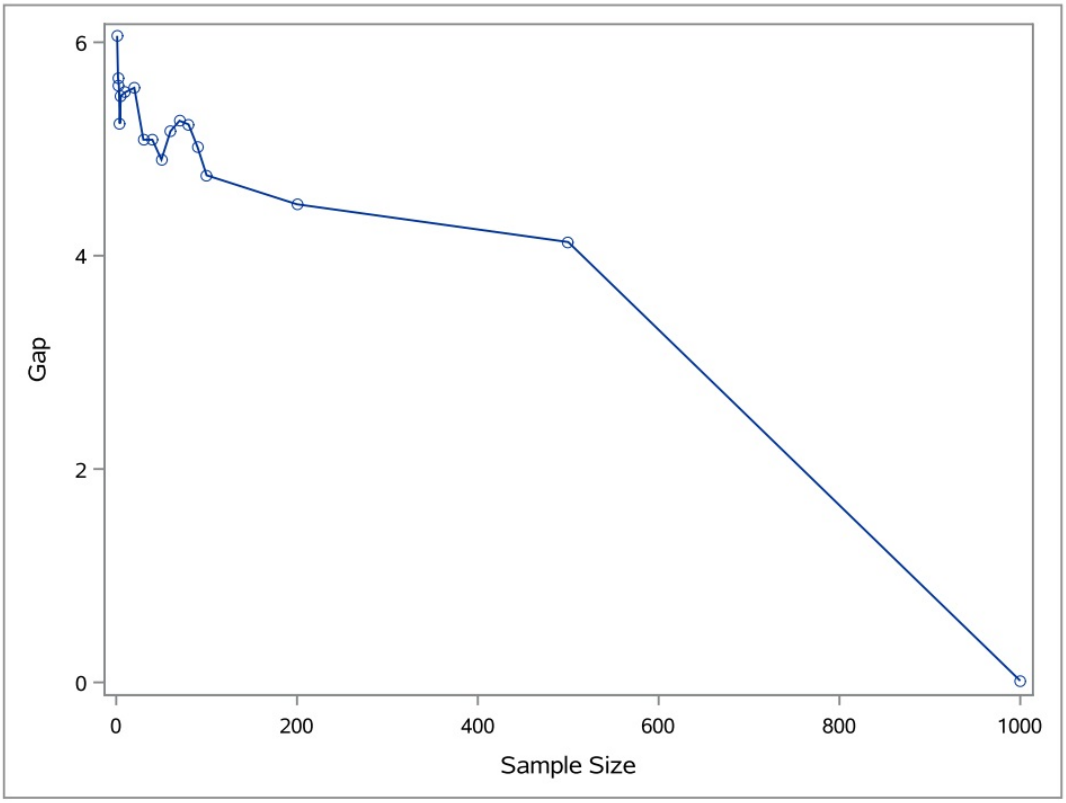

FIGURE 6.7: Optimality gap for different sample sizes for Scenario 1

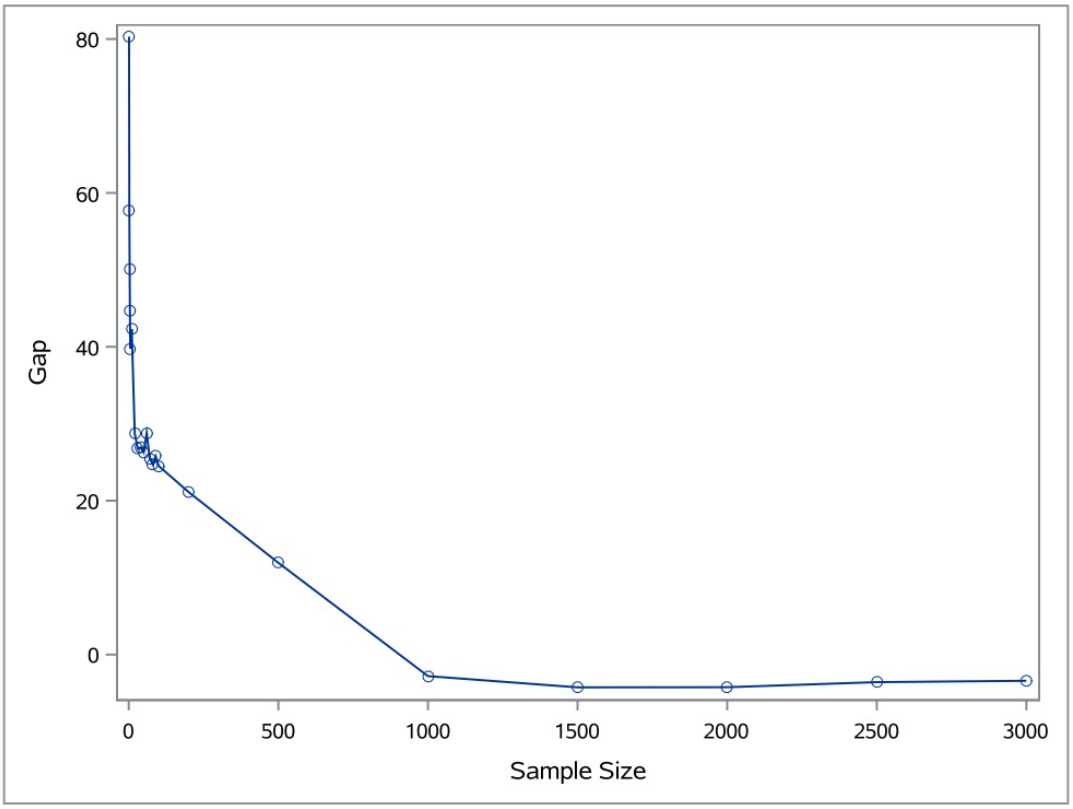

FIGURE 6.8: Optimality gap for different sample sizes for Scenario 12

6.3.3 Run Time

Figures 6.9 and 6.10 show how the run time increases with the sample size. We can see that the rate of increase in the largest scenario is considerably higher than the smallest one.

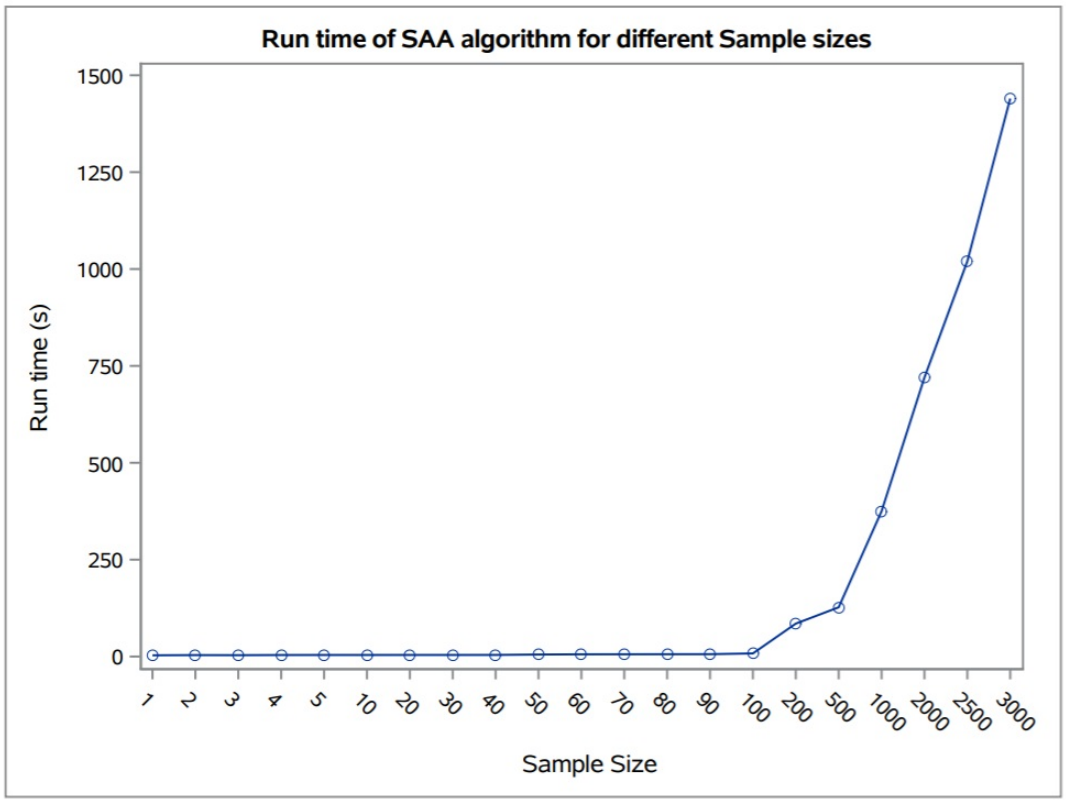

FIGURE 6.9: Run time for different sample sizes for Scenario 1

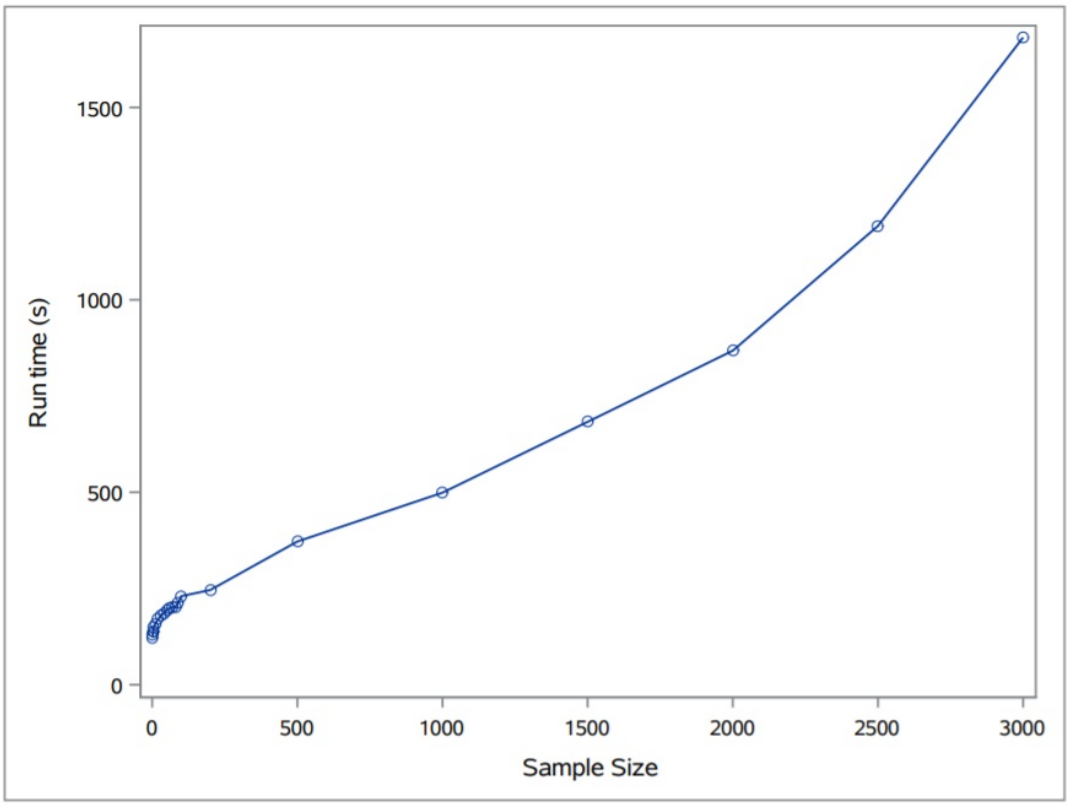

FIGURE 6.10: Run time for different sample sizes in Scenario 12

6.4 The Final Schedule

After solving the problem, the next goal would be translating its results into a meaningful schedule which can be used as a staff schedule for the clinic. Here, as a sample of the schedules created, the schedule obtained by using the solution approach we proposed is presented. This schedule corresponds to Scenario 8. The next two figures are the schedule from the clinic's and the individuals' perspectives. In this schedule, each care-provider has 5 shifts. Please note that in scenario 8 there is only one care-provider type and 15 individuals of this type are working in the clinic.

	Monday			Tuesday			Wednesday			Thursday			Friday			Saturday			Sunday		
Care-provider	Morning	Evening	Night	Morning	Evening	Night	Morning	Evening	Night	Morning	Evening	Night	Morning	Evening	Night	Morning	Evening	Night	Morning	Evening	Night
1	1											1						1		1	1
2									1	1			1		1	1					
3				1	1		1				1					1					
4					1									1		1			1	1	
5		1	1		1		1	1													
6	1									1	1			1					1		
7		1						1			1		1				1				
8	1			1					1										1	1	
9		1	1	1										1			1				
10	1							1		1			1								1
11							1					1	1		1	1					
12	1	1			1												1		1		
13				1		1	1				1							1			
14		1						1		1				1				1			
15						1										1	1		1	1	
	5	5	2	4	4	2	4	4	2	4	4	2	4	4	2	5	4	3	5	4	2

FIGURE 6.11: The proposed schedule for Scenario 8- individuals' view

Scehdule	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Morning	5	4	4	4	4	5	5
Evening	5	4	4	4	4	5	4
Night	2	2	2	2	2	2	2

FIGURE 6.12: The proposed schedule for Scenario 8- clinic's view

Chapter 7

Limitations and Additional Considerations

During the course of this applied research initiative, the main purpose has been to model and solve the problem, and of course learn from the process. However, like any other human endeavour, our work has limitations. In this chapter, we describe the limitations we faced during different phases of this study. In the end, a section is also dedicated to possible future work.

1. As explained in Chapter 6, we have used an artificial test dataset to evaluate the performance of the model we developed. While we tried our best to make the scenarios as realistic and comprehensive as possible, there are always some aspects that only show themselves in reality. Lack of real data is the first limitation of our study [1]. A study that tests and validates our models and solutions using real data would be a perfect complement to this work.

2. As mentioned in 6.1, all simulation and optimization models were run on a rudimentary PC. As a result the run times are higher than those that could be required in practice. Also, the sample sizes were not increased beyond $N_0 = 3000$. Running the models on a faster processors would be ideal for any similar study.

3. As mentioned in Chapter 4, we use a Binomial distribution for the number of people discharged from the clinic on a daily basis. Besides being mathematically correct, another benefit of using this distribution was the fact that it is simple to update. Although we have considered this in the model, we did not run such an experiment and evaluate the changes in the model's performance when considering probability updates.

[1] While my supervisors and I made several efforts to consider real data for my study, I was not able to find any real-world dataset. One of the reason for this was that I finished my thesis in the midst of COVID-19 outbreak in Winter and Spring of 2020.

7.1 Future Work

In this section, we propose some possible future work that can be done on the basis of our study.

1. Considering treatment modelling

While we consider a relatively detailed model of arrivals and discharges when compared to the models in Broyles, Cochran, and Montgomery, 2010, Bagheri, Devin, and Izanloo, 2016, and Vassilacopoulos, 1985), our methodology dose not consider what happens in the clinic in each shift. A complementary study could consider different routines in the clinic (for example fixed treatments on specific days or shifts) and change forecasts and care-provider requirements accordingly. At this point, I believe incorporating these into the model would increase accuracy of the scheduling decisions.

2. Probability updating experiment

As explained above, an aspect of the model that we were not able to test was the updating of the discharge probabilities based on predictions made by experts in the clinic. Here a brief review of that method is presented for possible future use.

Assume the discharge probability (p_is) patient type i. In order to update these probabilities based on experts' opinion, one can use use Bayesian Inference. This can possibly enhance the model's accuracy and practical behaviour over time. The idea can be implemented in the form of asking physicians, or any other care provider in the inpatient clinic, to evaluate patients' conditions and estimate their length of stay for patients of each type. Next, we observe what actually happens in the clinic on the shifts following the prediction and use it as measure of accuracy of the prediction.

Let the probability event of this observation be E. Then, we can use this observation to update the (p_is) on a regular basic. The process would go on for any new observation and, in the long run, would enhance the model's accuracy. Eventually, we use the most recent p_is to simulate the number of patients mix that is going to be fed into the SAA algorithm. At first, we can assume there is 50 percent chance that this prediction of discharge probability is correct and 50 percent chance that it is not correct (i.e. the complement set of this prediction is correct). After observing the situation in the clinic, we calculate the probability of the observation knowing the prediction and use it as the updated chance of prediction being correct. Finally, we find the updated probability of a patient remaining in the clinic by the Law of Total Probability. Mathematical description of what mentioned so far is presented in the next three equations.

First a review of Bayes Rule. Assuming two hypothetical events A and B, the general form of Bayes Rule would be:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \tag{7.1}$$

From updating probability assessment we have:

$$Posterior/prior \times likelihood \tag{7.2}$$

By applying the Bayes rule to our problem setting we have:

$$P(p_i|E) = \frac{P(E|p_i)P(p_i)}{P(E|p_i)P(p_i) + P(E|p_i^c)P(p_i^c)} \tag{7.3}$$

Finally,

$$\text{Updated } p_i = P(p_i|E) \times p_i + (1 - P(p_i|E)) \times p_i^c \tag{7.4}$$

3. Considering a larger system (Integral set of staffing models)
In a complex healthcare system, patients move between different settings regularly. Often a considerable amount of arrivals can be predicted from these transition probabilities. One can build a network of different settings and connect them as a an integrated body. Considering these probabilities for arrivals would also add to the accuracy of the model.

4. Decision Support System Finally, a future study/work project can consider the development of a proper interface for the core model we have developed. The ultimate goal of such a study could be turning the model into a decision support tool that can receive data in real time and help decision-makers make more accurate and efficient scheduling decisions.

Appendix A

A Discussion on Possible Additional Constraints to Be Incorporated into the MIP Model

As mentioned in Chapters 4, 5, and 6 the Mixed-Integer Programming model we have proposed is in a general form. This means that we have only considered constraints that are necessary to every staff scheduling model. We have considered demand coverage, wait-list capacity, and maximum daily and weekly number of shifts for care-providers. Obviously, there could be many other constraints and terms in the objective function based on the specific nature of each inpatient clinic. Union rules, workplace regulations, and practical limitations could be the potential reasons for this. In this appendix, we address some of these possible constraints to acknowledge the capability of the model to incorporate these changes. While we have not considered any of these in the simulation and tests, they can easily be adopted and incorporated into the model.

A.1 Manual Assignments

In some hospitals, a group of care-providers must be assigned to specific shifts. For example, head nurses must be assigned to the morning shifts (El Adoly, Gheith, and Fors, 2018; Bagheri, Devin, and Izanloo, 2016). The way of assigning an individual to a specific shift is as simple as forcing the S variable to 1 for that individual. Assuming H shows the set of care-providers who must go to work at every morning shifts we have:

$$S_{sjk} = 1 \quad \forall j, k = 1, s \ \in H \tag{A.1}$$

A.2 Consecutive Evening or Night Shifts

Sometimes it is important for the decision-makers that no two consecutive evening and night shifts are assigned to one person. The way of incorporating that into the model is as follows:

$$S_{sj2} + S_{sj3} \leq 1 \quad \forall s, j \tag{A.2}$$

A.3 No Day Shift Following a Night Shift

In many countries, union rules necessitate that care-providers cannot work the following morning shift they have fulfilled a night shift (NURSES'ASSOCIATION, 2013; Lim et al., 2016). We can simply incorporate this constraint in our model as well:

$$S_{sj3} + S_{s(j+1)1} \leq 1 \quad \forall s, j \tag{A.3}$$

A.4 Minimum Number of Night Shifts

In some cases, a model considers a minimum number of night shifts for each care-provider (Bagheri, Devin, and Izanloo, 2016). This is specifically important to ensure the fairness of the schedule. Obviously, if it is needed for feasibility, the model can penalize values less or more than the required number of night shifts. Considering A as the number of night shifts everyone must fulfil and slack variables m_s^- and m_s^- we have:

$$\sum_j S_{sj3} + m_s^- - m_s^- = A \quad \forall s \tag{A.4}$$

Appendix B

Sample Average Approximation Algorithm- Flow Chart

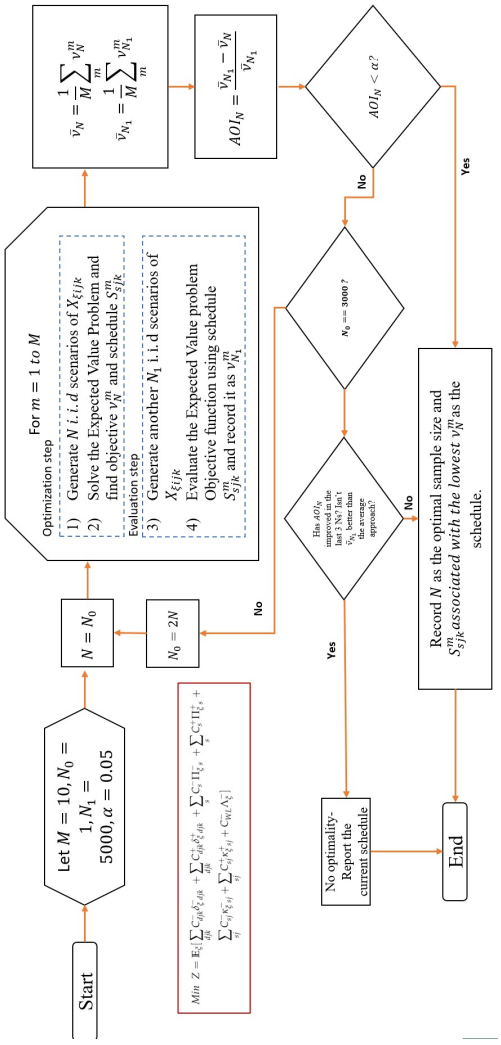

FIGURE B.1: Flow chart of the Sample Average Approximation algorithm

Bibliography

Agnihothri, Saligrama R and Patricia F Taylor (1991). "Staffing a centralized appointment scheduling department in Lourdes Hospital". In: *Interfaces* 21.5, pp. 1–11.

Ahmad, Parvez, Saqib Qamar, and Syed Qasim Afser Rizvi (2015). "Techniques of data mining in healthcare: a review". In: *International Journal of Computer Applications* 120.15.

Ahmed, Shabbir, Alexander Shapiro, and Er Shapiro (2002). "The sample average approximation method for stochastic programs with integer recourse". In: *Submitted for publication*, pp. 1–24.

Amaravadi, Ravi K et al. (2000). "ICU nurse-to-patient ratio is associated with complications and resource use after esophagectomy". In: *Intensive care medicine* 26.12, pp. 1857–1862.

Auditor-General, Ontario (2016). "Annual Report of the Auditor General of Ontario for the year 2016". In:

Bagheri, Mohsen, Ali Gholinejad Devin, and Azra Izanloo (2016). "An application of stochastic programming method for nurse scheduling problem in real word hospital". In: *Computers & Industrial Engineering* 96, pp. 192–200.

Balaratnasingam, Sivasankaran et al. (2011). "Mental health risk assessment: a guide for GPs". In: *Australian family physician* 40.6, p. 366.

Bianchi, Leonora et al. (2009). "A survey on metaheuristics for stochastic combinatorial optimization". In: *Natural Computing* 8.2, pp. 239–287.

Bidhandi, Hadi Mohammadi et al. (2019). "Capacity planning for a network of community health services". In: *European Journal of Operational Research* 275.1, pp. 266–279.

Broyles, James R and Jeffery K Cochran (2009). "A Markov chain methodology for predicting hospital inpatient census". In: *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), p. 832.

Broyles, James R, Jeffery K Cochran, and Douglas C Montgomery (2010). "A statistical Markov chain approximation of transient hospital inpatient inventory". In: *European Journal of Operational Research* 207.3, pp. 1645–1657.

Brunner, Jens O, Jonathan F Bard, and Rainer Kolisch (2009). "Flexible shift scheduling of physicians". In: *Health care management science* 12.3, pp. 285–305.

– (2010). "Midterm scheduling of physicians with flexible shifts using branch and price". In: *Iie Transactions* 43.2, pp. 84–109.

Callahan, Alison and Nigam H Shah (2017a). "Machine learning in healthcare". In: *Key Advances in Clinical Informatics*. Elsevier, pp. 279–291.

– (2017b). "Machine learning in healthcare". In: *Key Advances in Clinical Informatics*. Elsevier, pp. 279–291.

Capan, Muge et al. (2017). "From data to improved decisions: Operations Research in healthcare delivery". In: *Medical Decision Making* 37.8, pp. 849–859.

Chaari, Tarek et al. (2014). "Scheduling under uncertainty: Survey and research directions". In: *2014 International conference on advanced logistics and transport (ICALT)*. IEEE, pp. 229–234.

Chanchaichujit, Janya et al. (2019). "Optimization, Simulation and Predictive Analytics in Healthcare". In: *Healthcare 4.0*. Springer, pp. 95–121.

Cheang, Brenda et al. (2003). "Nurse rostering problems—-a bibliographic survey". In: *European journal of operational research* 151.3, pp. 447–460.

Cochran, Jeffery K and K Roche (2008). "A queuing-based decision support methodology to estimate hospital inpatient bed demand". In: *Journal of the Operational Research Society* 59.11, pp. 1471–1482.

Cooper, James K and Timothy M Corcoran (1974). *Estimating bed needs by means of queuing theory*.

El-Darzi, Elia, Revlin Abbi, et al. (2009). "Length of stay-based clustering methods for patient grouping". In: *Intelligent patient management*. Springer, pp. 39–56.

El-Darzi, Elia, Christos Vasilakis, et al. (1998). "A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department". In: *Health care management science* 1.2, p. 143.

De Bruin, Arnoud M et al. (2007). "Modeling the emergency cardiac in-patient flow: an application of queuing theory". In: *Health Care Management Science* 10.2, pp. 125–137.

Dua, Sumeet, U Rajendra Acharya, and Prerna Dua (2014). *Machine learning in healthcare informatics*. Vol. 56. Springer.

Eaton, William W and GA Whitmore (1977). "Length of stay as a stochastic process: A general approach and application to hospitalization for schizophrenia". In: *Journal of Mathematical Sociology* 5.2, pp. 273–292.

El Adoly, Ahmed Ali, Mohamed Gheith, and M Nashat Fors (2018). "A new formulation and solution for the nurse scheduling problem: A case study in Egypt". In: *Alexandria engineering journal* 57.4, pp. 2289–2298.

Ernst, Andreas T et al. (2004a). "Staff scheduling and rostering: A review of applications, methods and models". In: *European journal of operational research* 153.1, pp. 3–27.

– (2004b). "Staff scheduling and rostering: A review of applications, methods and models". In: *European journal of operational research* 153.1, pp. 3–27.

Eveborn, Patrik, Patrik Flisberg, and Mikael Rönnqvist (2006). "Laps Care—an operational system for staff planning of home care". In: *European journal of operational research* 171.3, pp. 962–976.

Farmer, RD and J Emami (1990). "Models for forecasting hospital bed requirements in the acute sector." In: *Journal of Epidemiology & Community Health* 44.4, pp. 307–312.

Fomundam, Samuel and Jeffrey W Herrmann (2007). *A survey of queuing theory applications in healthcare*. Tech. rep.

Fouskakis, Dimitris and David Draper (2002). "Stochastic optimization: a review". In: *International Statistical Review* 70.3, pp. 315–349.

Gallivan, Steve et al. (2002). "Booked inpatient admissions and hospital capacity: mathematical modelling study". In: *Bmj* 324.7332, pp. 280–282.

Green, Linda V et al. (2006). "Using queueing theory to increase the effectiveness of emergency department provider staffing". In: *Academic Emergency Medicine* 13.1, pp. 61–68.

Hachesu, Peyman Rezaei et al. (2013). "Use of data mining techniques to determine and predict length of stay of cardiac patients". In: *Healthcare informatics research* 19.2, pp. 121–129.

Homem-de-Mello, Tito and Güzin Bayraksan (2014). "Monte Carlo sampling-based methods for stochastic optimization". In: *Surveys in Operations Research and Management Science* 19.1, pp. 56–85.

Hyndman, Rob J and George Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts.

Ibrahim, Arief (2019). "Forecasting patient demand and predicting inpatient admission via machine learning techniques in acute care domain". MA thesis. University of Twente.

Isken, Mark W and Balaji Rajagopalan (2002). "Data mining to support simulation modeling of patient flow in hospitals". In: *Journal of medical systems* 26.2, pp. 179–197.

J Murray, Michael (2003). "The Canadian Triage and Acuity Scale: A Canadian perspective on emergency department triage". In: *Emergency medicine* 15.1, pp. 6–10.

Jain, RK (1989). "A semi-Markov model for the average length of stay in transient states and its application". In: *Computers and Biomedical Research* 22.3, pp. 209–214.

Jothi, Neesha, Wahidah Husain, et al. (2015). "Data mining in healthcare–a review". In: *Procedia Computer Science* 72, pp. 306–313.

Kao, Edward PC and Frank M Pokladnik (1978). "Incorporating exogenous factors in adaptive forecasting of hospital census". In: *Management Science* 24.16, pp. 1677–1699.

Kapadia, Asha Seth et al. (2000). "Predicting duration of stay in a pediatric intensive care unit: A Markovian approach". In: *European Journal of Operational Research* 124.2, pp. 353–359.

Katsaliaki, Korina and Navonil Mustafee (2011). "Applications of simulation within the healthcare context". In: *Journal of the Operational Research Society* 62.8, pp. 1431–1451.

Keller, Brian and Güzin Bayraksan (2009). "Scheduling jobs sharing multiple resources under uncertainty: A stochastic programming approach". In: *Iie Transactions* 42.1, pp. 16–30.

Kim, Sujin, Raghu Pasupathy, and Shane G Henderson (2015). "A guide to sample average approximation". In: *Handbook of simulation optimization*. Springer, pp. 207–243.

Kleywegt, Anton J, Alexander Shapiro, and Tito Homem-de-Mello (2002). "The sample average approximation method for stochastic discrete optimization". In: *SIAM Journal on Optimization* 12.2, pp. 479–502.

Koh, Hian Chye, Gerald Tan, et al. (2011). "Data mining applications in healthcare". In: *Journal of healthcare information management* 19.2, p. 65.

Lakshmi, C and Sivakumar Appa Iyer (2013). "Application of queueing theory in health care: A literature review". In: *Operations research for health care* 2.1-2, pp. 25–39.

Lapierre, Sophie D et al. (1999). "Bed allocation techniques based on census data". In: *Socio-Economic Planning Sciences* 33.1, pp. 25–38.

Leeftink, AG, IMH Vliegen, and Erwin W Hans (2019). "Stochastic integer programming for multi-disciplinary outpatient clinic planning". In: *Health care management science* 22.1, pp. 53–67.

Legrain, Antoine, Jérémy Omer, and Samuel Rosat (2018). "An online stochastic algorithm for a dynamic nurse scheduling problem". In: *European Journal of Operational Research*.

Liew, Don, Danny Liew, and Marcus P Kennedy (2003). "Emergency department length of stay independently predicts excess inpatient length of stay". In: *Medical Journal of Australia* 179.10, pp. 524–526.

Likas, Aristidis, Nikos Vlassis, and Jakob J Verbeek (2003). "The global k-means clustering algorithm". In: *Pattern recognition* 36.2, pp. 451–461.

Lim, Gino J et al. (2016). "Nurse scheduling with lunch break assignments in operating suites". In: *Operations Research for Health Care* 10, pp. 35–48.

Linderoth, Jeff, Alexander Shapiro, and Stephen Wright (2006). "The empirical behavior of sampling methods for stochastic programming". In: *Annals of Operations Research* 142.1, pp. 215–241.

Littig, Steven J and Mark W Isken (2007). "Short term hospital occupancy prediction". In: *Health care management science* 10.1, pp. 47–66.

Liu, Peng et al. (2006). "Healthcare data mining: Prediction inpatient length of stay". In: *2006 3rd International IEEE Conference Intelligent Systems*. IEEE, pp. 832–837.

Lodhia, Z, Akhtar Rasool, and Gaurav Hajela (2017). "A survey on machine learning and outlier detection techniques". In: *IJCSNS* 17.5, p. 271.

Mackay, Mark (2001). "Practical experience with bed occupancy management and planning systems: an Australian view". In: *Health Care Management Science* 4.1, pp. 47–56.

Mackay, Mark and Michael Lee (2005). "Choice of models for the analysis and forecasting of hospital beds". In: *Health Care Management Science* 8.3, pp. 221–230.

Marino, Paul L and Kenneth M Sutin (1998). *The ICU book*. Vol. 2. Williams & Wilkins Baltimore:

Marshall, Adele H, Sally I McClean, and Peter H Millard (2004). "Addressing bed costs for the elderly: a new methodology for modelling patient outcomes and length of stay". In: *Health care management science* 7.1, pp. 27–33.

Medicine, Institute of (2001). "Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century". In: *National Academies Press*.

Mehandiratta, Reetu (2011). "Applications of queuing theory in health care". In: *International Journal of Computing and Business Research* 2.2, pp. 2229–6166.

Mustafee, Navonil, Korina Katsaliaki, and Simon JE Taylor (2010). "Profiling literature in healthcare simulation". In: *Simulation* 86.8-9, pp. 543–558.

Nguyen, JM et al. (2007). "An objective method for bed capacity planning in a hospital department". In: *Methods of information in medicine* 46.04, pp. 399–405.

NURSES'ASSOCIATION, ONTARIO (2013). "Collective agreement". In:

Organization), WHO(World Health (1993). *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*. Vol. 2. World Health Organization.

Pardalos, Panos M et al. (2013). *Systems analysis tools for better health care delivery*. Vol. 74. Springer Science & Business Media.

Patrick, Jonathan (2012a). "A Markov decision model for determining optimal outpatient scheduling". In: *Health care management science* 15.2, pp. 91–102.

– (2012b). "A Markov decision model for determining optimal outpatient scheduling". In: *Health care management science* 15.2, pp. 91–102.

Raghupathi, Wullianallur and Viju Raghupathi (2014). "Big data analytics in healthcare: promise and potential". In: *Health information science and systems* 2.1, p. 3.

Rais, Abdur and Ana Viana (2011). "Operations research in healthcare: a survey". In: *International transactions in operational research* 18.1, pp. 1–31.

Ross, Sheldon M (2014). *Introduction to probability models*. Academic press.

Sato, Renato Cesar and Désirée Moraes Zouain (2010). "Markov Models in health care". In: *Einstein (São Paulo)* 8.3, pp. 376–379.

Shafqat, Sarah et al. (2018). "Big data analytics enhanced healthcare systems: a review". In: *The Journal of Supercomputing*, pp. 1–46.

Sitompul, D and SU Randhawa (1990). "Nurse scheduling models: a state-of-the-art review." In: *Journal of the Society for Health Systems* 2.1, pp. 62–72.

Spetz, Joanne et al. (2008). "How many nurses per patient? Measurements of nurse staffing in health services research". In: *Health Services Research* 43.5p1, pp. 1674–1692.

Squires, David and Chloe Anderson (2015). "US health care from a global perspective: spending, use of services, prices, and health in 13 countries". In: *The Commonwealth Fund* 15.3, pp. 1–16.

Tanuja, S, Dinesh U Acharya, and KR Shailesh (2011). "Comparison of different data mining techniques to predict hospital length of stay". In: *Journal of Pharmaceutical and Biomedical Sciences* 7.7.

Taylor, GJ, SI McClean, and PH Millard (2000). "Stochastic models of geriatric patient bed occupancy behaviour". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163.1, pp. 39–48.

Taylor, Gordon, Sally Mclean, and Peter Millard (1996). "Geriatric-patient flow-rate modelling". In: *Mathematical Medicine and Biology: A Journal of the IMA* 13.4, pp. 297–307.

Trilling, Lorraine, Alain Guinet, and Dominiue Le Magny (2006). "Nurse scheduling using integer linear programming and constraint programming". In: *IFAC Proceedings Volumes* 39.3, pp. 671–676.

Van den Bergh, Jorne et al. (2013). "Personnel scheduling: A literature review". In: *European journal of operational research* 226.3, pp. 367–385.

Vasilakis, Christos and Adele H Marshall (2005). "Modelling nationwide hospital length of stay: opening the black box". In: *Journal of the Operational Research Society* 56.7, pp. 862–869.

Vassilacopoulos, G (1985). "A simulation model for bed allocation to hospital inpatient departments". In: *Simulation* 45.5, pp. 233–241.

Xu, M, TC Wong, and Kwai-Sang Chin (2014). "A medical procedure-based patient grouping method for an emergency department". In: *Applied Soft Computing* 14, pp. 31–37.

Yankovic, Natalia and Linda V Green (2011). "Identifying good nursing levels: A queuing approach". In: *Operations research* 59.4, pp. 942–955.

Zhang, Xin-li et al. (2013). "Forecasting emergency department patient flow using Markov chain". In: *2013 10th International Conference on Service Systems and Service Management*. IEEE, pp. 278–282.