Data-Driven Analytics to Support Scheduling of Multi-Priority Multi-Class Patients with Wait Targets

by

Yangzi Jiang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2016

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The aim of dynamic scheduling is to efficiently assign available resources to the most suitable patients. The dynamic assignment of multi-class, multi-priority patients over time has long been a challenge, especially for scheduling in advance and under non-deterministic capacity. In this paper, we first conduct descriptive analytics on MRI data of over 3.7 million patient records from 74 hospitals. The dataset captures patients of four different priority levels, with different wait time targets, seeking treatment for one of ten classes of procedures, which have been scheduled over a period of 3 years. The goal is to serve 90% of patients within their wait time targets; however, under current practice, 67% of patients exceed their target wait times. We characterize the main factors affecting the waiting times and conduct predictive analytics to forecast the distribution of the daily patient arrivals, as well as the service capacity or number of procedures performed daily at each hospital. We then prescribe two simple and practical dynamic scheduling policies based on a balance between the First-In First-Out (FIFO) and strict priority policies; namely, weight accumulation and priority promotion. Under the weight accumulation policy, patients from different priority levels start with varying initial weights, which then accumulates as a linear function of their waiting time. Patients of higher weights are prioritized for treatment in each period. Under the priority promotion policy, a strict priority policy is applied to priority levels where patients are promoted to a higher priority level after waiting for a predetermined threshold of time. To evaluate the proposed policies, we design a simulation model that applies the proposed scheduling policies and evaluates them against two performance measures: 1) total exceeding time: the total number of days by which patients exceed their wait time target, and 2) overflow proportion: the percentage of patients within each priority group that exceed the wait time target. Using historical data, we show that, compared to the current practice, the proposed policies

achieve a significant improvement in both performance measures. To investigate the value of information about the future demand, we schedule patients at different points of time from their day of arrival. The results show that hospitals can considerably enhance their wait time management by delaying patient scheduling.

Acknowledgements

Foremost, I would like to express my greatest and most sincere gratitude to my supervisor, Professor Hossein Abouee Mehrizi, for his unwavering support, patient guidance, and continuous encouragement throughout my study and research. I have been blessed with such an amazing supervisor who has great passion towards research, immense knowledge of the field, and endless patience for students. He has made my MASc. experience memorable and gave me confidence to continue PhD study in Northwestern University.

I really appreciate the members of my reading committee, Professor Frank Safayeni and Professor Houra Mahmoudzadeh for their helpful comments and suggestions. I want to thank Professor Keith W. Hipel and his family for their warm consolations and encouragements throughout my stay in Waterloo. I would also like to convey my thanks to Professor Pengfei Li and Professor Qi-ming He for providing valuable advices and sharing vast knowledge.

A very special thanks to Cancer Care Ontario for allowing us the access to the much needed data, answering all tedious questions, and providing valuable feedbacks.

I warmly thank my colleague Yuhe Diao, without whom my simulation code would be 200 lines longer and 3 months late. Forough Pourhossein, I am so lucky to have you as an officemate when I arrived at Management Sciences. You are like a big sister to me and I wish you the best of luck in your future PhD study at UBC.

There is a very special group of people - my beloved debate team. It is at the debate meeting where I met my best friends- Sally Xu, Kaixin Liu, Zhen Gao, Kaixuan Ma. It had been a remarkable four years we spent together. We may be on different pathes now, but I hope we all can look back to the days we had together with fondness and a big smile.

Lastly, but most importantly, I want to thank my mother Haiyan Xu and father Ju Jiang for their unconditional love and unwavering confidence. Even though we are thousands of miles apart, they always listened to my troubles, cheered for my successes and encouraged my ambitions. There had been some difficult times in my life, but their love gave me confidences to pursuit my dreams. They have made scarifies to give me the life I have now and I will be forever grateful.

University of Waterloo has been a home to me for the past decade. It not only provided two degrees for me, but more importantly, it made me who I am today. Undergraduate study in Faculty of Mathematics taught me to be meticulous and built a solid foundation and Master study in Faculty of Engineering encouraged my creativity and gave me wings to fly. When you reach for the stars, you may not quite get one, but you won't come up with a handful of mud either. This is not the end, but the beginning of a new journey, goodbyes are not forever, so until we meet again.

Dedication

This is dedicated to my parents who changed my life so one day I can change the world.

Table of Contents

List of Tables xi

List of Figures xii

1 Introduction 1

2 Literature Review 5

3 Problem Definition 8

4 MRI Operation Flow: Descriptive Analytics 11

4.1 Type of Admission 11

4.2 Performance 15

4.2.1 Exceeding Time 15

4.2.2 Overflow Proportion 16

5 MRI Demand and Capacity: Predictive Analytics 18

5.1 Data Description 19

5.2 Arrival Process 19

5.3 Daily Procedures 22

6 Dynamic Scheduling: Prescriptive Analytics **27**

6.1 Scheduling Policies 28

6.1.1 Weight Accumulation Policy 28

6.1.2 Priority Promotion Policy 29

6.2 Simulation Model 29

6.3 Policy Evaluation 30

6.3.1 Exceeding Time Measure 31

6.3.2 Overflow Proportion 34

6.4 Advance Scheduling 37

6.4.1 General Structure 37

6.4.2 Results of Advance Scheduling 39

7 Sensitivity Analysis and Aggregation of Result **42**

7.1 Sensitivity Analysis 43

7.1.1 Capacity 43

7.1.2 Capacity Required to Satisfy 90% of Patients within Their Targets 45

7.2 Robustness of the Proposed Policies 47

8 Discussion and Conclusion **49**

9 Future Work **53**

References **55**

APPENDICES **59**

A Obtaining Optimal Coefficients **60**

A.1 Optimal α and β 60

A.2 Optimal $T_{i(i-1)}$ 61

List of Tables

4.1 Average Waiting Time (in days) for 10 Sample Hospitals by Priority level . 13

4.2 Average Exceeding Time for 10 Sample Hospitals by Priority Levels 16

5.1 Goodness-of-fit Statistic for Daily Arrivals 22

5.2 Statistic Parameters for Daily Arrivals 23

5.3 Goodness-of-fit Statistic 25

5.4 Statistic Parameters 26

6.1 Exceeding Day Detail Data Analysis 32

6.2 Overflow Proportion Detail Data Analysis 36

6.3 Scheduled Allocation Detail Data Analysis 41

6.4 Scheduled Allocation Results 41

7.1 Sensitiveness of Capacity 44

7.2 Sensitiveness of Variance in Capacity 45

7.3 Required Capacity for 90% Patients Treatment 46

7.4 Aggregated Results 48

List of Figures

4.1 Proportion of Each Priority level 12

4.2 Proportion of Procedure Classes 14

5.1 Four Goodness-of-fit Plots for Four Distributions Fitted to Hospital R's Empirical Data for Arrival 21

5.2 Four Goodness-of-fit plots for four distributions fitted to Hospital R's empirical Data for Average Daily Procedures Performed 24

6.1 Graphic Comparison of the Historical Result vs. the Proposed Policies for Exceeding Time 33

6.2 Graphic Comparison of the Historical Result vs. the Proposed Policies for Overflow Proportion 35

6.3 Results of Scheduling 0, 2 and 8 days in Advance 40

Chapter 1

Introduction

Appointment scheduling is mainly used to manage access to services by matching the existing resources to the most suitable demand. Many factors can influence the service performance including patient arrivals, available resources, and scheduling policies. Patient arrivals can be influenced through, for example, online wait time announcement or clustering. Capacity deficiencies, which are often caused by a mismatch between demands and available resources, are mitigated through capacity optimization and resource allocation. The scheduling policy governs the process of assigning appropriate patients to the available services units. A good scheduling policy leads to better queue management, more efficient use of resources, and reductions in waiting times. Poor scheduling mechanisms, however, cause inefficient allocation of resources and inadequate access to services, leading to prolonged waiting times. In this paper, we mainly focus on the scheduling of multi-class, multi-priority patients where different priority levels have different wait time targets, and different classes have different service durations leading to uncertainty over daily number of procedures performed.

We analyze over 3,700,000 records of data, gathered from 74 hospitals providing MRI services over three years. Patients are of different priority and classes, with specific wait

time targets for each class. The hospitals' goal is to treat over 90% of patients within their wait time target; however, based on the available data, the percentage of patients who exceeded their target increased from 56% in 2012 to over 67% in 2015. To address this gap in achieving wait time target, we propose practical scheduling policies that reduce the waiting time compared to current scheduling mechanisms.

In order to better understand the true disparity between demand and capacity, we first analyze two aspects of the data: the daily patient arrivals and the daily procedures performed. Due to the variable durations of the different types of operation and the randomness of which types are performed each day, the number of procedures performed each day in MRI hospitals is non-deterministic. We use statistical distributions to estimate and forecast patient arrivals and daily procedures performed. The historical data is tested against a large number of standard distributions and, by means of various goodness-of-fit tests, the best-fitting models are selected. Using training and validation sets to eliminate overfitting and estimation errors, we construct a descriptive and predictive model that can reflect the actual daily patient arrivals and forecast the future demand.

Even though there is no universal standardized scheduling policy, there are two commonly used appointment scheduling methods: First-In, First-Out (FIFO) and strict priority. Patients' treatment order in the former policy is based solely on their arrival time, while the latter gives preferential treatment to high-priority classes over lower-class patients. Under FIFO, the urgency and short wait time target of higher priority patients are overlooked because patients from all priority classes are considered as equals. In contrast, under the strict priority policy, lower priority groups are not treated until the pool of higher priority patients is exhausted. Therefore, the prompt treatment of high priority patients are often at the expense of lower priority ones. Sometimes, treating lower priority patients before the higher priority ones is more efficient, as long as all higher priority patients are

still treated within their wait time target. We aim to find a balance between the two scheduling mechanisms, by obtaining a point in time when a lower priority patient, that has been waiting for a long time, can be treated before a higher priority patient.

We prescribe two practical simple scheduling policies to determine such a point in time; namely, weight accumulation and priority promotion. Under the weight accumulation policy, patients from different priority levels start with different initial weights, and accumulate more weight as a linear function of their waiting time. The accumulated weight is a combination of the initial score, which is a function of the patients' priority levels, and rate of increase, which is a function of their waiting time. The weight is used to determine the order of treatment for patients, whereby those patients with higher weights are treated first. For the priority promotion policy, we propose a strict priority scheduling policy with non-static priorities. The time-dependent and dynamically changing priority allow patients to be promoted to a higher priority group after waiting for a predetermined threshold of time. The markedly different average waiting time, as experienced by different priority groups under strict priority, is reduced by the priority promotion policy.

In addition to the scheduling policy, another important but often overlooked factor that influences the performance of a scheduling process is the timing of the scheduling. We analyze two common practices in dynamic scheduling: allocation scheduling and advance scheduling. Under the allocation scheduling policy, we delay the patient scheduling to the actual procedure date when all the information regarding the patient pool is available. The more predominate choice in the real world is advance scheduling of patients where they are scheduled several days prior to their service date. However, when based on the forecasted model, advance scheduling can result in sub-optimal performance in the service system. We compare the results of scheduling on the day of treatment with scheduling various days in advance, including on the day of arrival, to examine the value of information lost due

to advance scheduling. This demonstrates that, as we schedule further in advance, there is more unknown information regarding the daily patient arrivals and capacity, causing the occurrence of inefficient allocations and increased waiting times. When scheduling on the day of arrival, the smallest amount of information is known, which necessitates the greatest level of estimations and leads to the highest uncertainty and longest average waiting time. The gradual shift from complete lack of information when scheduling on the day of arrival, to comprehensive information under allocation scheduling demonstrates the improvement in average waiting time with the availability of additional pieces of information. The value of information can be exploited by hospitals to determine the optimal scheduling time in order to achieve a balance between efficient usage of resources and the benefits of advance planning.

Utilizing the data and forecasted distributions, we construct a discrete-time, multi-period simulation model. The simulation model replicates the patient inflow from the forecasted demand, determines the order of treatment based on the proposed scheduling policies, obtains daily procedures based on the forecasted distributions, and evaluates the results against the proposed performance measures. The two measures proposed incorporate intricacies arising from the existence of multiple priority classes with different wait targets. The exceeding time performance measure calculates the total number of days by which patients exceed their wait time target, while the overflow proportion performance measure computes the percentage of patients within each priority level that exceed their target.

By implementing the proposed scheduling policies in the simulation and evaluating them based on the performance measures, we achieve an improvement of over 30% in the rate of timely treatments. The results are validated against all 74 hospitals and similar level of improvement is observed in 72 of the hospitals.

Chapter 2

Literature Review

There has been a significant number of operations management studies on the appointment scheduling process and its effects on patient waiting time. It is concluded by Chakraborty et al. (2010) that poor scheduling has been a major source of operational inefficiency, while a good scheduling policy helps to set the pace of access to service. Viewed from the resource perspective, a well-established scheduling process helps to achieve a more efficient use of existing resources (Lowery and Martin 1989). A flexible and dynamic scheduling system can accommodate the rapid changes in demand and ensure that services are conducted in a timely fashion (Ahn and Hyun 1989). Our paper mainly focuses on dynamic allocation and advance scheduling policies for multi-priority multi-class patients with different wait time targets under a non-deterministic daily capacity.

In general, multi-period dynamic scheduling can be divided into allocation scheduling and advance scheduling. Multi-period dynamic allocation scheduling is studied fairly extensively. One method of dynamic appointment scheduling is resource allocation. Gerchak et al. (1996) implement dynamic scheduling for two classes of patients: emergency and

elective surgery cases. Their paper analyzes the means of determining the number of elective surgeries of uncertain durations to accept each day, under random daily emergency surgery requirements. Qu et al. (2007) meanwhile assess the optimal static percentage of daily resources allocated to patients from different priority groups. The daily capacity is distributed between open-access and scheduled patients to determine the optimal resource allocation methods. In Gupta and Wang (2008), instead of static daily allocation, allocated spots are dynamically changed to satisfy the daily needs and are dependent on requests from previous periods. Ayvaz and Huh (2010), using a dynamic programming approach, design a model that handles multiple classes of patients with different reactions to delayed services under a limited capacity. Luo et al. (2012) consider an appointment scheduling problem with no-shows and service interruption and investigate the consequences of patients not attending an appointment by analyzing the wasted appointment spaces. More recently, Truong and Ruzal-Shapiro (2015) combine these ideas to design a decision process to dynamically assign daily resources to patients from different priority classes with various wait time targets. Feldman et al. (2014) incorporate the patient' preferences into the decision process to enhance service experiences. The patients are no longer assumed to accept all appointment times offered; they can either choose from a range of available appointments or wait for a later time.

Another method of dynamic appointment scheduling is the priority queue whereby the patients' priorities are used to determine the order of treatments. Kleinrock and Finkelstein (1967) introduces the idea of a priority queue scheduling method that modified the classical structures of FIFO and strict priority. The proposed time-dependent priority queue changes the perception that low priority patients cannot receive service until all the high-priority patients have been served. Later, Hay and Valentin (2006) combine the idea of an initial score and accumulated priority as a linear function of the patients' waiting

time to determine the order of treatments. Stanford et al. (2014) further improves this model to incorporate multiple priority groups and analyze the waiting time distribution to obtain a maximum priority process. Min and Yih (2014) also employ a time-dependent priority to schedule patients based on their initial urgency as well as their waiting time. The effects of this policy have been tested under various conditions and over an infinite time horizon.

Unlike the allocation scheduling, there are few papers in the literature that consider advance scheduling. Kopach et al. (2007) introduce the idea of open access where patients are treated close to their appointment request date to avoid no-shows. They analyze how scheduling closer to treatment dates can help to improve the patients' access and reduce uncertainty. Gocgun and Ghate (2012) model the advance scheduling problem as a Markov decision process and develop an approximate dynamic programming method to solve it. This model is extended by Liu et al. (2012) to consider advance scheduling with no-shows and cancellations. Patrick (2012) further demonstrates that the dynamic advance scheduling model with no-shows can be improved with shorter booking windows.

Chapter 3

Problem Definition

Patients' health status can deteriorate over time; therefore, timely treatment is an ongoing concern for hospitals. Each MRI (magnetic resonance image) patient, based on their priority level, has a wait time target which, if exceeded, can lead to dissatisfaction and health deterioration. It is recommended in the National Maximal Wait Time Access Target for Medical Imaging that the 90th percentile treatment rate is the "preferred retrospective measure" within each priority level. However, our data analysis of 74 hospitals show that under current practice, over 67% of patients exceed their wait time target and many exceed it by more than 100 days.

There are two methods of tackling the issue of prolonged waiting time: capacity expansion and improvement on the scheduling policy. In 74 hospitals that we study, the number of patients requiring MRI scans has been rapidly increasing in recent years; however, the available equipment and facility capacities have not kept up with the increases in demand. The shortfall in capacity has exacerbated wait times. Therefore, the limitation on capacity constitutes the first obstacle to reducing the waiting time. Another factor affecting waiting

time is the scheduling policy. Currently, hospitals employ a loosely enforced prioritization strategy, aiming to treat higher priority patients first, regardless of how long lower priority patients have been waiting. This leads to a considerably longer wait time for the lower priority patients. In summary, the inefficient usage of resources, combined with the mismatch between demand and capacity, causes prolonged waiting times for MRI patients.

The higher priority patients, if served within their respective wait time target, are not in imminent danger. Therefore, lower priority patients, who have been waiting for a long time, can be treated first as long as the higher priority patients are still treated within their wait time target. We aim to determine the point at which a lower priority patient can be treated before a higher priority one. In this way, we will enable a high number of patients to be treated within their wait time target and reduce the total number of days patients exceed their wait time target.

First we conduct descriptive analytics on patient data from 74 hospitals to understand the provision of MRI services under current practice. We provide some insights into how the existence of different classes of scans increase the variability in the available capacity and how different mix of priority levels impact the waiting time at each hospital. We then conduct predictive analytics to forecast the daily demand patient arrivals for the MRI services and estimate the daily capacity (number of procedures) at each hospital, considering that the length of service varies based on the class of scan.

Finally, we conduct prescriptive analytics to provide practical solutions to improving the MRI services. Specifically, we model the problem as a dynamic scheduling problem that reduces the total percentage of patients exceeding their wait time target for multi-class, multi-priority patients with different wait time targets. There are two aspects of scheduling to be considered: the dynamic process of assigning patients with different priority levels to the required services, and the time at which patients are scheduled for their appointments.

We propose two practical, simple, and efficient policies that consider both the number of patients in each priority level who are waiting for treatment and the the waiting time of each patient at each point in time. The policies can dynamically assign patients of different priority level to a treatment time that reduces the total number of overflow patients. The timing of scheduling determines whether we employ allocation scheduling, in which we schedule on the actual procedure date, or advance scheduling, whereby appointments are scheduled several days in advance.

Even though we cannot fully eliminate the overflow patients, we can mitigate the problem significantly without the need for additional capacity. In order to satisfy the hospitals' goal of a 90% treatment rate, additional capacities are required. We also provide an estimation of the minimum capacity required to satisfy the 90% service rate.

Chapter 4

MRI Operation Flow: Descriptive Analytics

This section briefly describes the MRI operational process flow of the 74 MRI hospitals based on the data analyzed. A total of 13 entries without a treatment date, regardless of abandonment, no-shows, or postponement till later time, are excluded from the dataset. We do not distinguish between returning patients and new a patient arrivals; all individuals arriving at hospital are treated equally, without preference for returning patients.

4.1 Type of Admission

Patients are referred to one of the 74 hospitals for MRI treatment by their primary physicians who typically choose a hospital that is closest to the patient's residential location or by patients' location preference. Patients are assigned one of four priority levels based on their clinical urgency. The most significant distinction among the priority level are their

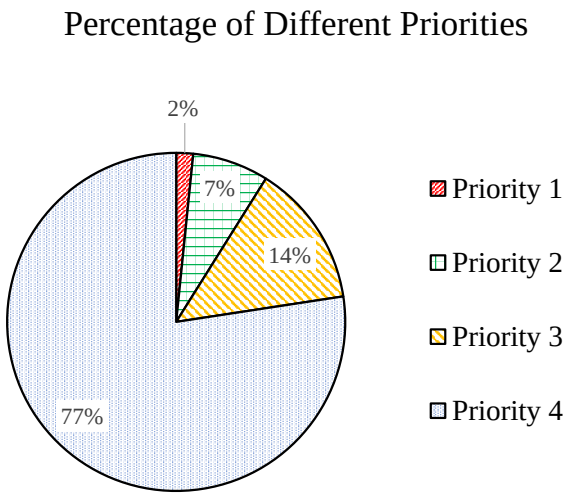

Figure 4.1: Proportion of Each Priority level

wait time targets. Priority 1, the highest priority levels, has a target of 24 hours, while the target for priorities 2, 3, and 4 are 48 hours, 10 days and 28 days, respectively. Figure 4.1 shows the four priority levels and the respective ratio of their demands based on our data analysis.

Figure 4.1 shows that over 77% of the patient arrivals are Priority 4 patients, while only 2% are Priority 1. However, despite the small population of Priority 1 patients, their extremely short wait time target makes their treatment a top priority in all hospitals. The percentage of patients exceeding their wait time target is more important than the actual number of patients exceeding. The average waiting time for 10 sample hospitals are calculated and summarized in Table 4.1.

As shown in Table 4.1, the average waiting time across hospitals varies significantly. For example, Hospital 6 and 7, each with two scanners available, operate for 8 hour every day.

Table 4.1: Average Waiting Time (in days) for 10 Sample Hospitals by Priority level

Hospital	Average WT	P1 Avg. WT	P2 Avg. WT	P3 Avg. WT	P4 Avg. WT
1	15.41	0.16	1.20	6.91	17.51
2	39.97	0.80	2.74	10.46	43.69
3	49.36	1.79	5.77	27.81	61.28
4	40.44	0	3.36	12.00	42.43
5	41.47	0.20	5.10	24.54	49.91
6	54.65	1.54	2.11	13.08	61.71
7	71.21	0.07	2.50	20.19	75.85
8	23.19	2.18	1.79	13.06	25.30
9	41.09	0.51	6.48	28.24	62.21
10	32.86	0.28	1.59	9.97	35.96

Hospital 6 has the highest average daily patient arrivals of 71.81 patients while Hospital 7 only has 20.25 average daily patient arrivals. However, Hospital 6's average treatment time of 54.65 days is much shorter than Hospital 7's 71.21 days. Therefore, the hospitals with high patient arrivals do not necessarily have longer waiting times, and small hospitals with fewer patient arrivals cannot necessarily guarantee quicker treatment. It is also observed in the data that hospitals within close proximity to each other do not necessarily have the same wait time. For example, Hospitals 8 and 9 are 20 minutes of driving distance apart, but differ in average waiting time to a notable degree.

The hospitals offer 10 classes of procedures: Abdomen, Breast, Cardiac, Extremities, Brain, Head and Neck, Pelvis, Peripheral Vascular, Spine and Thorax. The average pro-

portions of each class of procedures, for all hospitals, are illustrated in Figure 2. The actual scan durations may deviate from the scheduled duration. For example, heart procedure durations vary from 35 to 55 minutes and routine procedures' durations also fluctuate between 10 and 30 minutes.

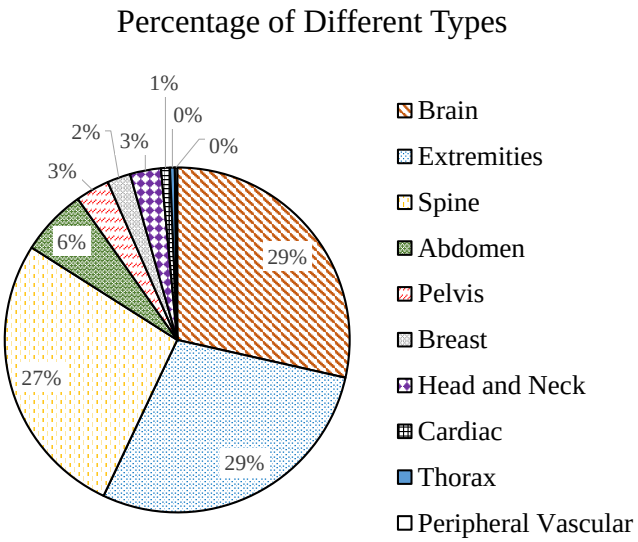

Figure 4.2: Proportion of Procedure Classes

As we can see from Figure 4.2, the most common MRI procedures performed are Head(Brain), Extremities and Spine. Even though the daily operating hours and average duration for each procedure is standardized, the class of procedures performed daily in each hospital does not follow any specific pattern. Therefore, the number of procedures performed daily in each hospital is not deterministic. On days with large quantities of long procedures, the total number of procedures performed that day will be significantly lower than on other days.

4.2 Performance

Historically, scheduling problems use waiting time, the time elapsed between the decision to order/request received date and the actual service date, as the most common performance measure in evaluating different scheduling policies. However, for patients with different wait time target, waiting time becomes a less effective performance measure. Column 2 in Table 4.1 displays the average overall waiting time for 10 sample hospitals. As indicated by the rest of the columns in Table 4.1, due to the varying wait time target for different priority levels, the comparison of average waiting time across patient priority levels is not meaningful. Hospital 2 and 4, for instance, display similar overall average wait times but significantly different levels of wait times by priority levels. Accordingly, we propose two additional performance measures that can be used to evaluate waiting time.

4.2.1 Exceeding Time

As displayed in Table 4.1, different priority levels' waiting times vary significantly. We propose exceeding time (ET), defined as the difference between the actual waiting time and their wait time target, as a more useful measure of the scheduling process than the wait time. One day of exceeding time indicates the same delay in service for all priority levels, while one day of waiting time has a different meaning among different priority levels. Table 4.2 displays the weighted average exceeding time for the same 10 hospitals as in Section 4.1 (weighted by the relative proportion of number of patients within each priority levels) and the average exceeding time for each priority levels within each hospital.

Table 4.2: Average Exceeding Time for 10 Sample Hospitals by Priority Levels

Region	Average ET	P1 Avg. ET	P2 Avg. ET	P3 Avg. ET	P4 Avg. ET
1	25.51	0.80	1.51	8.26	29.03
2	19.97	0.63	1.64	4.31	21.85
3	29.35	1.38	4.35	19.21	35.56
4	17.94	0	2.23	4.60	18.49
5	22.04	0.14	3.42	16.13	25.57
6	31.49	1.20	0.88	6.82	35.67
7	47.72	0	1.09	13.27	50.80
8	6.36	1.82	0.89	5.33	6.68
9	36.09	0.61	4.31	22.21	46.35
10	12.19	0.06	0.82	3.97	13.31

4.2.2 Overflow Proportion

When a patient exceed their wait time target, the patient is considered to be an overflow patient. Reducing the number of such patients is a main priority for hospitals. One way to assess the difference between overflow patients in different priority levels is by considering the overflow proportion, defined as the portion of patients within each priority levels that exceed their target. Using this as a performance measure, we can determine the gap between the hospital's target and its actual performance. The hospitals claim that the majority of their patients stay longer than their target wait time. This claim is supported by the observations from the data that over 67% of the patients exceed their target.

We observe that on average, the Priority 1 levels only has an overflow rate of 8.8%, while Priority 2, 3, and 4 have rates of 13.4%, 62%, and 71%, respectively. It can be concluded that the majority of the hospitals give significant preference to high priority patients while less emphasis is placed on and fewer preferential treatments are provided to the lower priority levels. The main reason for such decisions is that the high priority patients' health conditions may quickly deteriorate, whereas lower priority patients do not have high expectations or need for timely treatment and their health conditions do not tend to drastically change while they are waiting.

Chapter 5

MRI Demand and Capacity: Predictive Analytics

In this section, historical data analysis is provided and predictive models for forecasting the daily demand and capacity (number of procedures) are developed. The key dependent variable is the waiting time. There are two main factors that affect the dependent variable: daily patient arrivals and daily procedures performed. For the purpose of demonstration, a sample hospital (hereafter referred to as Hospital R), with average patient flow, is selected (we discuss the results obtained for the other 73 hospitals in Chapter 7). Historical data from Hospital R is used to forecast daily demand and capacity, as well as to evaluate the proposed scheduling policies outlined in the following sections. The research setting and the information contained in the data is discussed in Section 5.1. The estimation and prediction of the daily patient arrivals and daily procedures performed are demonstrated in Section 5.2 and Section 5.3, respectively.

5.1 Data Description

We use approximately two years' worth of MRI data for 18,105 patients treated in Hospital R with arrival dates starting from January 2, 2011 and continued for 642 days until October 5, 2012. The treatment dates could fall outside this time period. The records contain patient-level information including, but not limited to, the following: 1) priority score: a number between 1 to 4 that is assigned to patients upon arrival, indicating their clinical urgency; 2) wait time target: the number of days patients can wait without imminent risk to their current health condition, which is determined based on patient's priority score; 3) decision-to-treat date: the date the patient was added to the wait list (arrival date); 4) actual treatment date: the date the patient actually received their MRI procedures; 5) Class of MRI procedure: one of the ten classes of MRI procedures that patients may undergo. The waiting time is computed as the difference between the decision-to-treat date and the actual treatment date, while the exceeding time is computed as the difference between the waiting time and the wait time target.

5.2 Arrival Process

The patient arrival is the process of adding a request for an MRI procedures to the wait list. Understanding the pattern of patient arrivals and its fluctuations can help to construct the policies that bridge the gap between demand and capacity. In this section, we estimate the total number of arrivals per day independent of their priority levels. Then, in the simulation model discussed in Section 6.2, we use a Multinomial random variable based on the realized aggregate number of arrivals and the probability that a patient belongs to each priority levels to obtain the daily number of arrivals of each priority levels.

A wide range of discrete and continuous distributions are selected to compare against the historical data of Hospital R. First, we excluded the obvious mismatches, including Binomial, Degenerative, Geometric, and the like. A list of continuous distributions, including Beta, Logarithm, Gamma, Weibull, Normal, Half-Normal, LogNormal, and Exponential distributions are tested against the data. Only four of these, namely Gamma, LogNormal, Normal and Weibull, fitted the empirical data with relatively high accuracy and are analyzed in further detail.

The empirical distribution is fitted by the Maximal Likelihood Estimation ($L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$) (Le Cam 1990). Using statistical tools, we are able to obtain the estimated parameters, the estimated standard error, the loglikelihood, and the Akaika and Bayesian information criteria.

Figure 5.1 includes four graphs that demonstrate the goodness-of-fit for the four selected distributions. The density plot represents the density function of the fitted data against the empirical histogram distribution. The CDF graph plots the cumulated density function of both empirical data and fitted distribution. The Q-Q graph plots the fitted distribution's quantiles against the empirical data's quantiles. The P-P plot represents the empirical distribution function evaluated at each data point (y-axis) against the fitted distribution function (Delignette-Muller 2014).

Tables 5.1 and 5.2 provide summaries of the estimated parameters of the four selected distributions and their goodness-of-fit evaluation results. Three different testing methods, including the Kolmogorov-Smirnov Test ($sup|F_n(x) - F(x)|$), Cramer-von Mises Test ($n\int_{-\infty}^{\infty}(F_n(x) - F(x))^2 dx$), and Anderson-Darling Statistic ($n\int_{-\infty}^{\infty}\frac{(F_n(x)-F(x))^2}{F(x)(1-F(x))}dx$) are performed (Delignette-Muller 2014). It can be observed that the Normal distribution fits the data with the highest degree of accuracy.

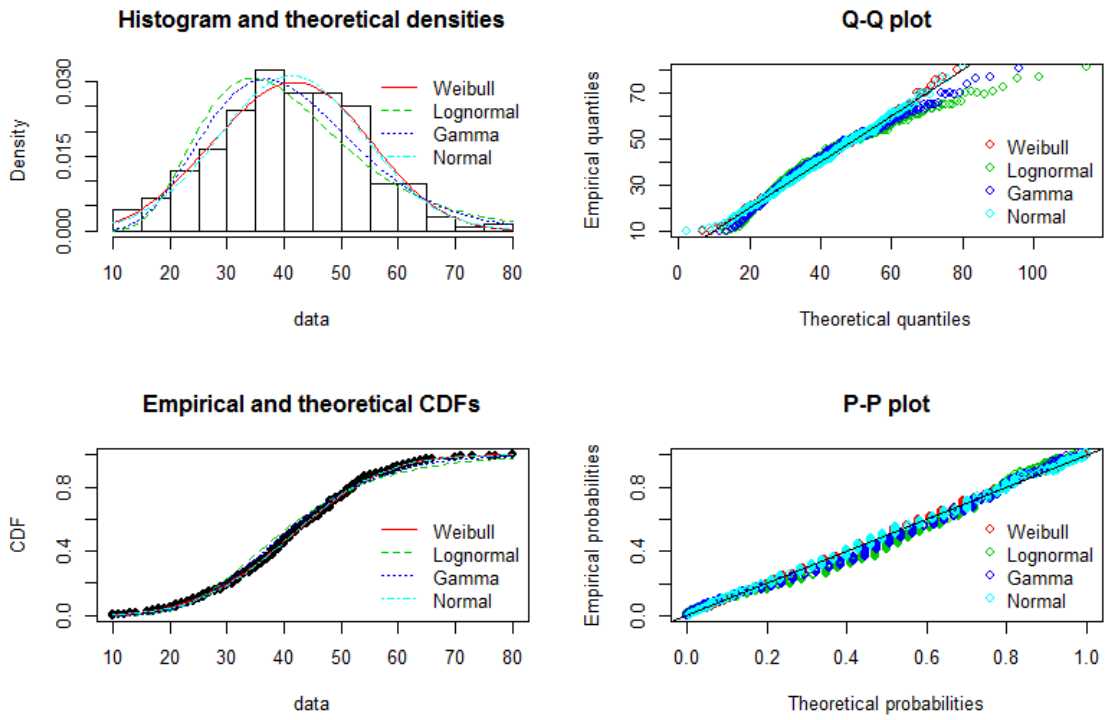

Figure 5.1: Four Goodness-of-fit Plots for Four Distributions Fitted to Hospital R's Empirical Data for Arrival

Therefore, a Normal distribution with parameters of μ =41.30 and σ = 12.81 is used to estimate Hospital R's daily number of patients' arrival. The Kolmogorov-Smirnov test provides a p-value of 0.718. In Kolmogorov-Smirnov tests, the p-value indicates the degree of compatibility between two distributions. The p-value represents the probability that the two cumulative frequency distributions will be similar if randomly sampled from identical populations. Therefore, a high p-value (larger than 0.5) indicates that the selected distribution can represent the empirical data to a high degree of accuracy (Lehmann 2006). The level of compatibility indicates that the estimated model is sufficient to replace the

Table 5.1: Goodness-of-fit Statistic for Daily Arrivals

Statistic Test	Gamma	Weibull	Lognormal	Normal
Kolmogorov-Smirnov	0.07694506	0.03834422	0.1004437	0.033568
Cramer-von Mises	0.47500763	0.0605623	0.9713341	0.04980234
Anderson-Darling	2.859687621	0.37056458	5.8741806	0.31821658
Goodness-of-fit Criteria	Gamma	Weibull	Lognormal	Normal
Aikake's Info Criteria	3428.719	3399.640	3469.135	3401.443
Bayesian Info Criteria	3426.837	3407.758	3477.254	3409.561

historical data in the simulation models. Since the daily patient arrival generated by the normal distribution can be a decimal value, it is truncated to the closest integer value.

In order to use the estimation of the historical data to forecast the future demand, we partition the data into training and validation sets. The first year's data is used as the training set and the remainder is used as the validation set. The estimation generated by the training set is cross-validated with those generated by the validation set to obtain an estimate with a low classification error. This error measures the estimated model against either a false negative where the estimated model reflects an event that did not happen or false positive response where the estimated model fails to capture an event that occurred.

5.3 Daily Procedures

The number of procedures performed each day reflects the capacity of the hospital and is a variable that influences the waiting time. Although each hospital has fixed operating

Table 5.2: Statistic Parameters for Daily Arrivals

	Estimated Parameters	Standard Error	Loglikelihood
Gamma	$k = 9.042$	0.607	-1712.359
	$\theta = 0.219$	0.015	-1712.359
Weibull	$k = 3.561$	0.133	-1697.82
	$\lambda = 45.850$	0.656	-1697.82
Lognormal	Meanlog = 3.665	0.017	-1732.568
	sdlog = 0.355	0.012	-1732.568
Normal	$\mu = 41.304$	0.619	-1698.721
	$\sigma = 12.808$	0.438	-1698.721

hours, due to the different durations for each class of procedures, the total number of procedures the hospital can perform each day is non-deterministic (as depicted in Figure 4.2 for Hospital R). Scan durations vary from 10 minutes for a minor assessment to over 4 hours for a particularly difficult heart MRI procedures.

Even though the daily classes of procedures and their respective durations cannot be accurately predicted, the daily number of procedures performed can be reliably estimated as the occasional long duration procedures can be offset by the short duration procedures. From the data, it is observed that the average daily procedures that hospital R performs can be described with a statistical distribution.

The same selections process as mentioned in Section 5.2 is utilized here. Again, the four statistical distributions, Gamma, LogNormal, Normal, and Weibull, fit the empirical data closely and are therefore analyzed in more details.

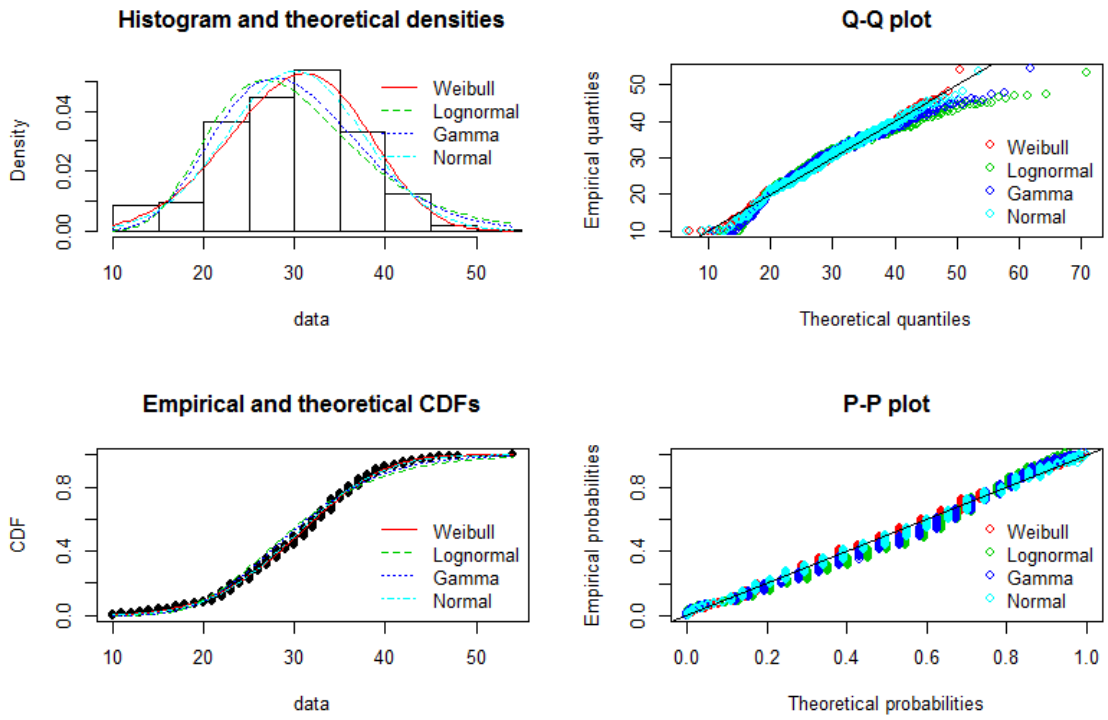

Figure 5.2: Four Goodness-of-fit plots for four distributions fitted to Hospital R's empirical Data for Average Daily Procedures Performed

Figure 5.2 includes the four graphs that are used to demonstrate the goodness-of-fit of the four selected distributions. The Weibull distribution is the best fit, with parameters $k = 4.58$ and $\lambda = 32.89$. The Normal distribution with parameters μ =30.05 and $\sigma = 7.47$ can also be used as its goodness-of-fit is very close to Weibull.

Table 5.3 and 5.4 contain summaries of the estimated distribution parameters and their goodness-of-fit evaluation results. The tests mentioned in Section 5.2 are performed for the goodness-of-fit. Compared to the daily patient arrivals, the daily number of procedures

Table 5.3: Goodness-of-fit Statistic

Statistic Test	Gamma	Weibull	Lognormal	Normal
Kolmogorov-Smirnov	0.09238612	0.04150228	0.1077583	0.05828262
Cramer-von Mises	0.92945912	0.13441592	1.5445012	0.23519675
Anderson-Darling	6.01227153	0.86836601	10.0074602	1.43810988
Goodness-of-fit Criteria	Gamma	Weibull	Lognormal	Normal
Aikake's Info Criteria	4160.118	4092.201	4216.335	4099.544
Bayesian Info Criteria	4168.902	4100.985	4255.119	4108.328

performed are more scattered and have more outliers. This is due to the different mix of classes of procedures performed each day as well as the fact that even though there is an expected duration for different classes of procedures, the actual treatment duration is rarely consistent with the expected duration and varies considerably; resulting in a highly variable daily number of procedures performed. Therefore, the results of the goodness-of-fit tests for estimating the number of daily procedures performed are not as strong as those for the estimations of daily patient arrivals. However, the distributions can be used as the basis for a rough estimation of the number of daily performed procedures that is required for simulation experiments in the latter sections.

Table 5.4: Statistic Parameters

	Estimated Parameters	Standard Error	Loglikelihood
Gamma	$k = 13.928$	0.797	-2078.056
	$\theta = 0.463$	0.027	-2078.056
Weibull	$k = 4.583$	0.146	-2044.101
	$\lambda = 32.892$	0.309	-2044.101
Lognormal	Meanlog = 3.367	0.012	-2106.168
	sdlog = 0.284	0.008	-2106.168
Normal	$\mu = 30.055$	0.306	-2047.772
	$\sigma = 7.472$	0.216	-2047.772

Chapter 6

Dynamic Scheduling: Prescriptive Analytics

In this section, we first propose two simple, practical dynamic scheduling policies that can be applied at any point from a patient's arrival until his or her actual service date, namely weight accumulation policy and priority promotion policy. We then apply these policies to the patient-level data of Hospital R on the date of the procedure, assuming that patients are kept on a waiting list meanwhile. Finally, we use the proposed policies to schedule patients in advance, at any time between the date of arrival and the actual procedure date. The numerical examples we obtained, based on the historical data, demonstrate the efficiency of the proposed policies compared to the current practice.

6.1 Scheduling Policies

In Section 6.1.1, we introduce a weight accumulation policy whereby the patient's weight, which is a function of their priority as well as the waiting time, is used to determine the order of treatment. In Section 6.1.2, we propose a priority promotion policy which is a strict priority policy with time-dependent priorities. Both policies provide a dynamic scheduling mechanism that can efficiently allocate the available resources.

6.1.1 Weight Accumulation Policy

This policy provides a mechanism to determine the point at which a lower priority patient can be treated before one of higher priority patient by finding a balance between the patients' priority and their waiting time.

Patients of Priority 1, 2, 3, and 4 arrive at the hospital at the beginning of each period. Upon arrival, they begin to accumulate weight at the following rate:

$$W_i(t) = \alpha \cdot p_i + \beta \cdot (t - A_i) \tag{6.1}$$

where $W_i(t)$ represents the accumulated weight of Patient i at time t. The arrival date of Patient i is denoted by A_i and their priority level upon the arrival is represented by p_i. The accumulated weight is used to determine the order of treatment, whereby patients with the highest accumulated weights are treated first.

The weight accumulation function is composed of two parts: the initial weight score and the accumulation rate. The former is a function of the patient's base priority while the latter is a function of the waiting time. Parameter $\alpha \cdot p_i$ denotes the initial weight score and parameter β indicates the rate at which the patient accumulates weight throughout their waiting time.

Note that when β is zero, the weight accumulation becomes the strict prioritization based on which high priority patients are always treated first. When α is zero, the accumulation function becomes a strictly FIFO policy wherein patients from all priority levels form a single queue, based purely on their arrival time.

6.1.2 Priority Promotion Policy

Based on this policy, patients are promoted to a higher priority level after waiting in the queue for a predetermined threshold of time, allowing them to be served sooner. Thresholds depend on the priority level and are determined so that the proportion of patients treated within their wait time targets is maximized.

Recall that a priority level is assigned to patients upon their arrival at the hospital and each priority level has a different wait time target. We define a threshold value $T_{i(i-1)}$ for priority level $i\,(= 2, 3, 4)$, so that once a patient's waiting time exceeds the threshold, they will be promoted to a higher priority level. This means that T_{21} represents the amount of time that Priority 2 patients will wait before they are promoted to Priority 1, T_{32} is the number of days taken for Priority 3 patients to be promoted to Priority 2, and T_{43} is the threshold for promoting Priority 4 individuals to Priority 3.

6.2 Simulation Model

We simulate the problem as a finite horizon multi-period scheduling system in which each period is defined as one day. At the beginning of each period, we forecast the total number of new patient arrivals based on the demand distribution obtained in Chapter 4. As mentioned, we then use a Multinomial random variable based on the realized total number

of arrivals and the probability that a patient belongs to each priority level to obtain the daily number of arrivals in each priority level. The new arrivals are added to the current queue, and their priority level is recorded. The patients' wait time targets are assigned according to their priority level. These patients will form the pool of all eligible patients available for treatment. The simulation model then uses the eligible patient pool as the input for the scheduling process. The information is fed to the proposed scheduling policies and the order of treatments is determined accordingly.

We assume that the scheduling system assigns patients of four priority groups to available servers in order to perform one of the ten classes of procedures. The scanners operate in parallel to and independent of each other, and each can perform all classes of procedures. Patients within each priority level are considered to be homogeneous in terms of clinical urgency and receive the same preference for treatment.

Once the patients receive their procedures, their final waiting time in the queue is recorded. The recorded wait times of all patients are then used to calculate the performance measures and evaluate the improvements brought by the proposed policies. The simulated results of the policies are then compared to those of the historical data to demonstrate the reduction in wait times.

6.3 Policy Evaluation

In this section we first define the objective function under each performance measure and obtain the thresholds as well as coefficients required to implement the scheduling policies. We then use the simulation model to evaluate the efficiency of the policies compared to the current practice.

6.3.1 Exceeding Time Measure

The exceeding time performance measures the total number of days exceeded by all patients in the given time horizon. The objective is to compare the proposed policies with the current practice under the exceeding time performance measure which is defined as,

$$\min_{p \in \Pi} \quad \sum_{D=1}^{T} D \times \mathbb{E}(M_D) \tag{6.2}$$

where T represents the total number of days in the time horizon, in the case of Hospital R, $T = 642$. Set Π represents the set of policies, including the proposed policies and the hospital's current practice. The number of days past the target date is denoted by D, while M_D is a random variable representing the number of patients who exceeded their wait time target by D days.

We first obtain the optimal thresholds and coefficients required to apply the proposed policies. For the weight accumulation function, there are two unknown coefficients, α and β, associated with it. Based on historical data from Hospital R, the set of (α, β) that minimizes the weight accumulation function under the exceeding time performance is $(6.063, 1.667)$. For the priority promotion policy, the thresholds obtained for the exceeding time measure are $T_{21} = 2, T_{32} = 3$, and $T_{43} = 3$. These thresholds mean that, for example, under the exceeding time measure the Priority 4 patients are promoted to Priority 3 after waiting in the queue for only 3 days. This is intuitively reasonable since for the exceeding time measure we are aiming to reduce the total number of days exceeding the target wait time. In Appendix A we discuss how we determine α, β, and $T_{i(i-1)}$ for $i = 2, 3, 4$.

Using the thresholds and coefficients obtained above, the two proposed policies are applied in the simulation model to determine the total exceeding time. To reduce the

variance in the simulation output for the total exceeding time, we take the average of the exceeding time for repeated simulation runs. We observe by taking the average of the output of 10 or more simulation runs the results converge such that the discrepancies in the outputs are negligible. Therefore, all the results discussed in the following sections are obtained as an average over 10 simulations runs. The results for the average number of patients exceeding their target by various number of days are summarized in Table 6.1.

Table 6.1: Exceeding Day Detail Data Analysis

Policy	<10 Days	10-19 Days	20-29 Days	30-39 Days	40-49 Days	≥ 50 Days
Empirical	2,111	1,927	2,642	2,274	2,213	2,239
Weight	2,326	1,296	1,621	1,349	308	15
Priority	2,351	1,848	1,352	1,032	228	0

The first column in Table 6.1 represents the policies and the rest of the columns display the number of patients exceeding their wait time target by the given number of days. For example, 2351 patients exceeded their wait time target for a duration fewer than 10 days under the priority promotion policy and 1621 patients exceeded their wait time target by 20-29 days under the weight accumulation policy. As shown in Table 7, both policies outperform the current practice. Under the priority promotion policy, no patient exceeded their wait time target by more than 50 days, and significantly fewer patients exceeded their wait time target by more than 30 days. Under the weight accumulation policy, fewer patients exceeded their wait time target by 10 to 20 days, but more exceeded their wait time target by 30 days or more.

The results of the simulation performed under the exceeding time measure are plotted

in Figure 6.1 which illustrates the number of patients exceeding their wait time target versus the number of days they exceeded their wait time target by.

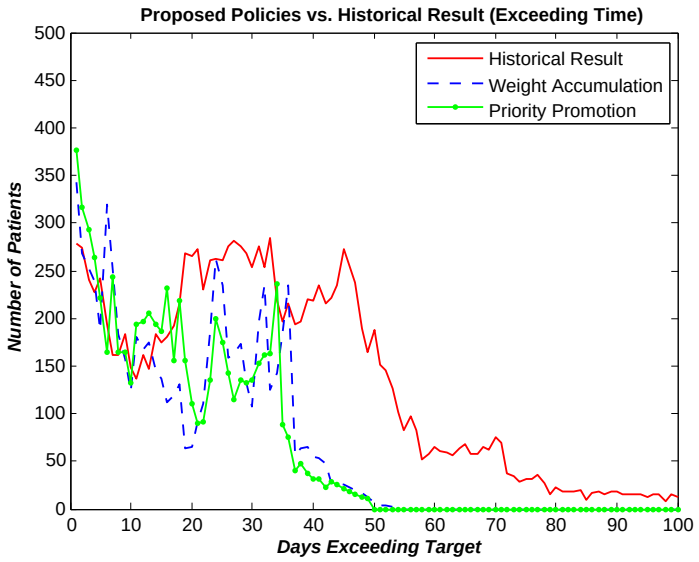

Figure 6.1: Graphic Comparison of the Historical Result vs. the Proposed Policies for Exceeding Time

The solid red line represents the exceeding time from the historical data evaluated under the exceeding time performance measure. The dashed line represents the results obtained using the weight accumulation policy, while the dotted line represents the results from the priority promotion policy. Considering that under the exceeding time measure the goal is to reduce the area under the curves depicted in Figure 6.1, it is evident that both proposed scheduling policies outperform the historical data.

Note that the proposed policies may increase the number of patients exceeding their wait time target for a few days as illustrated in Figure 6.1. But, they significantly reduce the number of patients who have exceeded their wait time target for a long time. Table

6.1 demonstrates that there is no significant differences between the performances of the proposed policies under the exceeding time measure and Figure 6.1 also shows that the area under the dashed curve is almost the same as the one under the dotted curve.

6.3.2 Overflow Proportion

In this section, we examine the performance of the overflow proportion measure. Under this measure, once patients exceed their wait time target, the number of exceeding days is no longer relevant. To reduce the proportion of patients exceeding their wait time target for each priority level, we use the following objective function

$$\min_{p \in \Pi} \quad \sum_{i=1}^{4} c_i \times \mathbb{E}(\frac{S_i}{N_i}) \tag{6.3}$$

where c_i indicates the relative importance of each priority level, S_i is a random variable representing the number of priority i patients who exceed their wait time target date, and N_i is a random variable that represents the total number of patients of priority level i that arrive at the hospital throughout the time horizon. The objective function calculates the weighted percentage of patients over all priority level that exceeded their wait time target. The value of c_i demonstrates the penalty of overflow for the various priority groups and acts as weighting factor in the weighted average. In our numerical examples, we set the values of c_i's relative to their wait targets, $c_1 = 28/41$, $c_2 = 10/41$, $c_3 = 2/41$, and $c_4 = 1/41$ where 41 is the total wait targets of all classes (1+2+10+28=41). The longer the wait time target, the less urgency is placed upon the patients' procedure times as their health is not likely to drastically deteriorate.

Using the objective function defined in 6.3, we obtain coefficients $\alpha = 10.067$ and

$\beta = 0.667$ for the exceeding time performance measure as well as promotion thresholds $T_{21} = 2, T_{32} = 8$, and $T_{43} = 21$ for the priority promotion performance measure. Note that under the overflow proportion measure more emphasis is placed on higher priority level, as the relative importance of the lower priority level exceeding wait time target is low. Therefore, the threshold at which patients of Priority 4 are promoted to Priority 3 under the overflow proportion measure, $T_{43} = 21$, is much higher than the one under the exceeding time measure, $T_{43} = 3$.

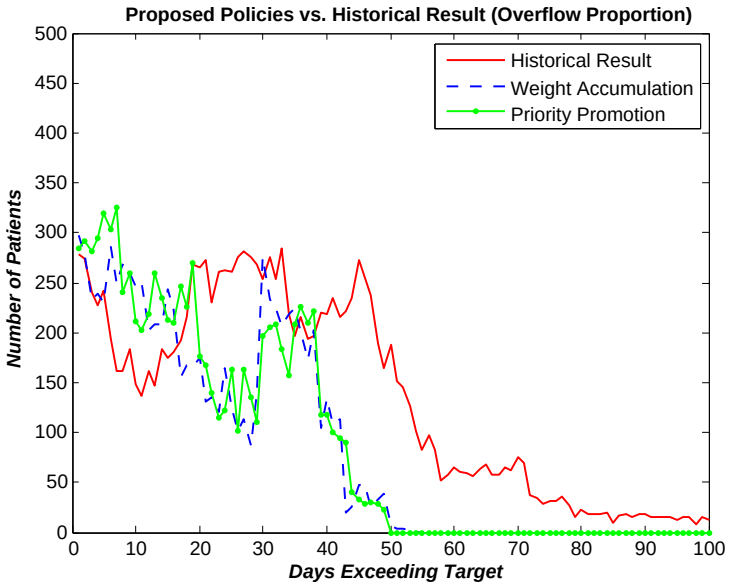

Figure 6.2: Graphic Comparison of the Historical Result vs. the Proposed Policies for Overflow Proportion

As Figure 6.2 illustrates, both proposed scheduling policies increase the number of patients who wait less than 15 days more than their wait targets compared to the current practice. However, they reduce the number of patients with a long waiting time notably.

Table 6.2 shows the decrease of overflow proportion under the proposed policies. In

Table 6.2: Overflow Proportion Detail Data Analysis

Policy	Exceed P1	Exceed P2	Exceed P3	Exceed P4	Overflow	Treatment
Empirical	68 (21%)	589 (46%)	2,539 (68%)	10,210 (80%)	0.3514	26%
Weight	13 (4%)	460 (36%)	1,560 (42%)	4,882 (38%)	0.1787	61.81%
Priority	15 (5%)	467 (36%)	1,525 (41%)	4,802 (37%)	0.1842	62.40%

Table 6.2, Columns 2-5 display the number and percentage of patients within each priority level that exceeded their wait time target. The treatment rate, displayed in the last column, is defined as the percentage of all patients treated within their wait time target (independent of their priority level).

As it is evident from the table, both policies outperform the current practice against the overflow proportion measure and there is no significant difference between the performances of the proposed policies.

Comparing the two proposed policies based on Table 6.2, we observe that under the priority promotion policy, slightly more high priority patients (Priority 1 and 2) exceeded their wait time target in exchange for a reduction in the number of Priority 3 and 4 patients who did so. By promoting lower priority patients, some high priority patients are treated a few days late, while more low priority patients can receive treatment sooner.

6.4 Advance Scheduling

The results discussed in Section 6.3 are obtained based on the allocation scheduling, whereby we schedule the patients on the actual procedure date. However, a more predominant paradigm in general practice is advance scheduling.

6.4.1 General Structure

Hospitals serve four priority levels of patients from which Priority 1 and 2 require immediate care, Priority 3 patients need to be served within 10 days, and Priority 4 patients can wait up to 28 days. We consider a dynamic advance scheduling model with two patient groups; an urgent demand class, who should be served on the day of arrival, and a regular demand class, who can be served at a future date. The *urgent class* includes Priority 1 and 2 patients while the *regular class* encompasses Priority 3 and 4. At the beginning of each period, the patients from the urgent class are scheduled for treatment on that day, while the patients from the regular class are kept in the queue until they can be scheduled for service. The number of days by which we are scheduling in advance determines when the regular class patients can receive notification for their treatment. For example, if we are scheduling two days in advance, patients will be notified on day t that they will be treated on day $t + 2$. We assume that patients will take the earliest appointments offered and expect to receive service on the day of their appointment.

A fixed amount of the daily capacity is always set aside for the patient arrivals of urgent class of patients. This allocation will not be used for other purposes, regardless of whether the daily patient arrivals in the urgent class exhaust the allocation. The purpose of this

allocation is to ensure that the urgent class patients can receive treatment immediately. Advance scheduling policies are not applied to urgent class patients, as they are scheduled upon their arrival and treated on the same day.

The detailed procedure for the advance scheduling process used in the simulation is as follows.

1. Advance scheduling policies are not applied to urgent class patients, as they are scheduled upon their arrival and treated the same day. Therefore, a fixed amount of the daily capacity is always set aside for the urgent class of patients. Due to capacity limitation or other factors, if some urgent class patients are not treated on the day of their arrivals, they will treated as soon as possible as it is explained below.

2. Due to over-scheduling or machine malfunction inefficient scheduling, etc., delays in the treatment process including deviations from expected durations of procedures, or other complications, it is possible that not all scheduled patients or urgent patients are treated on their appointment date. Therefore, we then schedule patients who were scheduled for the previous periods and have not been treated. This allocation is used to ensure all patients left from previous periods are treated as soon as possible.

3. Regular class patients are scheduled several days prior to their appointment. At the end of each period, a group of patients with the highest priority as determined by the proposed policies are scheduled for an appointment on a promised date. Due to uncertainty, if some scheduled patients are not treated on their appointment day, they will be treated as soon as possible as it is discussed in 2.

6.4.2 Results of Advance Scheduling

The analysis performed in Section 6.3 forms the base case for advance scheduling, wherein we schedule 0 days in advance, also known as allocation scheduling. Scheduling further in advance requires better forecasting and creates higher uncertainty, resulting in more inefficient planning. Therefore, the main concern is whether the trade-off between early notification and delays in the treatment process is acceptable.

We perform the simulation for scheduling on the day of treatment, 2, 4, 6, 8, and 10 days in advance, as well as scheduling on the day of arrival. In Figure 6.3, a graphical display of the result is generated and the trade-off between the number of days to schedule in advance and the increase in waiting time is demonstrated. However, to better illustrate the increase in waiting time, only the allocation policy, 2 days and 8 days in advance scheduling policy, and the historical data are displayed in the figure.

The solid line represents the historical data. The dashed line denotes the allocation policy on the day of treatment, the dotted line represents scheduling 2 days in advance, and the diamond one embodies scheduling 8 days in advance.

From Figure 6.3, we observe that the gap between the current practice and the proposed scheduling policies decreases as the number of days that the patients are scheduled in advance increases. For scheduling on the day of arrival, the area under the historical data is total 337,782 days for all patients, and the area under the diamond line is 242,853 days. The 94,929 days reduction in total exceeding time is a result of the additional information obtained over the 10 days prior to the day of the treatment. The gap between scheduling 8 days in advance and 2 days in advance is 84,121 days of exceeding time. Similarly, the exceeding time reduced by 43,040 days when we switched from scheduling two days in

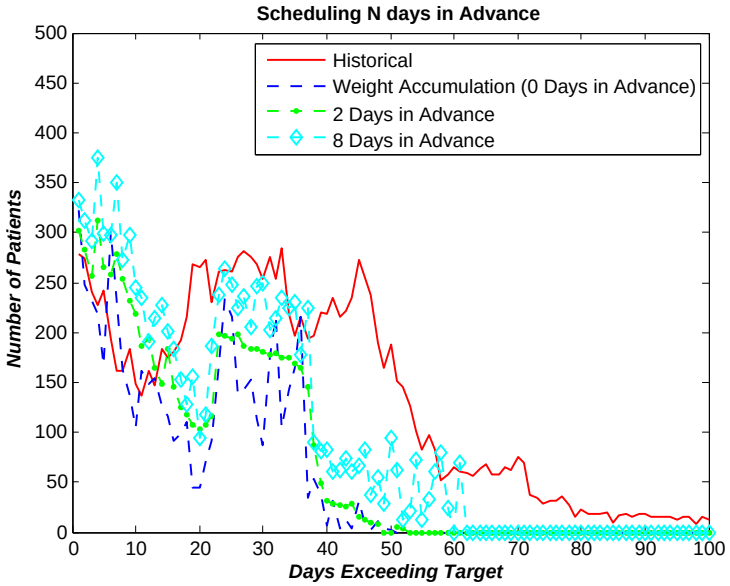

Figure 6.3: Results of Scheduling 0, 2 and 8 days in Advance

advance to scheduling on the day of operation.

It can be observed from Table 6.3 that the number of Priority 2 patients exceeding their wait time target date decreases under the advance scheduling policy for 2 days and 4 days. This is due to the urgent class allocation, which ensures that the majority of higher priority level patients are treated immediately. However, the regular class patients tend to wait longer under advance scheduling, and their waiting time increases as we schedule further in advance. The value of information is the main reason for the differences between the waiting times for the various number of days by which we schedule in advance. Scheduling on the date of treatment (allocation scheduling) allows for the greatest amount of information to be gathered, therefore enabling most efficient allocation of resources. However, when scheduling in advance, incomplete information can lead to inefficient scheduling.

As Table 6.4 indicates, regardless of the performance measure used, the waiting time

Table 6.3: Scheduled Allocation Detail Data Analysis

	Priority Promo	2 Days	4 Days	6 Days	8 Days	10 Days	On Arrival
Priority 1	13	14	14	15	16	16	18
Priority 2	457	455	455	458	457	462	582
Priority 3	1,572	1,689	1,854	1,982	2,134	2,546	3,171
Priority 4	4,848	5,871	5,964	6,061	6,098	6,358	9,523

Table 6.4: Scheduled Allocation Results

	Weight	2 Days	4 Days	6 Days	8 Days	10 Days	On Arrival
Exceeding	115,692	158,732	182,742	209,265	242,853	274,334	337,782
Overflow	0.1787	0.1830	0.1854	0.1902	0.1941	0.2013	0.2532

and number of patients exceeding their wait time target increases as we schedule further in advance. Therefore, it is up to the hospital and the patients to determine to what degree they are willing to delay their treatment in exchange for greater convenience in scheduling.

Chapter 7

Sensitivity Analysis and Aggregation of Result

In this section, we analyze how the waiting time can be influenced by changes in the parameters and distributions. We also examine the robustness of the proposed policies by applying them to data from other hospitals. In section 7.1, the sensitivity of the inputs used in the scheduling policies, including the distribution of the daily number of performed procedures and the daily patient arrivals, is discussed. In section 7.2, the policies are applied to others hospitals and the generalizability of the results is discussed. We demonstrate that the majority of the hospitals display similar patterns of improvement, regardless of the parameters of their input data and available capacity.

7.1 Sensitivity Analysis

7.1.1 Capacity

As mentioned in the first section, the limitation on the daily performed procedures is one of the bottlenecks of the congestion problem. Capacity expansion can be achieved through two methods: extending the daily operating hours or increasing the number of MRI scanners available at each hospital. However, there are constraints on the daily operating hours and each MRI scanner is extremely costly. Therefore, it is crucial to analyze the impact on waiting time per unit of increase in capacity.

First, we analyze how a change in capacity influences the weight accumulation policy under the two objective functions. Keeping all other input values constant, including the daily patient arrival rate, the following results regarding various capacities are obtained. All the results obtained are an average of ten simulation runs and are subjected to minor variations due to randomness of the numbers generated.

In Table 7.1, the first, third, fifth and seventh rows display the number of Priority 1, 2, 3, and 4 patients that exceeded their wait time target, respectively. The second, fourth, sixth and eighth rows represent the percentage of Priority 1, 2, 3, and 4 patients that exceeded their wait time target, respectively. In Table 7.1, $\aleph(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The second column ("Current") indicates the values obtained under Hospital R's current capacity and the next four columns indicate the overflowing results for four different average daily procedures performed. We observe that the overflow proportion is very sensitive to changes in the daily treatment numbers. When the current daily capacity with the average of 30.05 days is reduced by three procedures, the

Table 7.1: Sensitiveness of Capacity

Number of Patients	Current	א(27.5, 7.5)	א (31.5, 7.5)	א (35.5, 7.5)	א (40.5, 7.5)
Priority 1's Overflow	13	27	11	4	0
% Overflow	4.00%	8.31%	3.38%	1.23%	0
Priority 2's Overflow	460	621	358	289	0
% Overflow	35.63%	48.64%	27.73%	22.39%	0
Priority 3's Overflow	1560	2312	1218	872	0
% Overflow	41.83%	62.00%	32.66%	23.38%	0
Priority 4's Overflow	4882	6924	4966	4721	0
% Overflow	38.26%	54.26%	38.92%	37.00%	0

overflow percentage increases significantly. When the average daily number of treatments reaches 40.5, all patients will be treated within their wait time target. However, to increase the average capacity from the current level of 30.05 to the desired level of 40.5, we either need to extend the operation hour from 8 hours per day to 10.5 hours or to increase the number of MRI scanners.

Another important factor in improving the wait time is reducing the variance of the daily performed procedure number. In the following table, the sensitivity of waiting times to changes in variance is displayed.

From Table 7.2, the first, third, fifth and seventh rows display the number of Priority 1, 2, 3, and 4 patients that exceed their wait time target, respectively. The second, fourth, sixth and eighth rows represent the percentage of Priority 1, 2, 3, and 4 patients that

Table 7.2: Sensitiveness of Variance in Capacity

Number of Patients	Current $\aleph(30.05, 7.47)$	$\aleph(30.05, 4.50)$	$\aleph(30.05, 2.47))$	$\aleph(30.05, 0)$
Priority 1's Overflow	13	13	11	10
% Overflow	4.00%	4.00%	3.38%	3.08%
Priority 2's Overflow	460	405	386	379
% Overflow	35.63%	31.37%	29.90%	29.36%
Priority 3's Overflow	1560	1485	1348	1296
% Overflow	41.83%	39.82%	36.15%	34.75%
Priority 4's Overflow	4882	5187	4963	4818
% Overflow	38.26%	40.65%	38.89%	37.76%

exceeded their wait time target, respectively. The first column indicates the values obtained under Hospital R's current capacity and the second, third and fourth columns indicate the overflowing results for the same average daily performed procedures but different variances in the distribution. It is observed that the overflow proportion is relatively insensitive to the change in variance.

7.1.2 Capacity Required to Satisfy 90% of Patients within Their Targets

As mentioned before, the hospital's target is to treat over 90% of its patients within the

target waiting time. However, as calculated in Section 6.3 with the proposed scheduling policies, regardless of the timing of the policy (allocation scheduling or advance scheduling), we are unable to satisfy the desired 90% treatment rate. Therefore, it is important to estimate the capacity that can achieve the desired level of patient treatment. We observe that once the average capacity (number of performed procedures) increases to 38.5, the percentage of patients treated within their wait time target rises to 90%. Also, if the average capacity is increased to 36.75 while the variance decreases to 4.5, 90% of patients are satisfied within their wait time target. The results of the simulation under the new capacities are summarized in Table 7.3.

Table 7.3: Required Capacity for 90% Patients Treatment

Number of Patients	Current $\aleph(30.05, 7.47)$	$\aleph(38.5, 7.47)$	$\aleph(36.75, 4.50)$
Priority 1's Overflow	13	0	0
% Priority 1 Overflow	4.00%	0	0
Priority 2's Overflow	460	13	14
% Priority 2 Overflow	35.63%	10.00%	10.1%
Priority 3's Overflow	1,560	378	379
% Priority 3 Overflow	41.83%	10.14%	10.16%
Priority 4's Overflow	4,882	1,240	1,245
% Priority 4 Overflow	38.19%	9.72%	9.76%

Our numerical results show that if the hospital extends its operating hour by an additional 1.5 hours, 90% of patients will be satisfied within their wait time target. Variance reduction, accompanied by a smaller increase in capacity, is beneficial in the long-run as it

establishes a stable pattern that can be continued in the future.

7.2 Robustness of the Proposed Policies

The given dataset contains 74 hospitals, each with a different patient arrival rate and capacity. In the previous sections, Hospital R is randomly selected to test the effectiveness of the data. It is observed that the proposed scheduling policies, regardless of allocation scheduling or advance scheduling method, are able to reduce the overflow percentage. These policies are applied to the other 73 hospitals, of which 71 hospitals show a similar level of improvement in the exceeding time and the overflow proportion performance measures. The reason for the two exceptions is that one hospital has an extremely low daily patient arrival rate. Under their current systems, the daily patient arrivals often cannot fully exhaust their capacity. Therefore, there is no visible improvement under our proposed policies. Another hospital performs a significantly lower number of daily procedures than the capacity is able to handle. This can be the result of malfunctioning MRI machines. There is no significant improvement in this hospital under our proposed policies since the high congestions cannot be eliminated without significant capacity expansions or better scanner operation.

Several other hospitals, each with a different patient pool and MRI capacity, are analyzed and evaluated. The results of the two policies applied on the day of treatment are summarized in the following table.

Table 7.4 displays the results obtained from four randomly selected hospitals. The total number of days exceeded by all patients for the given time horizon, the number of patients exceeding their wait time target, and the proportion of patients exceeding

Table 7.4: Aggregated Results

	Hospital R	Hospital V	Hospital T	Hospital S
Overflow Proportion (Current)	0.3514	0.3278	0.0573	0.2169
Overflow Proportion (Weight)	0.1787	0.1117	0	0.1395
Overflow Proportion (Priority)	0.1842	0.1155	0	0.1486
Exceeding Number (Current)	436,165	287,854	1,268	20,265
Exceeding Number (Weight)	115,692	194,532	3	11,105
Exceeding Number (Priority)	109,759	185,633	0	9,843

their wait time target are significantly reduced under allocation scheduling policies of both weight accumulation and priority promotion policies. Similar results are also obtained for advance scheduling, but the details are not discussed here. It is concluded that the policies can be applied to, but not limited to, all hospitals with similar MRI scheduling structures and selection methods. The only differences in each hospital are the coefficients of the accumulation functions, the thresholds for priority promotion, and the relative importance of each priority level. Once those values are obtained, the evaluation and selection process will be identical for every hospital.

Chapter 8

Discussion and Conclusion

This study contributed to our understanding of how scheduling policies can be applied to effectively reduce the waiting time in MRI hospitals, so that the gaps between actual waiting times and wait time target for all priority levels are reduced. The essence of our contribution is to understand the current status of MRI services based on data from 74 hospitals, forecast the demand for MRI services, and provide two practical scheduling policies that can drastically reduce wait times compared to current practice.

We first analyzed patient-level MRI data from 74 hospitals and found that variability in the daily capacity (number of daily performed procedures) is very high, since the service durations for different types of procedures vary significantly. In addition, we observed that patients of Priority 1 and 2, who should be treated in 24-48 hours, respectively, account for a small proportion of patients. We then estimated the distributions of the number of daily patient arrivals at each hospital and showed that the accuracy of the estimation was high. Since the service duration varies for different classes of procedures and their future demand is unknown, we also estimated the distribution of the number of performed

procedures each hospital can perform daily.

We finally proposed two scheduling policies that aim to reduce the waiting time for the patients: the weight accumulation policy and the priority promotion policy. The weight accumulation policy assigned patients from various priority level a different initial score based on their priority level. Throughout their wait in the queue, they accumulated weight as a function of their waiting time. The total accumulated weight is a function of the combination of their priority and their waiting time. For the priority promotion policy, we used a modified version of strict priority scheduling. After waiting in the queue for a predetermined threshold of time, the patients were promoted to a higher priority level. This policy was implemented and formed a desirable ratio between high and lower priority patients' treatments.

We introduced two performance measures: the exceeding time and the overflow proportion. The former evaluated the total number of days by which the patients exceeded their wait time target. The latter calculated the percentage of patients within each priority level who exceeded their wait time target. These two performance measures were used to evaluate the effectiveness of the proposed policies. It was concluded that under both measures, the two scheduling policies outperformed the current practice. The overflow percentage decreased from 67% to less than 34%, while the waiting times were reduced from over 200 days for some patients to less than 54 days for all patients. When comparing the two scheduling policies, we observed that the total number of patients exceeding their wait time target in both policies was very similar. Under a more detailed analysis, the composition of the overflow patients changed. The priority promotion policy had more Priority 1 patients exceeding their wait time target by a short period of time in exchange for a greater number of Priority 4 patients reducing their waiting time significantly.

The timing of scheduling also played an important role. For allocation scheduling, the

patients were scheduled on the actual procedure date, while for advance scheduling, they were scheduled a varying number of days in advance. We observed a gradual reduction in the number of patients treated in their wait targets as we shifted from scheduling on the day of arrival to doing so on the actual procedure date. The shifts reflected the additional information gained by delaying the scheduling of patients: as we schedule further in advance, we have less information regarding the patient arrivals and the available resources. The trade-off between the convenience of advance planning and the increased waiting time under advance planning should be considered by hospitals in determining the optimal time to schedule.

A sample hospital R was used to demonstrate how we successfully forecasted the future demand and the daily number of performed procedures. Using the estimations, we constructed a simulation that reflected the process of deciding which patients to treat. The simulation demonstrated reductions in prolonged waiting times and the high proportion of overflow patients under proposed scheduling policies. The simulation also demonstrated the trade-off between the days to schedule in advance and the increase in the waiting time.

In order to achieve the hospitals' goal of serving over 90% of patients within their wait time target, we analyzed the sensitivity of the controlled parameters. The capacities required by each hospital to achieve their goal were computed. The extension of the operation hour by 1.5 hours each day resolved the congestion problem for hospital R and ensured a smooth patient inflow and outflow. Unfortunately, the increase in capacity is quite expensive and assessing its feasibility is beyond the scope of our analysis.

In general, our model efficiently resolved the problem of prolonged waiting times by proposing a standardized selection process. Even though we were not able to achieve the goal of a 90% service rate, we reduced the waiting time. Through our discrete-time, multi-period simulation model, we validated our policies against all 74 hospitals. Similar

rates of improvements and significant reductions were demonstrated in 72 of the hospital, indicating that our proposed scheduling process may help to resolve the current congestion problems in MRI hospitals in Ontario.

Chapter 9

Future Work

Firstly, in order to incorporate multi-class, multi-priority patients with non-deterministic capacity, we utilize simulation models to validate our model and obtain numerical results. For both allocation and advanced scheduling, we use data to validate the improvements in waiting time. However, we are unable to obtain an analytical solution. Based on the capacity constraint and scheduling targets, we hope to find a close-form, analytical solution to the two proposed objective functions in the future.

Secondly, for application purposes, the dynamic scheduling models of resource assignments are simplified. The transition phase between the current stage and future stage is not optimal. The current advance scheduling policy allows for fixed daily allocations to the urgent class and the regular class patients. However, the daily requirement for urgent class patients is not stationary. It will improve the performance further if the daily allocation of resources to each group can be dynamically determined and recursively generated.

Thirdly, as mentioned before, grid search methods are used to find the optimal α, β, and $T_{i(i-1)}$s. However grid search methods often suffer from the "curse of dimensionality" when

the dataset becomes too large. We hope to improve the parameter searching techniques by incorporating other innovative machine learning methods including Bayesian optimization or genetic algorithms. This way we can increase the accuracy of the optimal parameters, reduce the computation time, and obtain better results under the performance measures.

Fourthly, resource pooling is the next step we will pursuit to further reduce the waiting time. We can view the 74 hospitals or a subset of the hospitals as a network. The resources and patients can be shared and transferred between the hospitals. Currently, the hospitals operate independent of each other and information are not exchanged between them. Some hospitals' MRI scanners are idling while a neighbouring hospital is suffering from overcrowding. By forming a resource pooling between the hospitals, we can integrate the available MRI scanners together to achieve optimality on the aggregated level.

Lastly, we do not partition the patients base on their scan types during the scheduling process. The patients are divided into four priority groups and assume to be homogenous within each priority group. However, a further classification of patients based on both priority groups and scan classes can reduce the uncertainty in the daily number of procedures performed and the overall exceeding time. A clustering mechanism can be employed to divide the patients into groups where they share similar patterns.

References

References

[1] Ahn DS, Hyun CJ, Ahn JK, Yim CY. 1989. Impaired Interleukin-2 Receptor Expression on Lymphocytes from Patients with Chronic Active Hepatitis Type B. *Korean J Intern Med.* **4(1)** 34-41.

[2] Ayvaz N, Huh WT. 2010. Allocation of Hospital Capacity to Multiple Types of Patients. *Journal of Revenue & Pricing Management* **9(5)** 386-398.

[3] Bergstra J, Bengio Y. 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* **13(2012)** 281-305.

[4] Canadian Association of Radiologists. 2013. National Maximum Wait Time Access Targets for Medical Imaging (MRI and CT). *Canadian Association of Radiologists*.

[5] Chakraborty S, Muthuraman K, Lawley M. 2010. Sequential Clinical Scheduling with Patient No-shows and General Service Time Distributions. *IIE Transaction* **42(5)** 354-366.

[6] Delignette-Muller M, Lyon de U, Dutang C, Strasbourg U. 2014. fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software* **64(4)** 1-34.

[7] Feldman J, Liu N, Topaloglu H, Ziya S. 2014. Appointment Scheduling under Patient Preference and No-show Behavior. *Operations Research* **62(4)** 794-811.

[8] Gerchak Y, Gupta D, Henig M. 1996. Reservation Planning for Elective Surgery under Uncertain Demand for Emergency Surgery. *Management Science* **42(3)** 321-334.

[9] Gocgun Y, Ghate A. 2012. Lagrangian Relaxation and Constraint Generation for Allocation and Advanced Scheduling. *Computer & Operations Research* **39(10)** 2323-2336.

[10] Gupta D, Wang L. 2008. Revenue Management for a Primary-care Clinic in the Presence of Patient Choice. *Operations Research* **56(3)** 576-592.

[11] Hay AM, Valentin EC, Bijlsma RA. 2006. Modeling Emergency Care in Hospitals: A Paradox - The Patient Should Not Drive the Process. *Proceedings of the 2006 Winter Simulation Conference* pp. 439-445.

[12] Kleinrock L, Finkelstein RP. 1967. Time Dependent Priority Queues. *Operation Research* **15** 104-116.

[13] Kopach R, DeLaurentis PC, Lawley M, Muthuraman K, Ozsen L, Rardin R, Wan H, Intrevado P, Qu X, Willis D. 2007. Effects of clinical characteristics on successful open access scheduling. *Health Care Management Science* **10(2)** 111-124.

[14] Le Cam L. 1990. Maximum Likelihood: An Introduction. *International Statistical Review***58(2)** 153-171.

[15] Lehmann E, Romano J. 2006. Testing Statistical Hypotheses. *Springer*.

[16] Liu N, Ziya S. 2012. Panel Size and Overbooking Decisions for Appointment-based Services under Patient No-shows. *Production and Operations Management* **23(12)** 2209-2223.

[17] Lowery J, Martin JB. 1989. Evaluation of an Advance Surgical Scheduling System. *Journal of Medical System* **13(1)** 11-23.

[18] Luo J, Kulkarni VG, Ziya S. 2012. Appointment Scheduling under Patient No-shows and Service Interruptions. *Manufacturing & Service Operation Management* **14(4)** 670-684.

[19] Min D, Yih Y. 2014. Managing a Patient Waiting List with Time-dependent Priority and Adverse Events. *RAIRO-Opeartions Research* **48(01)** 53-74.

[20] Patrick J. 2012. A markov decision model for determining optimal outpatient scheudling. *Health care management science* **15(2)** 91-102.

[21] Qu X, Rardin R, Williams J, Willis D. 2007. Matching Daily Healthcare Provider Capacity to Demand in Advanced Access Scheduling Systems. *European Journal of Operational Research* **183 (2)** 812-826.

[22] Shi P, Chou MC, Dai JG, Ding D, Sim J. 2015. Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time. *Management Science* **62(01)** 1-28.

[23] Song H, Tucker AL, Murrell KL. 2015. The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay. *Management Science* **61(12)** 3032-3053.

[24] Stanford DA, Taylor P, Ziedins I. 2014. Waiting time distributions in the accumulating priority queue. *Queueing System* **77** 297-330.

[25] Truong V, Ruzal-Shapiro C. 2015. Optimal Advanced Scheduling. *Management Science* **61(7)** 1584-1597.

APPENDICES

Appendix A

A.1 Optimal α and β

In-order to obtain the set of α and β that creates the optimal solution, a traditional hyper-parameter optimization approach is employed (Bergstra 2012). Hyperparameter optimization is where a set of hyperparameters is selected for a model with the goal of optimizing the results of such algorithm. In this case, a grid search method (parameter sweep) is used to obtain the optimal parameters. There are two inputs associated with grid search techniques: the boundary and the grid step. The boundary is the upper and lower bounds of the parameter values. The grid step determines the number of steps we need to cycle through to obtain an optimal solution.

The boundary of the grid space for optimal set of (α, β) is determined as follows. The lower bound of the grid space is (0,0), since the parameters cannot take a negative value. To determine the upper bound, we first realize that the waiting time and exceeding proportion both converge. The point of convergence is at $\alpha = 16$, $\beta = 2$. For all $\alpha \geq 16$ and $\beta \geq 2$, the weight accumulation policy would produce the same result for the objective function. Therefore, the boundary of the grid space is set as $\alpha \in [0, 16], \beta \in [0, 2]$. The grid step

is set as 1/6 which means the solution space is divided into 96×12 small grid steps for determining the optimal coefficients.

The parameter sweep starts with the obvious worst choice $\alpha = 0$, $\beta = 0$, where all patients are selected at random to receive treatments. The value obtained through random selection is assigned as the cross-validating set where the value found from next evaluation cycle can be compared against. The grid search then cycle through (α, β) determined by the grid steps and updates the cross-validating set each time a set of better performing parameters is found. After the completion of the grid search, the optimal coefficient pair is obtained. The optimal coefficients are unique to each hospital and different when calculated under different performance measures.

A.2 Optimal $T_{i(i-1)}$

As defined before $T_{i(i-1)}$, $i = 2, 3, 4$ are the thresholds of days Priori ty i patients have to wait before promoted to Priority $i - 1$. The lower boundaries for all $T_{i(i-1)}$ are 0 as the patients cannot be promoted before they arrive at the queue. The upper boundary of each $T_{i(i-1)}$ are defined as the target wait time for priority $i + 1$ patients. The reason for the upper bound is that once the patients overstay their target, there is no reason to promote them to a higher priority group. Therefore, the range for $T_{i(i-1)}$'s are as follows: $T_{21} \in [0, 2], T_{32} \in [0, 10]$, and $T_{43} \in [0, 28]$. The size of the grid steps is set to 1, since the priority promotion only happens at the beginning of each period.

The parameter sweep starts with the obvious worst choice of $T_{21} = 2$, $T_{32} = 10$, $T_{43} = 28$, the patients are never promoted to a higher priority group before they had already exceeded their target waiting time. The value obtain through random selection is

assigned as the cross-validating set where the value obtained from next evaluation cycle can be compared against. The grid search then cycle through (T_{21}, T_{32} and, T_{43}) determined by the grid steps and updates the cross-validating set. Similar to parameters α and β, the optimal $T_{i(i-1)}$'s are unique to each hospital and different when calculated under different performance measures.