Reducing Disruptive Effects of Patient No-shows: A Scheduling Approach

Mingang Fu

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Richard Storch, Chair

Norman Beauchamp

Archis Ghate

Program Authorized to Offer Degree:

Industrial and Systems Engineering

University of Washington

Abstract

Reducing Disruptive Effects of Patient No-shows: A Scheduling Approach

Mingang Fu

Chair of the Supervisory Committee:
Professor Richard Storch
Industrial & System Engineering

Appointment scheduling systems have been studied for nearly 60 years. From a decision making point of view, related problems can be classified into two categories: static and dynamic. In a static scheduling problem, all decisions are made before a clinic session starts; in a dynamic scheduling problem, the schedule of future arrivals is revised constantly during the clinic session. I categorize my problem as static. Within the research field of static appointment scheduling, little attention has been paid to patient no-show until the past decade. As an important aspect of patient arrival behaviors, the phenomenon of patient no-show has resulted in huge economic loss industry wide. I aim to explore static scheduling approaches to alleviate negative effects of patient no-show, with consideration of nonhomogeneous patients, overbooking, and nonconventional patient waiting cost structure.

One primary contribution of this dissertation is a static analytical model I developed for the problem of scheduling patients to queues with consideration of quadratic patient waiting costs, nonhomogeneous patient no-show probabilities, and nonhomogeneous patient waiting cost ratios. By relaxing the assumptions of constant and identical no-show probabilities and waiting cost ratios, Hassin and Mendel's [25] model becomes a special case of my model. Another major contribution lies in my study on a set of heuristics that sequence patients based on their no-show probabilities. My numerical studies on three heuristics suggest scheduling patients with higher no-show probabilities in front of patients with lower no-show probabilities. It achieves best overall system performance as well as patient waiting performance. Last, I integrated the static model with a nonconventional overbooking strategy to formulate a problem with a hybrid overbooking model which not only determines number of patients to schedule but also determines scheduled inter-arrival times. It enables outpatients, inpatients, and emergency patients to be considered within a static scheduling environment. By comparing performances of three booking heuristics, I recommend scheduling inpatients first when no-show probability is low, while scheduling outpatients first when no-show probability is high.

Patient waiting is reported to be an important index of patient satisfaction in various surveys. Almost all appointment scheduling studies assume a linear relationship between patient waiting cost and patient waiting time, which might not be correct. The waiting cost of a system with one patient waiting for 40 minutes is not equal to another one with 20 patients each waiting for 2 minutes [34]. Furthermore, it also involves issues of goodwill, service, and "costs to the society", which place a value on patients' waiting time [8]. Therefore, a nonlinear cost structure of patient waiting is desired. To control the complexity of target problems, a majority of the static

scheduling literature assumes homogenous patients, which might be oversimplified. For the same amount of waiting time, waiting cost varies from one patient to another, due to various occupations held by different patients. Similarly, no-show probability needs to be patient specific as it's determined by various patient level attributes (Age, sex, marital status, income, appointment delay, etc.).

I solve a static scheduling problem with patient no-show probability varied among patients. To represent the nonlinear nature of the relationship between waiting cost and patient waiting time, I formulate the objective function as a total of quadratic patient waiting cost and linear sever idle cost. By comparing it to a model with linear waiting cost, I find quadratic waiting cost may change my decision of sequencing patients when no-show probability is nonhomogeneous. I solve another problem with both patient no-show probability and patient waiting cost ratio varied among patients, and compare the performance of three no-show probability based booking heuristics: lower no-show first, higher no-show first, and higher no-show in the middle, with the purpose of providing simplified heuristics to medical scheduling practices.

Next, I address a daily scheduling problem of allocating relatively flexible diagnostic capacities among three categories of patients: inpatients, who have low level of no-show probability and waiting cost ratio; outpatients, who have medium level of no-show probability and waiting cost ratio; and emergency patients, who usually show up as walk-in, with extremely high waiting cost ratio. To incorporate walk-in emergency patients into the model, I employ an overbooking strategy with server overtime allowed. The objective is to maximize system performance in terms of net revenue which consists of service revenue, server idle cost, patient waiting cost, and patient deny penalty cost. I analyze the model from three perspectives: behavior of optimal

schedules, overall system performance, and customer experience. To make the model easy to apply, I analyze the model performances under three heuristic booking strategies: all outpatient, inpatient first and outpatient first, with three environmental factors (outpatient no-show probability, equipment hourly idle cost, and inpatient service fee) are varied. The system is found to perform better when server hourly idle cost is greater. This phenomenon is more significant when outpatient no-show probability is relatively low. For clinics which also schedule inpatients, I recommend using the inpatient first policy when outpatient no-show probability is low; and using outpatient first policy when outpatient no-show probability is high. To a certain extent, overbooking can alleviate the negative effects brought by patient no-show, but system performance still decreases as no-show probability increases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Throughout my Ph.D. studies, I have received encouragement and generous support from many kind people around me, without whom it would not be possible for me to complete this dissertation.

Above all, I would like to express my deepest gratitude to my Ph.D. supervisor, Prof. Ricard Storch. The insights and support I received from him has been invaluable to my academic advancement and personal development. His guidance on various aspects such as choosing a research direction, setting a bar for my dissertation, and developing methodologies has made this dissertation a thoughtful and fruitful journey for me. At the times when I was struggling to balance between life, research, and work, his patience and friendship greatly helped me continue to persevere.

I would like to acknowledge the academic and technical support of the department of Radiology at the University of Washington, especially my reading committee member Prof. Norman Beauchamp, for his enthusiasm of my research topic and insights on choosing a right modality as my research subject. I also thank Mr. Daniel Lane and Mr. Erik Christianson from the University of Washington Medical Center for their support in my work flow observation and empirical data collection.

My thanks also go to the member of my reading committee, Prof. Archis Ghate, for sharing his ideas on improving the general structure of my dissertation, and Prof. Santosh Devasia for

accepting to be a GSR, and Prof. Christina Mastrangelo for accepting to be my committee member.

I would like to thank the following colleagues at UW ISE and Amazon Outbound Transportation for their sincere friendship: Wei Wang, Pengbo Zhang, Lihui Shi, Hongrui Liu, Lei Chen, Huan Nguyen, Stephen Swan, Eric Jones, Michael Campion, Raghav Mehra, Marc Armstrong, Yumin Deng, Di Wu. My thanks also go to Daisy Fu, my friend in China, for her trust and understanding.

Finally, I would like to thank my parents, who always supported me during my entire academic career.

DEDICATION

To my family

Chapter 1 **INTRODUCTION**

1.1. Background

In 2009, the total U.S. health expenditure reached $2.5 trillion, which represents 17.6% of the nation's Gross Domestic Product (GDP) [10]. Although decelerating, the U.S. health care spending growth rate was constantly higher than the GDP growth rate in the past 10 years. Obviously, more public investment is needed to finance health care, which makes it a huge burden for the country. In such a fast-growing industry, hospitals that fail to maintain cost effective operations struggle to survive financially [21]. Therefore, health care providers face a great deal of pressure to reduce costs and to improve efficiency.

On the other hand, timely access to medical services becomes a significant business concern in the health care industry. A study by Strunk and Cunningham [62] shows an increase from 24.4% to 27.4% for general examination appointments with Appointment Delay longer than three weeks. In a major report on health care quality conducted by the Institute of Medicine [30], “timeliness” is identified to be a key aim for improvement. Patient waiting is costly, not only because of the direct economic losses they cause, but also because of the potential losses in patient satisfaction they may bring. Therefore, it is regarded as key to delivering good clinical outcomes (i.e. volume of patients receiving services); it is also reported to be an important index of patient satisfaction in various surveys. Hospitals and clinics have been struggling to reduce long patient waiting time. Depending on time horizon, patient waiting can be classified into two types: indirect waiting and direct waiting. Indirect waiting, also known as appointment delay, is

defined as the time span between an appointment request is made and actual appointment time; direct waiting is defined as the time a patient spends in the queue, waiting for receiving medical services. In this study, I focus on direct waiting, which I name as patient waiting for the rest of this paper, unless the contrary is explicitly stated.

One plausible way to provide timely access appears to be hiring more physicians and supporting staff, which is not only costly but also unrealistic in terms of the insufficient supplies. Thus, it's challenging to maintain cost effectiveness while offering timely access to patients. Appointment scheduling systems deal with balancing between maintaining cost efficiency and providing timely access to care. An effective scheduling system can help reach a reasonable level of appointment delay/patient waiting without increasing costs.

1.2.Motivation

Since the 1950s, numerous studies on appointment scheduling have been done. However, within this research field, little attention has been paid to patient no-show until 2000s. As an important aspect of patient's arrival behaviors, the phenomenon of patient no-show has a significant impact on overall efficiency of the appointment scheduling systems. High no-show rates have been reported to result in huge economic loss industry wide. Pesata et al. [56] reported 14,000 missed appointments in a pediatric practice which resulted in an estimated loss of over a million dollars. Moore et al. [49] reported 31% of the total appointments at a family practice clinic were missed or cancelled, which caused a corresponding estimated a loss between 3% and 14% of annual revenues. BBC News [3] reported 12 million general practice appointments missed in the UK in a single year, which costs 250 million GBP. Since 2000s, appointment scheduling systems with consideration of patient no-show have been intensively studied. Nevertheless, a majority of the

literature has assumed constant and identical no-show probabilities. In medical practices, it's unrealistic to estimate no-show probabilities of different patients with a fixed value, as it ranges from as low as 3% to as high as 80% [1]. Even within the same service type, it varies significantly depending on various other factors (appointment delay, patient attributes etc.). There is a need for nonhomogeneous no-show probabilities in static appointment scheduling problems.

As an important component of overall appointment scheduling system performance, patient waiting cost has been assumed by almost all related literature to be linearly related with patient waiting time, which might not be correct. As pointed out by Klassen and Rohleder [34], the waiting cost of a system with one patient waiting for 40 minutes doesn't equal to another one with 20 patients each waiting for 2 minutes. From the point view of heath care providers, the former is usually considered as a quality incident, which costs much more than the latter, as tolerated waiting. Furthermore, it also involves issues of goodwill, service, and "costs to the society", which place a value on patients' waiting time [8]. Even for the same amount of waiting time, waiting cost varies from one patient to another, due to various occupations held by different patients. Therefore, nonlinear patient waiting cost and nonhomogeneous waiting cost ratios are desired to provide a more realistic representation of patients.

The radiology department of a hospital typically faces demand from three groups of patients: (i) outpatients, who are scheduled to come, usually featured with low level of no-show probability and waiting cost; (ii) inpatients, who either wait for a call or are scheduled with reserved slots, usually featured with medium level of no-show probability and waiting cost; and (iii) emergency patients, who show up as walk-in, featured with extremely high waiting cost, are usually served

depending on urgency. However, limited research attention has been paid to the static scheduling problems dealing with inpatients, outpatients, and emergency patients.

1.3. Contributions

I first demonstrate that there is a need for quadratic patient waiting cost measurement, from operational, psychological, and social perspectives. With this purpose, I formulate a static scheduling problem which determines patient inter-arrival times to minimize a total of quadratic patient waiting cost and linear server idle cost, with patient no-show probability and waiting cost ratio fixed. By enumerating values of no-show probability, I find that sensitivity of scheduled patient inter-arrival time to patient no-show probability varies from one patient to another, with the inter-arrival times among first several patients more sensitive to no-show probability than the inter-arrival times among the last several patients. I relax the assumption of constant and identical no-show probabilities used in previous studies. By comparing the performance of three no-show probability based booking heuristics, I recommend scheduling patients with higher no-show probabilities to the first several slots of a clinical session. This heuristic can help clinics achieve better system performance while reducing overall and patient specific waiting times. I further relax the assumption of constant and identical patient waiting cost ratio. Under the new model, the higher no-show first heuristic still outperforms the other two heuristics, in terms of overall system cost; although it has the highest waiting time it achieves the shortest completion time. I observe that the higher no-show first is a robust heuristic against a wide range of parameters.

The generalized model developed in Chapter 4 can be well applied to a practical case with inpatients and outpatients, as both of these two types of patients can be scheduled in advance, no-

show probabilities and waiting cost ratios could be estimated based on patient attributes. In Chapter 5, to include walk-in emergency patients in a static scheduling problem, I develop a hybrid overbooking model with relatively flexible appointment capacity, by relaxing the assumption of scheduling fixed number of patients, including service revenue and patient deny costs in the objective function, and allowing a certain level of server overtime. Under the new model, I find that the optimal schedules do not follow the well-known dome shape, moreover, expected patient waiting time tends to increase non-monotonically, which is contradictory with my observation from the previous work without overbooking in Chapter 4. For a very limited improvement on system net revenue, the optimal solution results into huge sacrifice of patient waiting, therefore, it's not recommended to book patients to maximum. The numerical study results of three type based patient sequencing policies indicate that the model performs better when server hourly idle cost is greater. This phenomenon is more significant when outpatient no-show probability is relatively low. For clinics which also schedule inpatients, I recommend using the inpatient first policy when outpatient no-show probability is low and using outpatient first policy when outpatient no-show probability is high. To a certain extent, overbooking can alleviate the negative effects brought by patient no-shows, but system performance still decreases as no-show probability increases.

1.4. Outline of the dissertation

The remainder of this dissertation is organized as follows.

Chapter 3 gives review of relevant literature.

Chapter 4 first solves a static scheduling problem with fixed no-show probability. Under the quadratic waiting cost assumption, the effects of variable waiting cost ratios and no-show probabilities are analyzed. It then relaxes the assumption of constant and identical no-show probabilities and compares the performance of three no-show probability based booking heuristics: lower no-show first, higher no-show first, and higher no-show in the middle. It another problem with both patient no-show probability and patient waiting cost ratio varied among patients. The following three questions are answered in this chapter

- How should patients been sequenced based on estimated no-show probabilities?
- Does quadratic waiting cost affect the decision on sequencing patients? If yes, how?
- Do variable waiting cost ratios impact the decision on sequencing patients? If yes, how?

Chapter 5 extends the generalized queuing model developed in Chapter 4 to a hybrid overbooking model. The objective is to maximize system net revenue which consists of service revenue, server idle cost, patient waiting cost, and patient deny penalty cost. Solutions are analyzed from three perspectives: behavior of optimal schedules, overall system performance, and customer experience. The performance of the model is tested under three heuristic booking policies: all outpatient, inpatient first and outpatient first, with three environmental factors (outpatient no-show probability, server hourly idle cost, and inpatient service fee) are varied.

Chapter 6 summarizes my major findings throughout this dissertation and discusses their potential implementations to practices.

The dissertation ends in Chapter 7 with potential directions for future research extensions.

Chapter 2 **METHDOLOGY**

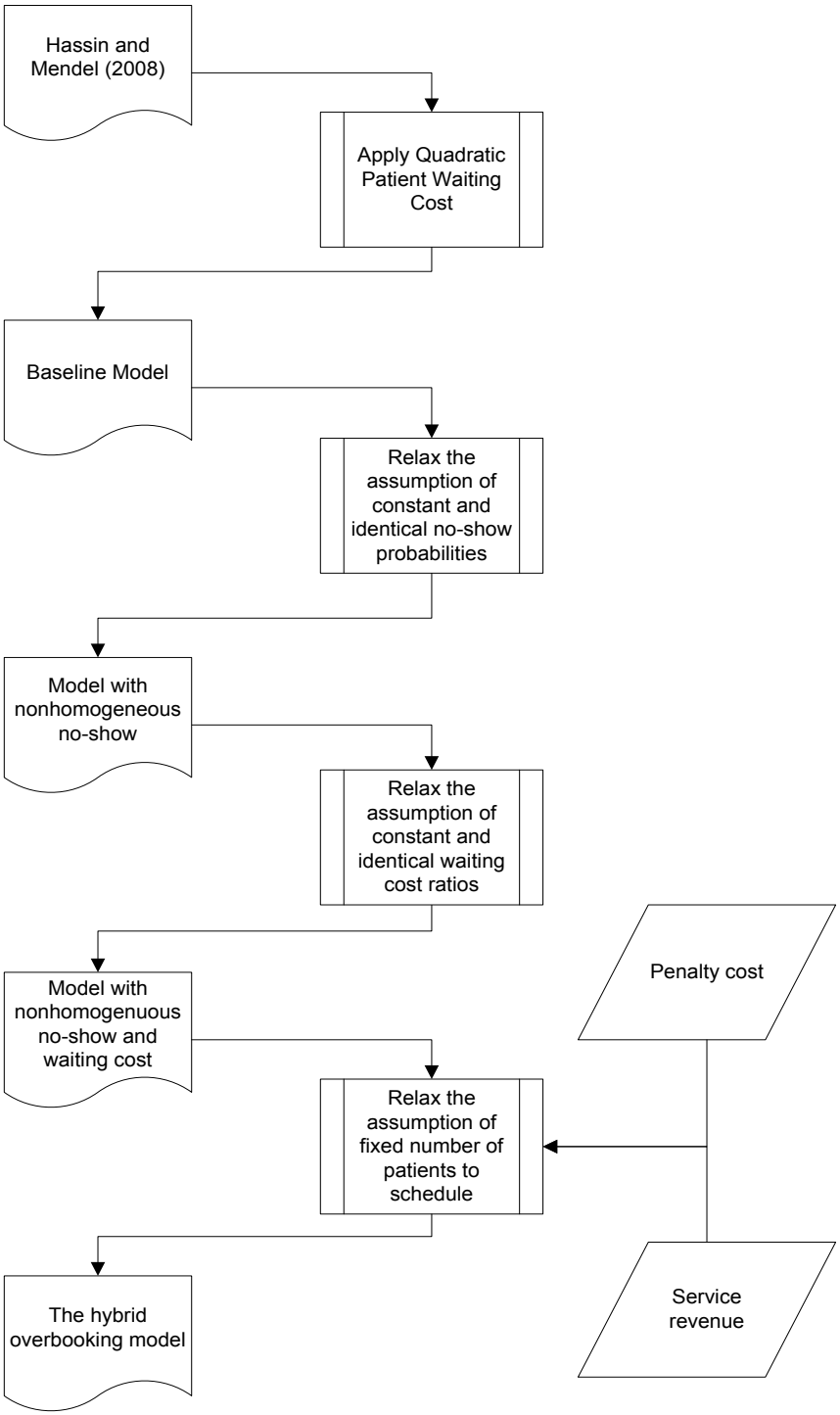
From a decision making point of view, appointment scheduling problems can be static or dynamic. In a static scheduling problem, all decisions are made before a clinic session starts; in a dynamic scheduling problem, the schedule of future arrivals can be revised during the clinic session. All the problems studied in this dissertation are categorized as static.

The objective of this dissertation is to find good static appointment scheduling systems (including decisions on inter-arrival times, sequence of patients based on no-show probabilities or patient types, and number of patients to schedule), which attempt to alleviate negative effectives of patient no-show, by optimizing pre-defined system performance measures (consisting of patient waiting, server idle time, service revenue, and patient deny penalty), with a more realistic representation of radiology environment. To achieve this goal, nonhomogeneous patients, overbooking, and a nonconventional patient waiting cost structures are considered.

2.1.Evolution of analytical models

The research part of this dissertation starts with a review on the following streams of literature: i) a set of problems of scheduling arrivals to queuing systems, ii) general appointment scheduling problems with patient no-shows, iii) scheduling multiple categories of patients, iv) scheduling rules, v) studies on patient no-shows, vi) time based cost measurement, and vii) new appointment policies. It extends from the first literature stream, with a series of assumptions relaxed. Figure 2.1 depicts the stepwise approach it takes to develop the analytical models.

Figure 2.1 Flow chart of model development



Starting with the analytical model by Hassin and Mendel [25], I first apply the concept of Taguchi's loss function to modeling patient waiting cost as a quadratic function of patient

waiting time. Then I relax the assumption of constant and identical no-show probabilities. I further relax the assumption of constant and identical waiting cost ratios among patients. Finally, I relax the assumption of fixed number of patients to schedule, by applying a nonconventional overbooking strategy which allows a certain level of server overtime.

2.2. Problem formulation

The subject radiology facility is modeled as a queuing system, with a single Markovian server. The scheduling horizon is one static non-evolving clinical session consisting of a fixed number of equal-length appointment slots. Patients are scheduled to arrive punctually at their appointment times, with certain chances of being no-shows.

2.2.1. Performance measurement

To be comparable with previous studies on scheduling arrivals to queuing systems, Chapter 4 uses a classic combination of two time based cost measurements: patient waiting cost and server idle cost, but in a different way, where patient waiting cost is not linearly related with patient waiting time. By employing an overbooking strategy, Chapter 5 takes service revenue and patient deny penalty cost into consideration.

- Time based cost measurement

In Chapter 4, the objective function is a total of quadratic patient waiting cost and linear server idle time cost. Patient waiting time is reported to be a direct indicator of patient satisfaction, while server idle time is considered as a waste of valuable medical resource. A combined cost

function of patient waiting and server idle time has been used by a majority of appointment scheduling papers, to name a few: [16], [43], [40], [57], [14], [15], [36], [25], and [50].

However, all studies mentioned above assume a linear relationship between patient waiting cost and patient waiting time, which might not be correct. To address this issue, I decided to model patient waiting cost as quadratic to waiting time. To the best of my knowledge, only Laganga and Lawrence [37] apply a quadratic patient waiting cost to the objective function. They did a thorough comparison between linear and quadratic cost functions. I distinguish my study by comparing the two functions within a context of nonhomogeneous no-show probabilities, and I find that quadratic waiting cost may change my decision on choosing the best scheduling policy.

- Service revenue

From each patient served, the MRI facility collects a certain amount of service revenue, in the form of an insurance fee charge. In most general hospitals, scanning fees charged on outpatients are much higher than those charged on inpatients. For the numerical study in Chapter 5, I keep the parameter setting for outpatient revenue at $1,500, and vary inpatient revenue from a low of $300 to a high of $1,000.

- Patient deny penalty cost

Unlike overbooking in hotel/airline industries, denying a scheduled patient usually don't incur direct penalty cost. Instead, it involves various indirect costs which are difficult to quantify. For inpatients, I approximate it with the opportunity cost of one extra day stay in the hospital, due to the fact of denying an inpatient causes longer unpaid stay; while for outpatients, I estimate it with a combination of cost of scheduling, potential staff overtime cost, and loss of goodwill.

2.2.2. Decision variable

The most commonly used decision variable for appointment scheduling problems is the schedule itself. A review of various representations of schedules is given in Chapter 3. As part of the hybrid overbooking model, the number of patients to schedule is also considered in Chapter 4.

- Schedule

Depending on the nature of a problem that is of interest, there are primarily two types of formulations for appointment schedules: i) inter-arrival times, and ii) patients to slots. The former is a vector of times between scheduled arrivals for fixed number of patients, on a continuous time basis. It can be expressed as $X = (x_1,\ldots, x_i,\ldots, x_{N-1})$ where x_i is scheduled inter-arrival time between ith patient and $(i + 1)$st patient, and N is the number of patients scheduled (see Figure 2.2); the latter is a vector of the numbers of patients assigned to each appointment slot, usually patients are required to arrive at the beginning of his/her slot, on a discrete time basis (see Figure 2.3). It can be expressed as $N = (n_1,\ldots, n_i,\ldots, n_K)$ where n_i is the number of patients scheduled to the ith slot, and K is the number slots for a clinic session. A common assumption shared by these two formulations is that appointment slots are of constant equal-length.

Figure 2.2 Schedule in form of inter-arrival times

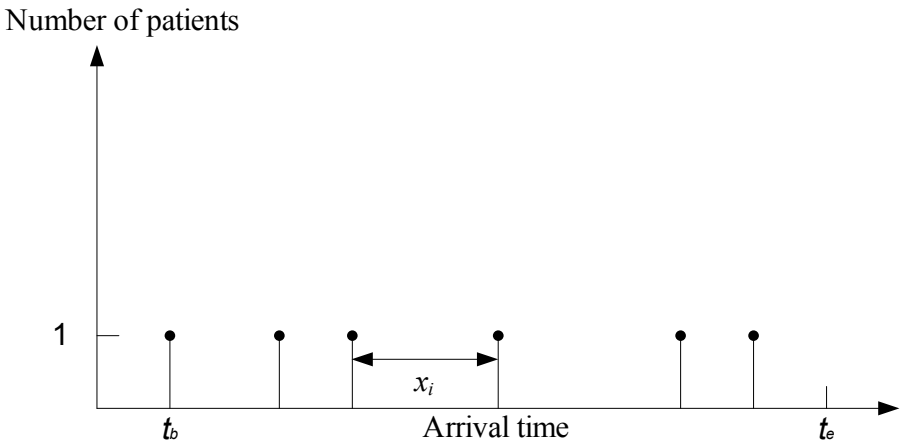
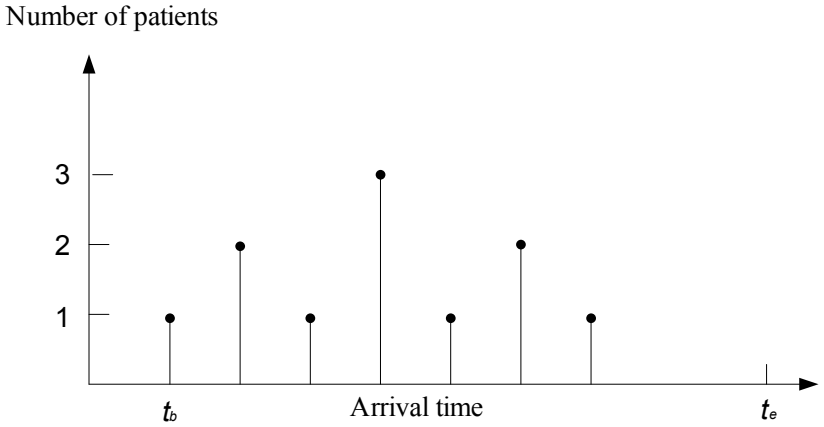


Figure 2.3 Schedule in form of patients to slots assignment



In general, the latter is frequently used in medical overbooking studies, for example, [37], [38] and [75]. However, in this dissertation, the former is chosen as part of decision variables in a hybrid overbooking model in Chapter 5.

- Number of patients to schedule

The hybrid overbooking model developed in Chapter 5 schedules more patients than the capacity of a clinic session. Thus, determining the optimal number of patients to schedule becomes one of the objectives.

- Patient sequence

One major contribution of this dissertation is a set of heuristics gained from sequencing patients based on their no-show probabilities or types. The numerical studies suggest scheduling patients with higher no-show probabilities in front of patients with lower no-show probabilities when there is no overbooking. In a case with overbooking, the numerical studies suggest scheduling inpatients first when no-show probability is relatively low, and outpatients first when no-show probability is relatively high.

2.2.3. Service time

Identical and constant service times are used by many studies for all patients or one category of patients, to reduce the complexity of the problems. It's a common assumption especially for dynamic scheduling problems such as [23], [22], and [54]. It could be justified by the fact that appointment slots are generally fixed and doctors try to finish their service on time. In general, the use of equal-length appointment slots is more suitable in the scenarios that medical services are delivered with less uncertainty. For example, it could be reasonable to assume equal-length service times for diagnostic imaging, while unrealistic to assume it for surgical services. Even for radiology, preparing patients takes significantly different times for different groups of patients. For example, female patients with more jewelry take a longer time than male patients. For static scheduling problems, due to the nature of its simplicity, variable service times are usually used.

Most studies assume independent and identically distributed exponential service times to make the models tractable. Throughout this dissertation, I assume service times are independent and identically distributed.

2.3.Data collection

Operations data on MRI service times and no-show is collected from the Radiology Information System (RIS) [12], which is owned and maintained by the University of Washington Medical Center (UWMC) and its affiliated facilities. The RIS data is stored in form MS SQL databases, which can be accessed via Crystal Reporting.

By specifying scheduled date between March 1st 2013 and March 31st 2013, modality as "MR", and department as "MR", 1,136 MRI appointment records are obtained. After further excluding the appointment records with reason code of "DUPLICATE EXAM SCHEDULED IN ERROR" and "WRONG DATE", I obtain a sample of 1,130 effective data points. Among the 1,130 MRI appointments, 93 fall into one of the following reason codes: "Patient Did Not Show", "Patient Cancelled", "Rescheduled", and "Patient Discharged", which are considered as no-shows in my numerical studies. Within the same sample, 791 appointments have both effective values of begin datetime and end datetime. Service time is calculated by subtracting begin datetime from end datetime. Average service time for these 791 appointments is approximately 46 minutes.

Revenue and cost related data is collected from public statistics or reviewed literature.

2.4.Solution method

All the models are coded in Matlab. Solutions are obtained by using the *fmincon* function in the Matlab optimization toolbox, which is designed to solve problems of minimizing a constrained nonlinear multivariate objective function. By choosing the "Active Set" optimization strategy, the function performs line search using Sequential Quadratic Programming (SQP) method.

SQP method is one of the most successful methods for computing numerical solutions of nonlinearly constrained optimization problems. At each iteration, it models a nonlinear programming problem by a Quadratic Programming (QP) sub-problem and solves it with an approximate solution. It then uses the approximate solution to generate a better approximate solution in the next iteration. This process creates a sequence of approximates which, when certain conditions are satisfied, converges to an optimal solution.

Given an appropriate choice of QP sub-problem, SQP can be viewed as an extension of Newton and quasi-Newton methods to constrained optimization problems. It shares two common attributes with Newton like methods: i) rapid convergence when iterates are close to the solution, and ii) possible erratic behavior when iterates are far from the solution [6]. The success of SQP methods depends heavily on the speed and accuracy of the algorithms chosen for solving QP sub-problems.

Chapter 3 **LITERATRATURE SURVEY**

Patient appointment scheduling has been studied intensively since 1950s. As the first to draw research attention to this research area, [2], [73], and [42] demonstrate the use of quantitative and simulation tools can significantly improve the performance of an appointment scheduling system. From then on, numerous studies have been conducted to approach this topic from various perspectives. Before the 2000s, most of the research efforts had been devoted to the problems of scheduling homogeneous outpatients to find an optimal balance between patient costs and server costs. Recent research trends lie in taking inpatients and emergency patients into consideration, exploring appointment policies that can handle multiple patient classes with a variety of arrival behaviors (no-show, late cancellation, and walk-in, etc.) Refer to [8], [14], and [48] for comprehensive reviews of outpatient appointment scheduling literature.

This dissertation is related with the following distinct streams of literature:

3.1.A set of problems of scheduling arrivals to queuing systems

I start this dissertation with a review of the studies on S(N)/M/1 problems which schedule arrival times of N homogeneous patients to a single Markovian server with independent and identical exponential service times. Pegden and Rosenshine [55] solve the cases of $N = 2$ and $N = 3$, they obtain explicit expressions of solutions, and prove them to be optimal by demonstrating that the objective functions are convex. For the problem with $N \geq 4$, they provide a recursive algorithm using gradient search to obtain a solution numerically. Stein and Cote [60] formulate the recursive algorithm with a transition matrix, and compute solutions with reduced-gradient search.

Hassin and Mendel [45][25] incorporate a constant no-show probability p into the S(N)/M/1 problem. Based on an assumption that the objective function is convex, they solving the problem by using line search. A key contribution of their work is a recursive representation of expected patient waiting time of the ith patient w_i. For all $i > j \geq 0$, w_i is calculated as the following

$$w_i = \theta \sum_{j=0}^{i-1} j \mathrm{P}\left(N_i = j\right), \tag{3.1}$$

where $\mathrm{P}(N_i = j)$ is the probability of j patients in the system right before the ith scheduled arrival.

When $j = 0$

$$\begin{aligned}
\mathrm{P}\left(N_i = 0\right) &= \left(1-p\right)\sum_{k=0}^{i-1}\left[\mathrm{P}\left(N_{i-1} = k-1\right)\sum_{l=k}^{\infty}\mathrm{P}\left(N\left(x_{i-1}\right) = l\right)\right] \\
&\quad + p\sum_{k=0}^{i-2}\left[\mathrm{P}\left(N_{i-1} = k\right)\sum_{l=k}^{\infty}\mathrm{P}\left(N\left(x_{i-1}\right) = l\right)\right] \\
&= \left(1-p\right)\sum_{k=0}^{i-1}\left[\mathrm{P}\left(N_{i-1} = k-1\right)\left(1-\sum_{l=0}^{k-1}e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{l}}{\theta^k l!}\right)\right] \\
&\quad + p\sum_{k=0}^{i-2}\left[\mathrm{P}\left(N_{i-1} = k\right)\left(1-\sum_{l=0}^{k-1}e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{l}}{\theta^k l!}\right)\right] \\
&= \sum_{k=0}^{i-2}\left[\mathrm{P}\left(N_{i-1} = k\right)\left(1-\sum_{l=0}^{k-1}e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{l}}{\theta^k l!} - \left(1-p\right)e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{k}}{\theta^k k!}\right)\right]
\end{aligned}, \tag{3.2}$$

When $j > 0$

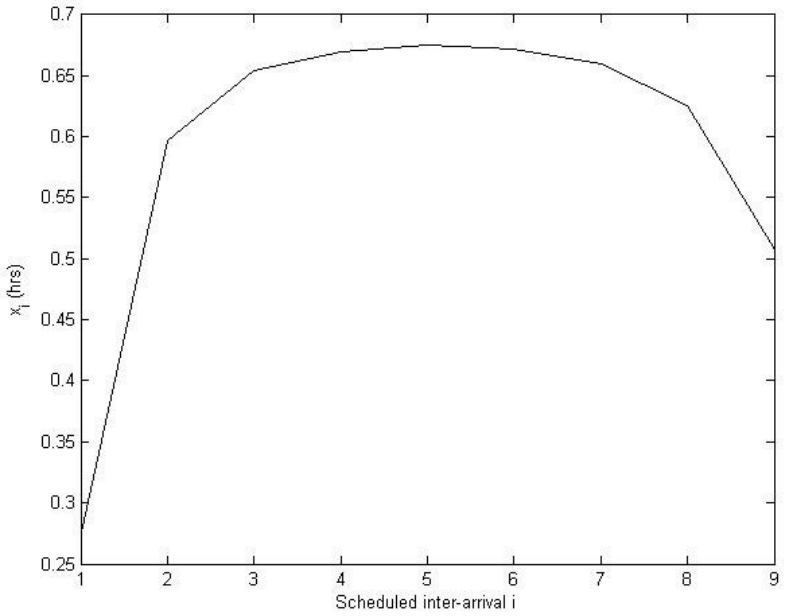
$$
\begin{aligned}
\mathrm{P}\left(N_i = j\right) &= \left(1-p\right)\sum_{k=0}^{i-j-1}\left[\mathrm{P}\left(N_{i-1} = j+k-1\right)\mathrm{P}\left(N\left(x_{i-1}\right) = k\right)\right] \\
&\quad + p\sum_{k=0}^{i-j-2}\left[\mathrm{P}\left(N_{i-1} = j+k\right)\mathrm{P}\left(N\left(x_{i-1}\right) = k\right)\right] \\
&= \left(1-p\right)\sum_{k=0}^{i-j-1}\left[\mathrm{P}\left(N_{i-1} = j+k-1\right)e^{-\frac{x_{i-1}}{\theta}}\frac{{x_{i-1}}^{k}}{\theta^k k!}\right] \\
&\quad + p\sum_{k=0}^{i-j-2}\left[\mathrm{P}\left(N_{i-1} = j+k\right)e^{-\frac{x_{i-1}}{\theta}}\frac{{x_{i-1}}^{k}}{\theta^k k!}\right] \\
&= \sum_{k=1}^{i-j-1}\left[\mathrm{P}\left(N_{i-1} = j+k-1\right)e^{-\frac{x_{i-1}}{\theta}}\frac{{x_{i-1}}^{k-1}}{\theta^{k-1}\left(k-1\right)!}\left(\left(1-p\right)\frac{x_{i-1}}{\theta k}+p\right)\right] \\
&\quad + \left(1-p\right)\mathrm{P}\left(N_{i-1} = j-1\right)e^{-\frac{x_{i-1}}{\theta}}
\end{aligned}
\quad , \qquad (3.3)
$$

where P($N(x_i) = l$) denotes the probability that l patients are served during scheduled inter-arrival time between ith and (i+1)th patients x_i.

A common observation from this set of problems is the well-known "dome shaped" optimal schedule in form of a vector of inter-arrival times $X = (x_1,\ldots, x_i,\ldots, x_{N\text{-}1})$. Compared to arrival times, it is chosen to give a better illustration of the Poisson distribution that $N(x_i)$ follows.

Refer to Figure 3.1 for an example of dome shaped schedule, where i of X axis denotes inter-arrival number (i.e. i = 1 denotes inter-arrival between 1st patient and 2nd patient), x of Y axis denotes the corresponding scheduled inter-arrival time in unit of hour. A point (i, x) in the graph represents x_i, the scheduled inter-arrival time between ith patient and (i + 1)st patient. In a dome shaped schedule, inter-arrival time increases among the first several patients, keeps relatively constant at a certain level thereafter, and then drops significantly between last two patients.

Figure 3.1 An example of the "dome shaped" optimal schedule



Intuitive explanations for this behavior have been given by both [60] and [25]. The scheduling of first several patients to arrive closer is to avoid the server being idle after finishing the first patient. The relatively fixed inter-arrival times among patients in the middle is due to the fact that the person scheduling customers at time 0 will anticipate a probabilistic steady-state at a time distant in the future. Scheduling last several patients closer is to avoid server being idle while only a few customers remain to arrive.

The dome shaped optimal schedule is first observed by Stein and Cote [60] in a numerical study on the cases of $N = 50$ and $N = 10$ without no-shows. [45] and [25] demonstrate that the adding of constant and identical patient no-show probabilities does not change the dome shape of optimal schedules.

Dependency of the dome shape for optimal schedules is summarized as patients being homogenous with non-deterministic service times. Specifically, it needs to satisfy the following two requirements: i) service times are non-deterministic, but independent and identically distributed, and ii) no-show probabilities and waiting cost ratios are constant and identical.

3.2.Scheduling patients with no-shows

The phenomenon of no-show has an inevitable impact on efficiency and accessibility of health care systems. Although no-show itself has been an active research area with productive results, throughout nearly 60 years research history of appointment scheduling, limited attention has been paid to scheduling problems with patient no-shows until the past decade.

3.2.1. General appointment scheduling problems with patient no-shows

[46] and [47] are considered as pioneering works in this research area. In these two studies, no-show is modeled together with patient lateness, a no-show happens when the lateness probability is greater than a threshold value. Recent studies considering no-show include [23], [36], [25] and [22]. The first three assume constant no-show probability, which might be oversimplified. Koole and Kaandorp [36] solve a problem of scheduling patients with constant and identical no-show probabilities to a server with independent and identically distributed exponential service times; a local search algorithm is developed to compute the schedule with lowest objective value. They report the well-known dome shape schedules, which is also observed by Hassin and Mendel [25] in a similar problem. The major difference between these two studies lies in decision variables, where the former study uses patients to slots while the latter uses inter-arrival times. I differentiae my dissertation from these two studies with quadratic patient waiting cost structure,

nonhomogeneous no-show probabilities and nonhomogeneous patient waiting cost ratios. In my model, if patients are not sequenced in ascending or descending order of their no-show probabilities, schedules are not dome shaped anymore. Green and Savin [22] model patient no-show probability as a function of indirect waiting/appointment delay in a dynamic scheduling problem, which supports one of my primary assumptions that no-show probability a patient could be predetermined at the time of booking. Refer to [9] and [27] for schedule evaluation studies which consider no-shows.

Most closely related to my work is a study by Zeng et al. [75], which considers nonhomogeneous patients with different no-show probabilities in an overbooking model. In this study, they report a solution where patients with lower no-show probabilities are scheduled in front of patients with higher no-show probability, which is contradictory to my finding that patients with higher no-show probabilities are preferred to be scheduled in front of patients with lower no-show probabilities. I will discuss the discrepancies in further detail in Chapter 4.

3.2.2. Overbooking applications to alleviate patient no-show

In Chapter 5, in order to take the unknown arrivals of emergency patients into consideration within a static scheduling environment, where all decisions are made before a clinic session starts, I choose to use a hybrid overbooking strategy, which aims to determine optimal number of patients to schedule and inter-arrival times. Overbooking as an important strategy for overcoming customer no-shows, has been a common practice in airline and hotel industries for decades. It recently draws research attention to outpatient appointment scheduling.

Besides the fact of suffering from no-shows, clinic scheduling problems are quite different from the airline/hotel booking problems. Muthuraman and Lawley [50] summarize the major differences between these two types of problems from three perspectives: objective function, decision variable, and system dynamics. They formulate the objective function with patient waiting time, staff overtime and patient revenue. Patients to slot vector $N = (n_1, n_2, \ldots, n_K)$ is chosen as decision variable, a different increasing pattern of expected patient waiting for various numbers of patients is observed. Zeng et al. [75] develop a sequential scheduling procedure to solve a similar problem, with essentially the same objective function, and same decision variable. In this study, patient waiting is modeled as number of patients who overflow from one slot to the next, patient deny penalty cost, as a major concern in overbooking problems, is not taken into consideration. Laganga and Lawrence [37], [38] evaluate performance of an overbooking strategy within a wide range no-show probabilities, schedule sizes, and overbooking ratios. Based on rich numerical studies results, they demonstrate the advantages of applying overbooking to clinic scheduling. In these studies, however, patients are assumed to be homogeneous with constant and identical no-show probabilities and waiting cost ratios, which limit their applications to practices. Another representative study in clinical overbooking is [32].

3.3.Scheduling multiple categories of patients

Traditionally, appointment scheduling papers only take outpatient into consideration. In practice it is possible that inpatients are also scheduled to come, therefore, a recent research trend of appointment scheduling lies in scheduling multi-category patients. A typical radiology department in a general hospital faces demand from three patient groups: outpatients, who are not hospitalized, visit a hospital for diagnosis or treatment; inpatients, who are admitted to stay

overnight in a hospital for treatment or observation; and emergency patients, the unexpected walk-in patients who need urgent treatment. Demands from all three groups arrive randomly, but are treated in different ways. Outpatients are scheduled in advance. Emergency patients are usually served with next available slot, by postponing scheduled appointments. Rules of accepting and serving inpatients requests vary from hospital to hospital. Even within the same hospital, it varies from department to department. One common practice is to reserve a certain number of appointment slots or time periods within a clinic session for inpatients. In some hospitals, inpatients requests are held till an empty slot becomes available. Among the three groups, emergencies are usually served with highest priority, while the priorities of the rest two groups vary from case to case. In some cases, outpatients are assigned with different levels of priorities, based on which the appointment delay is determined. In British Columbia, Canada, there exist three outpatient priority classes with allowable wait times of 7, 14, and 28 days, respectively [54]. The three patient groups also have different associated costs. From the point of view of management, outpatients are often considered as revenue source while serving inpatients are regarded as cost.

Limited research attention has been paid to the problem of scheduling multiple categories of patients for constrained medical resources. Gerchak et al. [19] explore this problem in a situation of allocating operating rooms between elective and emergency patients, with an emphasis on finding an optimal policy for scheduling elective patients. Refer to [61] and [33] for similar studies on booking operating time slots for different patient classes. A subset of this literature stream deals with this problem in radiology environment.

3.3.1. Applications to radiology practices

Almost all studies on scheduling multiple categories of patients model the problem as dynamic scheduling. I distinguish my study by dealing booking multiple category patients within a static booking environment. Among these studies, Markov Decision Process (MDP) is the most frequently used modeling technique. Reviewed literature includes modalities of Computer Tomography (CT), Magnetic Resonance Imaging (MRI) and chest radiography. Patrick et al. [54] consider a case of CT department in a general hospital which serves inpatients with high priority and outpatients with lower priority. They use MDP to model it as a dynamic scheduling problem which aims to minimize overall patient waiting cost. Due to the extremely large state space, approximate dynamic programming is employed to solve an equivalent linear program of this problem. No-show is not considered in this study. Kolisch and Sickinger [35] also use MDP to model dynamic resource allocation between two CT scanners to the three categories of patients, with an objective function to maximize the total rewards consist of revenues, waiting costs and penalty costs. Walk-in is considered in this study which differentiates this paper from most existing outpatient appointment scheduling papers. The demand from walk-in emergency patients and inpatients are modeled as random arrivals with fixed show up probabilities. Green et al. [23] formulate the problem of managing demand from inpatients, outpatients and emergency patients for MRI services as a finite-horizon dynamic program. They develop a linear heuristic as an alternative to the optimal policy, and use numerical studies based on empirical data to show the robustness of this heuristic. Refer to [53] for more work on allocating diagnostic resources to multi-category patients. Table 3.1 summarizes radiology appointment scheduling studies with consideration of multi-category patients.

Table 3.1 Radiology appointment scheduling studies with consideration of multi-category patients

Year	Author	Imaging Service	Patient Type	No-show
1973	Walter [68]	Chest Radiography	OP, IP	Without
2006	Green et al. [23]	MRI	OP, IP and EMP	Constant p
2008	Kolisch and Sickinger [35]	CT	OP, IP and EMP	Constant p
2008	Patrick et al. [54]	CT	Multi-priority OP, IP and EMP	Without
2010	Gocgun [20]	CT	OP, IP, and multi-priority EMP	Without

As pointed out by Patrick et al. [54], the major challenge faced in this type of problems "is that the low-priority demand must be booked before knowing the high-priority demand. Therefore, a significant portion of the total capacity must be reserved for this unknown high-priority demand leading inevitably to unused capacity on those days when the high-priority demand is lower than expected." To alleviate the waste caused by higher than needed reserved capacity, in Chapter 5, instead of reserving capacity for high priority emergency patients, I overbook one emergency to each appointment slot.

3.4. Time based cost measurement

In general, there are two broad categories of time based measures: patient time and server time.

The most commonly used patient time is patient waiting time, which has been included in a majority of schedule evaluation type of studies. To name a few, [42], [31], [59], [16], [55], [27], [43], [66], [40], [57], [14], [36], [37], [25], [50], and [75]. Some studies also use patient flow time, which is defined as the total time from the moment when a patient enters the queue to the moment when the patient is released by server. A breakdown of total expected patient flow time is illustrated by the following formula:

$$E(t_F) = E(t_W) + \sum_{i=1}^{N} E(t_{si}), \tag{3.4}$$

where t_F denotes total patient flow time, t_W denotes total patient waiting time, t_{si} denotes service time of the ith patient. From optimization point of view, it's essentially equivalent to patient waiting time, as total expected service time $\sum_{i=1}^{N} E(t_{si})$ is fixed. Refer to [70] for a use case of server completion time as part of objective function.

Server time includes but not limited to three types: server idle time, server completion time, server overtime.

Server idleness happens when there is no patient in the system before a clinic session ends. It is regarded as waste of valuable medical resource. To summarize, it has been used by the following studies: [31], [16], [27], [41], [43], [40], [57], [14], [15], [36], and [25].

Server overtime is defined as the timespan between scheduled finished time and actual completion time of one service session. Studies employed server overtime includes: [16], [55], [41], [69], [60], [70], [43], [71], [66], [14], [15], [36], [37], [75].

Server completion time is defined as the time that a doctor spends from the beginning of the schedule until the last patient scheduled for the period has been served. It has been used by [69] and [70].

3.5.Scheduling rules

In this section, to illustrate scheduling rules, I adopt the graphical representation used in [16]. t_b denotes beginning time; t_e denotes end time; n_1 denotes the number of patients scheduled to 1st

slot, n_i denotes the number of patients scheduled to ith slot ($i \geq 1$); T denotes clinical session length; N is number of patients scheduled to one clinical session.

Starting from [2], a number of studies attempted to find scheduling rules, under certain circumstances, that achieve a reasonable balance between server idle cost and patient waiting cost.

In early times, doctor's time was considered to be much more valuable than a patient's time. Therefore, most hospitals/clinics used scheduling rule $n_1 = N$ / $n_i = 0$ / no x_i (see Figure 3.2), which minimizes server idle time while maximizing patient waiting. On the other hand, scheduling rule $n_1 = 1$ / $n_i = 1$ / constant x_i (see Figure 3.3) pays more attention to patient waiting, by scheduling one patient to arrive at the beginning of each equal-length appointment slot, significantly reduces patient waiting, but degrades doctors' utilization.

Figure 3.2 Rule $n_1 = N$ / $n_i = 0$ / no x_i

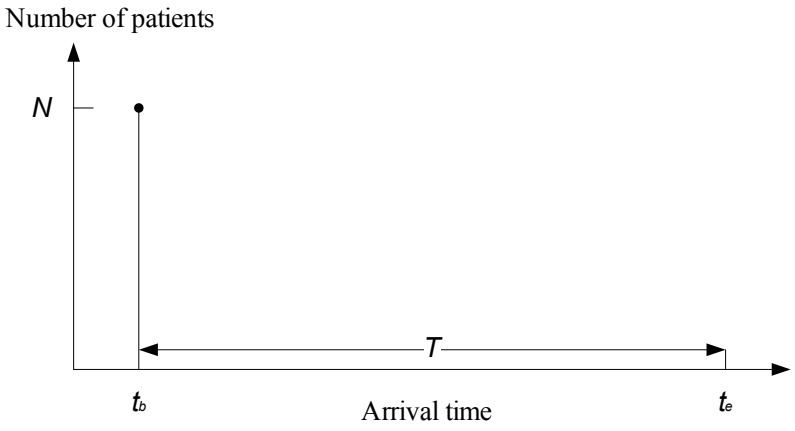
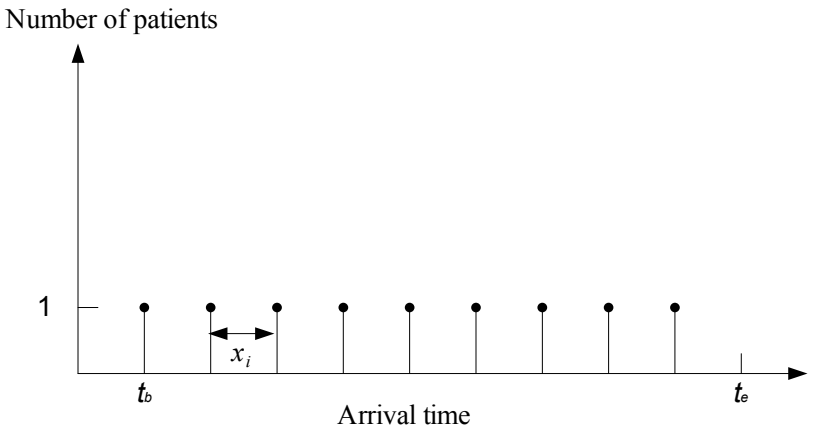


Figure 3.3 Rule $n_1 = 1$ / $n_i = 1$ / constant x_i



Bailey [2] in his study suggests the rule of $n_1 = 2$ / $n_i = 1$ / constant x_i, which attempts to schedule two patients to the first slot, and one patient to each slot thereafter. It significantly reduces the risk of server being idle at the beginning of a clinic session. Later known as "Bailey's rule" (See Figure 3.8), this rule has been evaluated by numerous studies in this field. "Bailey's rule" falls into the general category of $n_1 > 1$ / $n_i = 1$ / constant x_i (see Figure 3.4), it's a special case where $n_1 = 2$. Recent research trend lies in the rules of $n_1 \geq 1$ / $n_i = 1$ / variable x_i (see Figure 3.5), which allows variable spaced inter-arrivals, it aims to determine the best combination of x_i that achieves optimum or sub optimum objective function values. Rules of $n_1 = n_i > 1$ / constant x_i (see Figure 3.6) and of $n_1 > n_i > 1$ / constant x_i (see Figure 3.7) are generally used in clinic overbooking studies, in which each appointment is booked with more than one patients.

Figure 3.4 Rule $n_1 > 1$ / $n_i = 1$ / constant x_i

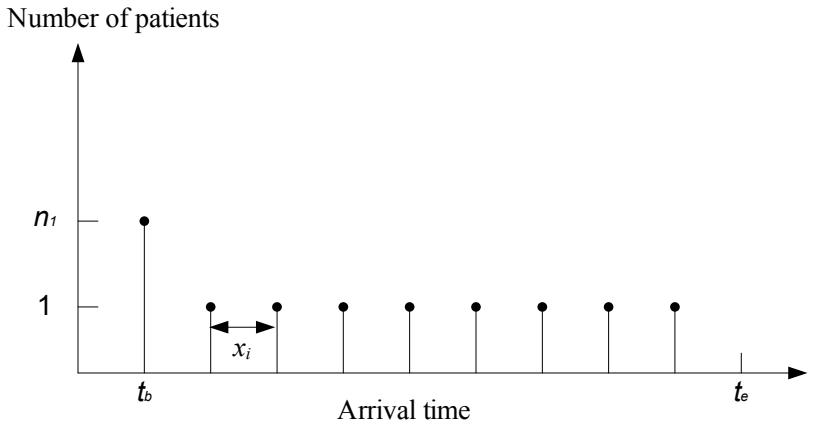


Figure 3.5 Rule $n_1 \geq 1$ / $n_i = 1$ / variable x_i

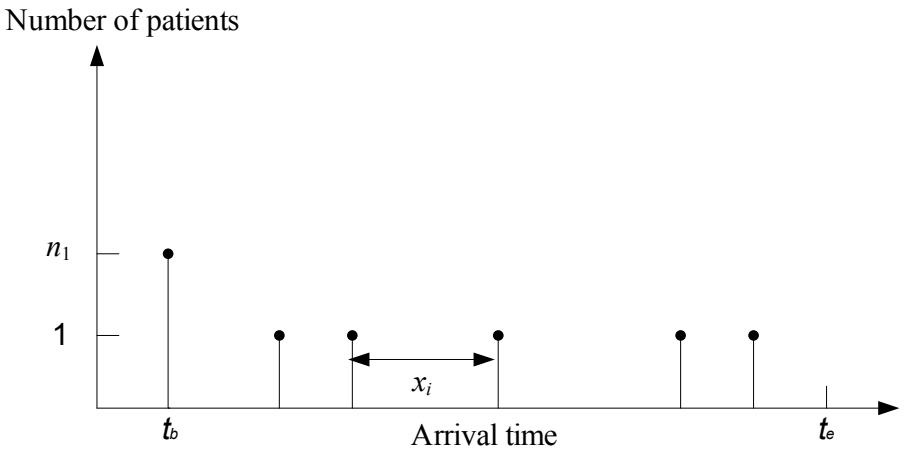


Figure 3.6 Rule $n_1 = n_i > 1$ / constant x_i

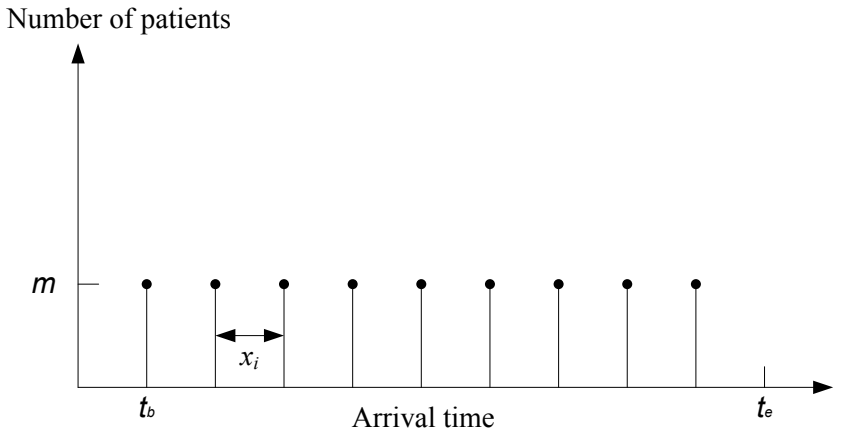
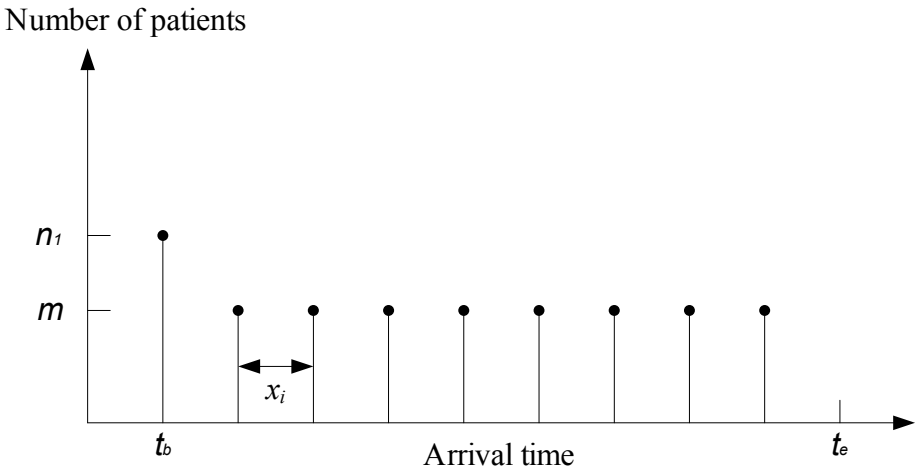


Figure 3.7 Rule $n_1 > n_i > 1$ / constant x_i



3.5.1. Rules in radiology practice

There are three popular scheduling rules in radiology practice. They use simulation to evaluate these three rules for scheduling outpatients:

- Bailey's rule: also known as front-loading, it schedules 2 to the first slot, and one patient to each of the remaining slots (see Figure 3.8);

- Block: a variation of the primitive rule $n_1 = N$ / $n_i = 0$ / no x_i. It can be described as $n_1 = \frac{N}{2} / n_2 = \frac{N}{2} / x_1 = \frac{T}{2}$. Patients are scheduled to only two slots, one at the beginning of the day and one at the beginning of the afternoon. In this case, instead of an exact times, patients who make appointments are only scheduled to come during the morning or during the afternoon;
- Threshold: all patients are scheduled to the slots before a specified threshold time, the remaining slots are left open.

The first two rules are used by two university medical centers in Germany, as study objectives in [35]. The third one is discussed in detail by Green et al. [23], as a popular practice of scheduling patients to MRI services which are under the pressure of high equipment cost.

Figure 3.8 Bailey's rule

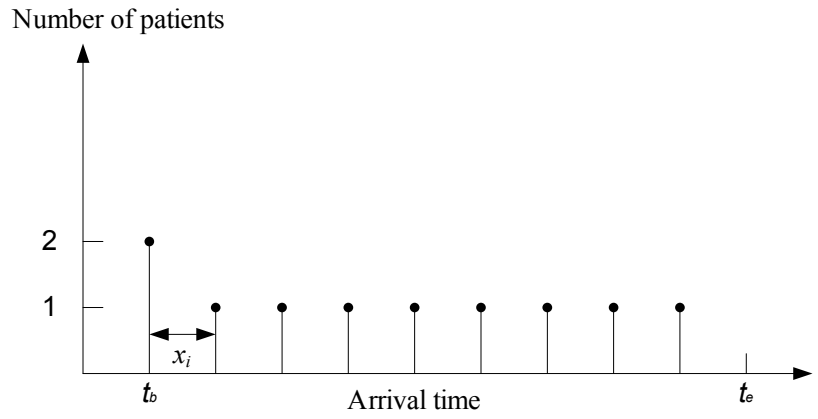


Figure 3.9 Block rule

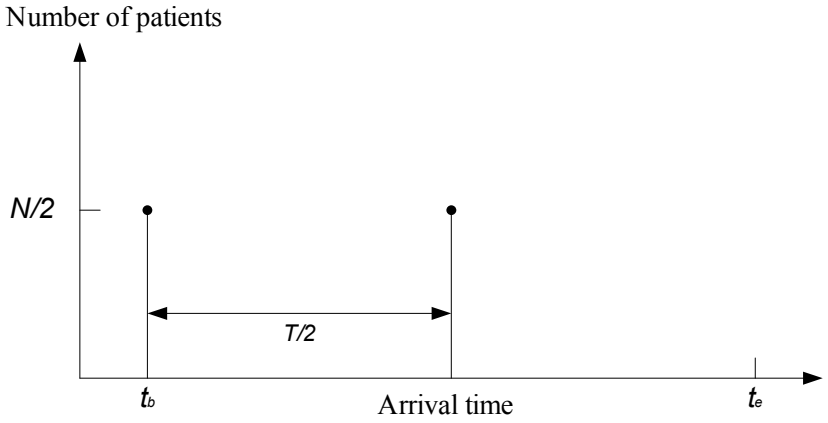


Figure 3.10 Threshold rule

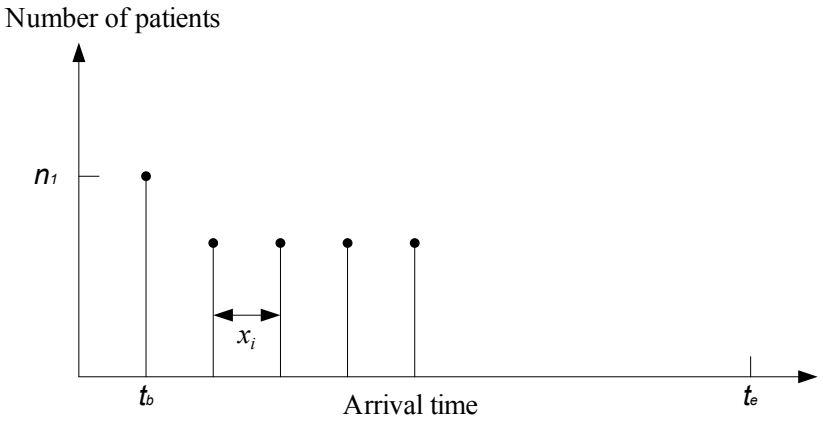


Table 3.2 summarizes the three rules mentioned above.

Table 3.2 Summary of scheduling rules in radiology practice

Rules	First slot (n_1)	Remaining slot (n_i)	Inter-arrival time (x_i)
Bailey's rule	2	1	Constant
Block	N/2	N/2	constant (T/2)
Threshold	≥1	≥1 or 0	constant

3.6.Studies on patient no-show

In this section, I first summarize reviewed surveys on no-show rates of different practices, and then give a brief description of a prevailing method to model patient no-show probability with appointment delay.

3.6.1. Empirical studies on no-show rates

U.S. based surveys on patient no-shows indicate that a majority of clinic services experience no-show rates within a range between 20% and 30%. One of the most comprehensive national

surveys on no-show rate is conducted by Hixon et al. [26]. In this study, 60% of 468 surveyed U.S. family practice residency programs responded, and that over a third of them report experiencing a no-show rates higher than 20%. Moore et al. [49] report a calculated no-show rate around 31% out of the 4,055 patient appointments scheduled to a period of 20 days. Xakellis and Bennett [74] report a 25% of no-show rate in a family practice clinic. Similar numbers are obtained by Ulmer and Troxler [64], whose study considers seasonal variations of no-show rates, which range from 22% to 26.5%. Table 3.3 gives a detailed summary to the above studies.

Table 3.3 No-show rates of family and primary care clinics

Study	Sample size	Type of clinic	Length of study period	Including cancellation?	No-show rate
Moore et al. [49]	4,055	family practice clinic	20 days	Yes	31%
Xakellis and Bennett [74]	555	Family medicine teaching clinic	1 week	No	25%
Ulmer and Troxler [64]	3618 and 3101	Primary care practice	3 months in 2001 and 3 months in 2002	No	22% and 26.5%

3.6.2. Modeling no-show probability with appointment delay

No-show probability is used in many studies to quantify the arrival behavior of patient no-show. Most studies assume constant no-show probabilities, for the sake of simplicity. However, in practice, no-show probability is found to be a variable that depends on other factors such as patient's age, sex, marital status, income level, and so on. A few studies investigate this functional representation of no-show rate. Galucci et al. [18] conduct a study at a public mental health clinic at the Johns Hopkins Bayview Medical Center in Baltimore, on the dependence of

cancellation and no-show rate on appointment delay in number of days, they identified a strong relationship between no show and appointment delay described as the following:

i. the no-show rate starts at a minimum value (about 12%) for same day appointments;
ii. the no-show rate increases (by about 12% per extra day of waiting) monotonically with length of appointment delay;
iii. the no-show rate stabilizes when it reaches a maximum value (about 42%).

Green and Savin [22] extended this study by finding similar features based on data from the Columbia MRI facility and building a no-show function depending on backlog length. Their minimum and maximum no-show rates are 4% and 37%, respectively. By comparing the no-show rates in these two studies, they claim that patients of a mental health care center are less reliable than the patients of an MRI diagnostic facility. They develop an exponential no-show function as the following:

$$\gamma(k) = \gamma_{\max} - (\gamma_{\max} - \gamma_{\min})e^{-k/C}, \tag{3.5}$$

where k denotes the appointment delay in days; $\gamma_{\max}$ and $\gamma_{\min}$ denote the maximum and minimum observed no-show rates; C is a no-show appointment delay sensitivity parameter. Best fit values of the no-show function parameters $\hat{C}$, $\hat{\gamma_{\max}}$ and $\hat{\gamma_{\min}}$ are obtained by minimizing the sum of the squared deviations between the predicted and actual no-show rates. The best fit no-show functions are applied to both cases and compared with original data, respectively. Table 3.4 (adapted from [22]) compares the estimated parameters of no-show function for these two studies, it reinforces the claim that patients of mental health is more sensitive to appointment delay than patients of diagnostic imaging.

Table 3.4 Estimated parameters for no-show probability functions

Data source	$\hat{C}$	$\hat{\gamma}_{\min}$	$\hat{\gamma}_{\max}$
Galucci et al. [18]	9	0.15	0.51
Green and Savin [22]	50	0.01	0.31

3.7.New appointment policies

The positive correlation between patient no-show rate and appointment delay provides us a hint on how to reduce no show: shortening the appointment delay. For this sake, a new appointment policy named Advanced Access (AA) or Open Access (OA) emerges and becomes popular in clinical practices.

Developed by Murray and Tantau [51], the AA/OA system sticks to one core rule "Do today's work today", which enables patients to see their own personal physicians on same day they make their appointments, by eliminating the functional relationship between appointment delay and urgency/request service type (routine or preventive). Under AA/OA, patients are also allowed to schedule preferred future appointment.

Many success stories of implementing AA/OA systems have been reported, with significant improvements on various measures. In their own practice, Murray and Tantau [51] report a reduction of patient wait for appointment from 55 days to 1 day, an increase of patients' chance to see their own physicians from 47 percent to 80 percent, and significantly improved patient satisfaction. As a result of reduced appointment delay, no-show rates decrease to minimum levels.

Successful implementations of AA/OA systems depend on a key factor, that is, to keep demand and supply in balance. In practice, it is difficult to determine how much is "in balance". As daily patient demand fluctuates, simply ensuring supply is more than average demand is not enough. Without sufficiently high capacity relative to demand, severe overloads may happen frequently [44]. Green and Savin [22] try to answer this question by developing a method with two queuing models and one simulation model to determine the largest panel sizes that work sustainably under AA/OA system. Their results show that the panel sizes are much smaller than theoretical values required for a queuing system to stabilize, which supports the common feeling that the long run supply rate is greater than the long run demand rate is not sufficient for AA/OA to work.

For many clinics, however, it's not realistic to maintain a capacity that is much higher than the average daily demand. As a result, the effect of implementation of AA/OA system is controversial.

Chapter 4 **STATIC SCHEDULING FOR PATIENTS WITH NONHOMOGENEOUS NO-SHOW PROBABILITIES AND NONHOMOGENEOUS WAITING COST RATIOS**

Appointment scheduling systems have been studied for nearly 60 years. However, within this research field, limited attention has been paid to patient no-show until the past decade. In this Chapter, I focus my study on a set of problems of scheduling patients with no-shows to the queuing systems with exponentially distributed service times. To represent the nonlinear nature of the relationship between waiting cost and patient waiting time, I formulate the objective function as a total of quadratic patient waiting cost and linear server idle cost.

I first solve a problem of the baseline model with homogeneous patients, and compare it with Hassin and Mendel's model, which assumes linear patient waiting cost. Then, I solve a problem with patient no-show probability varied among patients. By comparing it to a model with linear waiting cost, I find quadratic waiting cost may change my decision of sequencing patients when no-show probability is nonhomogeneous. Last, I solve another problem with both patient no-show probability and patient waiting cost ratio varied among patients. I compare the performance of three no-show probability based patient sequencing heuristics: lower no-show first, higher no-show first, and higher no-show in the middle, with the purpose of providing simplified heuristics to medical scheduling practices.

4.1. The baseline model with quadratic waiting cost

The objective is to minimize a total cost of expected patient waiting cost and expected server idle cost by determining an optimal schedule $X = (x_1, \ldots, x_i, \ldots, x_{N-1})$ of N homogeneous patients with constant no-show probability p. Denote the scheduled inter-arrival time between ith patient and $(i + 1)$st patient with x_i. Patients are served by a single server on a First Come First Served (FCFS) basis, with independent and identically distributed exponential service times. In later sections of this chapter, I study models with nonhomogeneous no-show probabilities and waiting cost ratios. Late arrival is not considered in this model. Given that a patient shows up, he/she must be punctual.

4.1.1. Notation

p	no-show probability
c_W	patient waiting cost ratio
c_I	server idle cost ratio
α	relative patient waiting cost ratio = $c_W/(c_W + c_I)$
θ	expected service time
w_i	expected waiting time of ith patient given showing up
N_i	number of patients in the system right before ith scheduled arrival
$N(x_i)$	number of patients served during the inter-arrival time x_i
β	patient waiting cost coefficient

4.1.2. Model description

The general objective function with linear costs assumption is formulated as the following:

$$\min \quad c_W (1-p)\sum_{i=1}^{N} w_i + c_I \left(\sum_{i=1}^{N-1} x_i + w_N + (1-p)\theta - (1-p)N\theta \right), \tag{4.1}$$

where $c_W (1-p)\sum_{i=1}^{N} w_i$ represents total expected patient waiting cost,

$c_I \left(\sum_{i=1}^{N-1} x_i + w_N + (1-p)\theta - (1-p)N\theta \right)$ represents total expected server idle cost. The term $\sum_{i=1}^{N-1} x_i$ denotes the scheduled arrival time of the last patient, combined with w_N, $\sum_{i=1}^{N-1} x_i + w_N$ defines the expected time when the server can start to serve last patient, given he/she shows up. By adding $(1-p_N)\theta$, I obtain $\sum_{i=1}^{N-1} x_i + w_N + (1-p_N)\theta$ as expected service completion time.

The objective function assumes that all the costs are linearly related with time, which holds well for server idle cost, as the unit time costs of equipment and personnel are relatively stable within a short time frame. However, it might not be fair to assume that patient waiting cost is linearly related with patient waiting time. Besides the fact that waiting cost of 40 minutes waiting time for one patient doesn't equal the case where 20 patients each waits for 2 minutes [34], for a long time, it has been suspected that there might be a threshold value of patient waiting time, above which the tolerance of patients decreases steeply. Furthermore, it also involves issues of goodwill, service, and "costs to the society", which place a value on patients' waiting time [8]. As a patient continues to wait, unit time waiting cost increases.

With the purpose to better address the issues mentioned above, I employ the concept of quadratic loss function, developed by Taguchi, where I regard patient waiting cost as quality loss of health

care services provided to patients. The target value of expected waiting time is set at 0, hence, it's a "smaller the better" characteristic.

I define the quadratic waiting cost representation as the following:

$$c_W (1-p)\sum_{i=1}^{N}\left(w_i^2\right). \tag{4.2}$$

There are two ways to determine the waiting cost ratio c_W: relative cost ratio value and absolute monetary value. For the former, the value of relative cost ratio $\alpha = c_W/(c_W + c_I)$ is usually determined by healthcare decision makers, from their perception or standard cost accounting. For the latter, I will discuss in detail in the numerical study of Section 4.3. In this section, since patient waiting cost ratio c_W is constant, I decide to use relative cost ratio α.

With quadratic patient waiting cost, my objective function becomes a nonlinear combination of expected patient waiting time, total expected server idle time (see the equation below), which differentiates the baseline model from the model developed by Hassin and Mendel [25].

$$\min \quad Z = c_W (1-p)\sum_{i=1}^{N}\left(w_i^2\right) + c_I\left(\sum_{i=1}^{N-1} x_i + w_N + (1-p)\theta - (1-p)N\theta\right). \tag{4.3}$$

The constant terms $(1-p)\theta$ and $(1-p)N\theta$ do not impact the optimal solution. By removing them, the objective function can be simplified as the following

$$\min \quad Z = c_W (1-p)\sum_{i=1}^{N}\left(w_i^2\right) + c_I\left(\sum_{i=1}^{N-1} x_i + w_N\right). \tag{4.4}$$

Apply relative patient waiting cost ratio to the equation above, by dividing it with ($c_W + c_I$), I obtain

$$\min \quad Z = \alpha(1-p)\sum_{i=1}^{N}\left(w_i^2\right) + (1-\alpha)\left(\sum_{i=1}^{N-1} x_i + w_N\right). \tag{4.5}$$

Besides the nonnegative requirement for x_i, the constraint to this problem is the representation of w_i. For all $i > j \geq 0$, w_i is calculated as the following

$$w_i = \theta\sum_{j=0}^{i-1} j\mathrm{P}\left(N_i = j\right). \tag{4.6}$$

Refer to Hassin and Mendel [25] for development of the recursive representation of P($N_i = j$).

To solve this problem, I use the *fmincon* function in Matlab optimization toolbox, which is designed to solve problem of minimizing constrained nonlinear multivariate objective function. By choosing "Active Set" optimization strategy, the function performs line search using Sequential Quadratic Programming (SQP), it generate ad solves a Quadratic Programming (QP) sub-problem at each iteration.

4.1.3. Impact of no-show probability and relative waiting cost ratio on schedule

In this section, I solve a problem of scheduling N = 10 homogeneous patients with expected service time θ = 0.5 hours to a clinic session of 10 half-hour length appointment slots, at five levels of patient no-show probability p (0.1, 0.2, 0.3, 0.4, and 0.5) and five levels of relative waiting cost ratio α (0.1, 0.2, 0.3, 0.4, and 0.5).

Figure 4.1 illustrates the optimal solutions under five levels of relative waiting cost ratios, with constant patient no-show probability $p = 0.1$; Figure 4.2 illustrates optimal solutions under five levels of no-show probabilities, with constant relative waiting cost ratio $\alpha = 0.1$. Refer to Section 3.1 for detailed description of optimal solution figure.

Figure 4.1 Schedules of the baseline model at $N = 10$, $p = 0.1$, $\theta = 0.5$

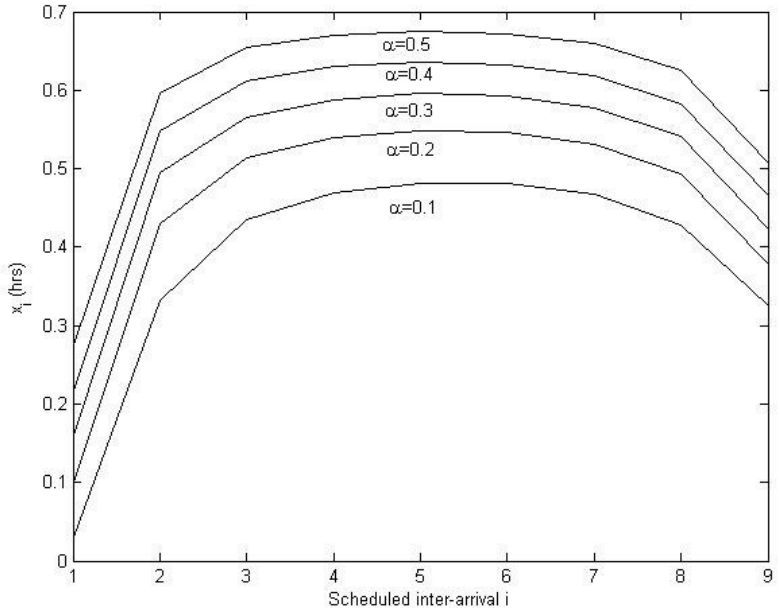


Table 4.1 Schedules of the baseline model at $N = 10$, $p = 0.1$, $\theta = 0.5$

α	X								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.1	0.03	0.33	0.44	0.47	0.48	0.48	0.47	0.43	0.33
0.2	0.10	0.43	0.51	0.54	0.55	0.55	0.53	0.49	0.38
0.3	0.16	0.50	0.57	0.59	0.60	0.59	0.58	0.54	0.42
0.4	0.22	0.55	0.61	0.63	0.64	0.63	0.62	0.58	0.47
0.5	0.27	0.60	0.65	0.67	0.68	0.67	0.66	0.63	0.51

Figure 4.2 Schedules of the baseline model at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

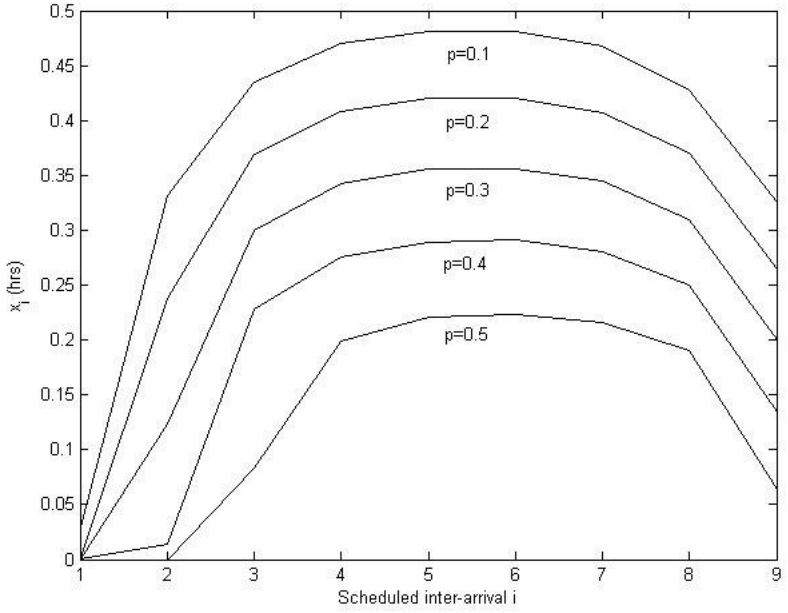


Table 4.2 Schedules of the baseline model at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

p	X								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.1	0.03	0.33	0.44	0.47	0.48	0.48	0.47	0.43	0.33
0.2	0.00	0.24	0.37	0.41	0.42	0.42	0.41	0.37	0.26
0.3	0.00	0.12	0.30	0.34	0.36	0.36	0.35	0.31	0.20
0.4	0.00	0.01	0.23	0.28	0.29	0.29	0.28	0.25	0.13
0.5	0.00	0.00	0.08	0.20	0.22	0.22	0.22	0.19	0.06

As we can see from Figure 4.1, scheduled inter-arrival time increases among the first several patients, it keeps relative constant at a certain level thereafter, and then drops significantly between last two patients. In general, the schedule forms a dome shape, which has been observed by many previous studies. The dome shape "expands" when α increases. The explanation for this phenomenon is straightforward: higher patient waiting cost will cause the system to schedule them with longer inter-arrival times.

As p increases, the first several patients are scheduled closer. When it reaches 0.2, (see Figure 4.2), it starts to schedule the first two patients to arrive together. This is first suggested by Bailey [2], known as Bailey's rule, to reduce the chance of the server being idle at the beginning of a clinic session. As p reaches 0.5, the first three patients are scheduled to the first slot. However, no matter how p changes, only the last two patients are scheduled significantly closer than prior patients, which indicates that the ending part of a schedule is relatively insensitive to no-show probability p. In a case where all patients must show up (without no-show), the solutions of Stein and Cote [60] demonstrate a similar feature of the last two patients scheduled close or together, which supports my observation that inter-arrival time among the last several patients is insensitive to no-show probability.

Comparing Figure 4.1 with Figure 4.2, I observe that the increase of relative waiting cost ratio doesn't change the shape of the schedules, but simply "magnifies" the schedules, whereas the increase of no-show probability "squeezes" the shape of the schedule, by scheduling the first several patients closer to each other.

4.1.4. A comparison with the general model of linear waiting cost

To illustrate how quadratic waiting cost impacts the outcome the optimal schedule, I compare my baseline model with the model developed by Hassin and Mendel [25], which assumes a linear relation between patient waiting cost and waiting time. Their objective function is

$$\min \quad Z = \alpha(1-p)\sum_{i=1}^{N} w_i + (1-\alpha)\left(\sum_{i=1}^{N-1} x_i + w_N\right). \tag{4.7}$$

The solution schedules share similar features with the baseline model, that the first several patients and the last two patients are scheduled closer than the rest of the patients in the middle of the schedule.

Figure 4.3 Schedules of Hassin and Mendel's model at $N = 10$, $p = 0.1$, $\theta = 0.5$

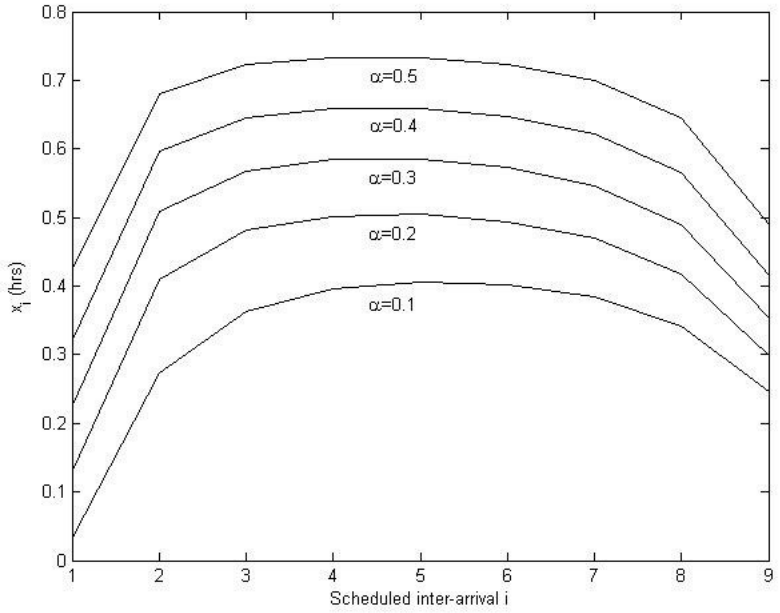


Table 4.3 Schedules of Hassin and Mendel's model at $N = 10$, $p = 0.1$, $\theta = 0.5$

α	X								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.1	0.03	0.27	0.36	0.40	0.41	0.40	0.38	0.34	0.25
0.2	0.13	0.41	0.48	0.50	0.51	0.49	0.47	0.42	0.30
0.3	0.23	0.51	0.57	0.58	0.59	0.57	0.55	0.49	0.35
0.4	0.32	0.60	0.65	0.66	0.66	0.65	0.62	0.57	0.42
0.5	0.42	0.68	0.72	0.73	0.73	0.72	0.70	0.65	0.49

Figure 4.4 Schedules of the Hassin and Mendel's model at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

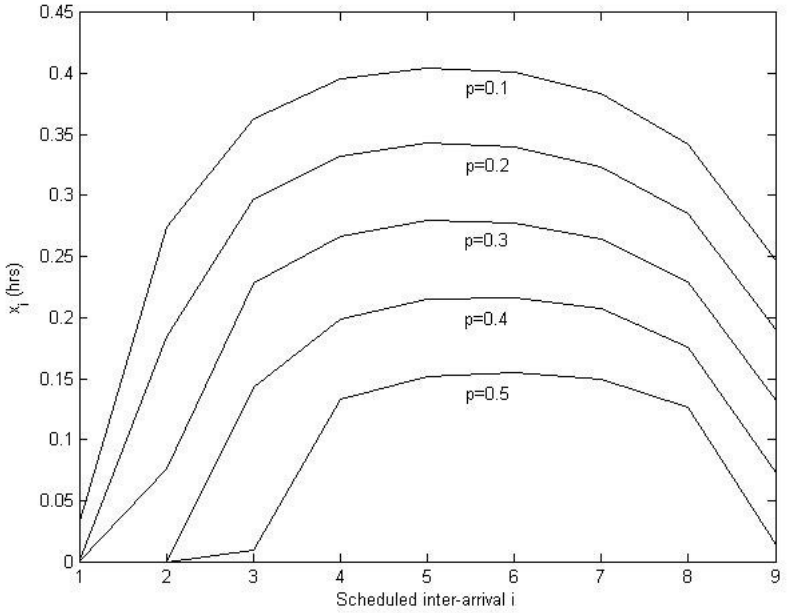


Table 4.4 Schedules of the Hassin and Mendel's model at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

p	X								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.1	0.03	0.27	0.36	0.40	0.41	0.40	0.38	0.34	0.25
0.2	0.00	0.18	0.30	0.33	0.34	0.34	0.32	0.29	0.19
0.3	0.00	0.08	0.23	0.27	0.28	0.28	0.26	0.23	0.13
0.4	0.00	0.00	0.14	0.20	0.22	0.22	0.21	0.18	0.07
0.5	0.00	0.00	0.01	0.13	0.15	0.16	0.15	0.13	0.01

Comparing Figure 4.3 with Figure 4.1, I observe that when relative waiting cost ratio is low, Hassin and Mendel's model has generally shorter scheduled inter-arrival times than the corresponding times in the baseline model; conversely, when relative waiting cost ratio is high, it has generally longer scheduled inter-arrival times than the corresponding values in the baseline

model. Comparing Figure 4.4 with Figure 4.2, I observe that the scheduled inter-arrival times of Hassin and Mendel's model are generally higher than corresponding values in the baseline model, for all levels of no-show probability p.

Figure 4.5 Expected server completion time of the two models at $N = 10$, $p = 0.1$, $\theta = 0.5$

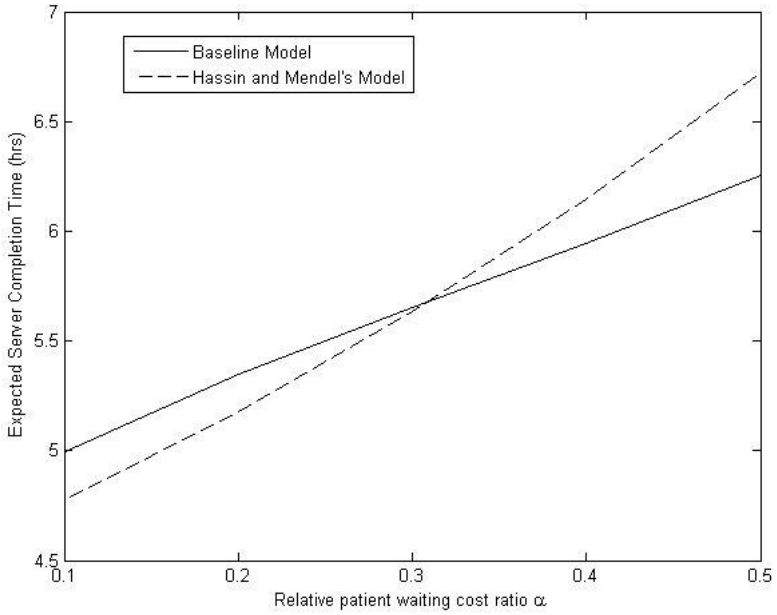


Table 4.5 Expected server completion time of the two models at $N = 10$, $p = 0.1$, $\theta = 0.5$

α	Expected Server completion time	
	Baseline model	Hassin and Mendel's model
0.1	5.00	4.78
0.2	5.35	5.18
0.3	5.65	5.64
0.4	5.95	6.14
0.5	6.25	6.72

Figure 4.6 Expected server completion time of the two models at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

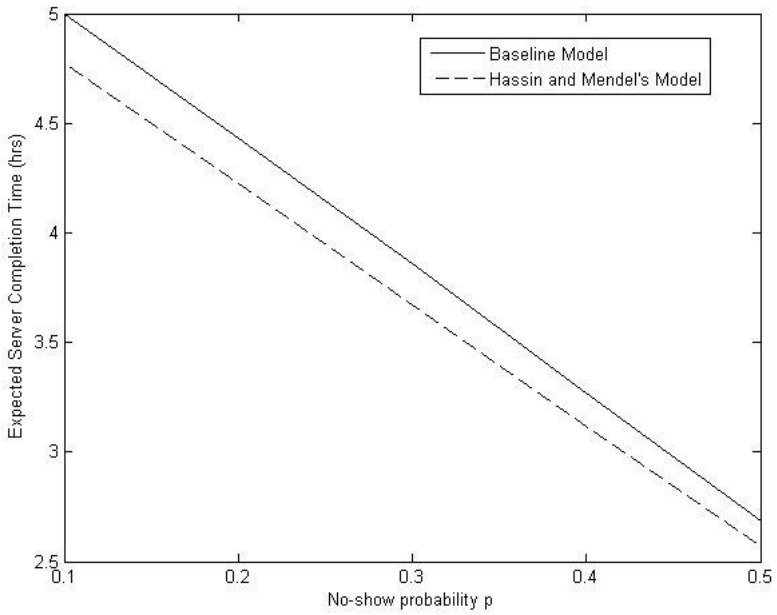


Table 4.6 Expected server completion time of the two models at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

p	Expected Server completion time	
	Baseline model	Hassin and Mendel's model
0.1	5.00	4.78
0.2	4.43	4.23
0.3	3.86	3.67
0.4	3.27	3.12
0.5	2.68	2.57

Figure 4.5 and Figure 4.6 present server completion times of the two models under five levels of relative cost ratio and no-show probability, respectively. They validate my observations above from a perspective of total server's time.

As an important indicator of patient satisfaction, patient waiting time is paid more attention in my study. Figure 4.7 and Figure 4.8 compare total expected patient waiting times of the two models under different values of relative cost ratio and no-show probability, respectively. When the relative cost ratio is lower than 0.3, the baseline model achieves better performance in terms of patient waiting. If the relative cost ratio is fixed at 0.1 and no-show probability changes, the baseline model always outperforms the linear waiting cost model. Intuitively, with quadratic patient waiting cost, the baseline model will try harder to avoid excessive waiting.

Figure 4.7 Total expected patient waiting time of the two models at $N = 10$, $p = 0.1$, $\theta = 0.5$

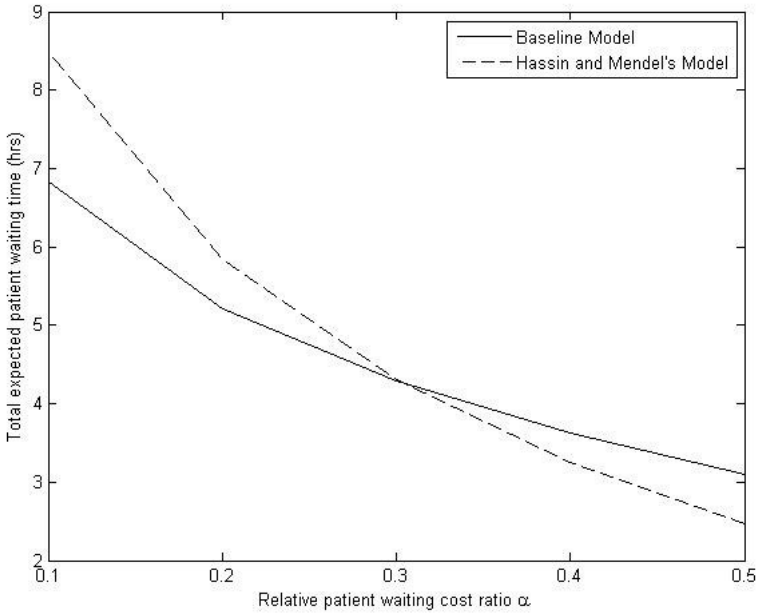


Table 4.7 Total expected patient waiting time of the two models at $N = 10$, $p = 0.1$, $\theta = 0.5$

α	Total expected patient waiting time	
	Baseline model	Hassin and Mendel's model
0.1	6.84	8.49
0.2	5.22	5.84

0.3	4.29	4.30
0.4	3.63	3.25
0.5	3.10	2.46

Figure 4.8 Total expected patient waiting time of the two models at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

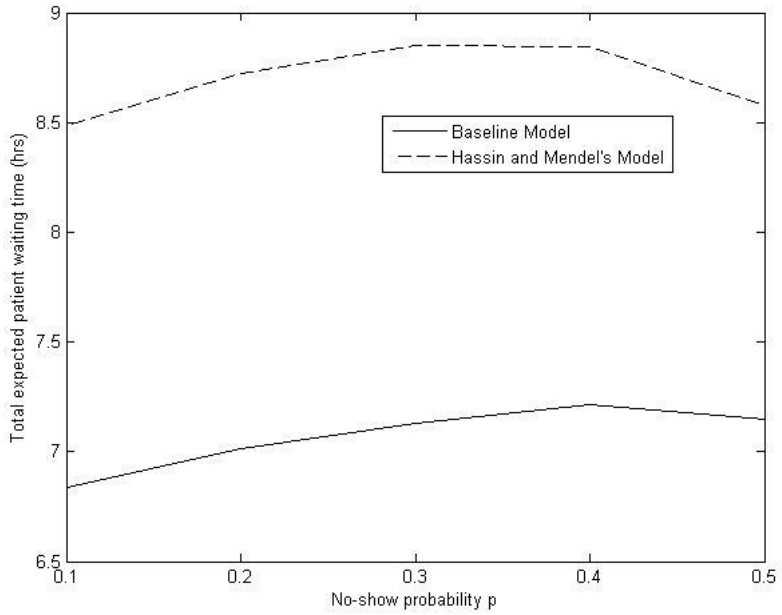


Table 4.8 Total expected patient waiting time of the two models at $N = 10$, $\alpha = 0.1$, $\theta = 0.5$

p	Total expected patient waiting time	
	Baseline model	Hassin and Mendel's model
0.1	6.84	8.49
0.2	7.01	8.73
0.3	7.13	8.85
0.4	7.22	8.84
0.5	7.15	8.58

4.2. The queuing model for patients with nonhomogeneous no-show probabilities

In this section, I relax the assumption of constant and identical patient no-show probability by replacing p with a no-show probability vector $P = (p_1,\ldots, p_i,\ldots, p_N)$, where p_i is the no-show probability of the ith patient. The resulting objective function becomes to be:

$$\min \quad Z = \alpha \sum_{i=1}^{N} (1-p_i) w_i^2 + (1-\alpha)\left(\sum_{i=1}^{N-1} x_i + w_N \right). \tag{4.8}$$

Accordingly, the recursive representation of $P(N_i = j)$ is modified to be as the following:

When $j = 0$

$$P(N_i = 0) = \sum_{k=0}^{i-2} P(N_{i-1} = k)\left(1 - \sum_{l=0}^{k-1} e^{-\frac{x_{i-1}}{\theta}} \frac{x_{i-1}^{\ l}}{\theta^k l!} - (1-p_{i-1}) e^{-\frac{x_{i-1}}{\theta}} \frac{x_{i-1}^{\ k}}{\theta^k k!} \right), \tag{4.9}$$

When $j > 0$

$$\begin{aligned} P(N_i = j) = & \sum_{k=1}^{i-j-1} P(N_{i-1} = j+k-1) e^{-\frac{x_{i-1}}{\theta}} \frac{x_{i-1}^{\ k-1}}{\theta^{k-1}(k-1)!} \left((1-p_{i-1}) \frac{x_{i-1}}{\theta k} + p_{i-1} \right) \\ & + (1-p_{i-1}) P(N_{i-1} = j-1) e^{-\frac{x_{i-1}}{\theta}} \end{aligned}. \tag{4.10}$$

4.2.1. Numerical results

Given a patient's no-show probability can be predetermined based on his or her profile at the time of booking [18] and [22], I compare the following three booking heuristics:

- Lower no-show first: patients are scheduled in ascending order of estimated no-show probability;

- Higher no-show first: patients are scheduled in descending order of estimated no-show probability;
- Higher no-show in the middle: patients with higher estimated no-show probability are scheduled closer to the middle slots.

To determine right no-show probability values to use, I take the reviewed empirical data on no-show rates of MRI patients. Each schedule includes 10 patients with no-show probability ranges from 0.04 to 0.4, with 0.04 unit increment patient by patient from the lowest to the highest. Table 4.9 summarizes the no-show probability vectors of the three booking heuristics.

Table 4.9 No-show probability vectors of the three booking heuristics

Patient number i	Patient no-show probability p		
	lower no-show first heuristic	higher no-show first heuristic	higher no-show in the middle heuristic
1	0.04	0.40	0.04
2	0.08	0.36	0.12
3	0.12	0.32	0.20
4	0.16	0.28	0.28
5	0.20	0.24	0.36
6	0.24	0.20	0.40
7	0.28	0.16	0.32
8	0.32	0.12	0.24
9	0.36	0.08	0.16
10	0.40	0.04	0.08

Optimal schedules at five levels of α (0.1, 0.2, 0.3, 0.4, and 0.5) under the three booking heuristics are illustrated by Figure 4.9, Figure 4.10, and Figure 4.11, respectively. A common

feature shared by all three figures is that the increase of relative waiting cost ratio doesn't change the general shape of optimal schedule, it simply "magnifies" the schedule. The well-known dome shaped structure of schedules has been observed again in Figure 4.9 and Figure 4.10. However, Figure 4.11 exhibits a significant drop of scheduled inter-arrival time in the middle among the patients with highest no-show probabilities, at all levels of α. It implies that higher no-show tends to create shorter scheduled inter-arrival time, which can prevent server from excessive idleness when no-show happens. With two domes exhibited in Figure 4.11, the probabilistic steady-state in the middle of a clinic session is broken, due to the non-ascending and non-descending order of no-show probabilities of the patients scheduled.

Figure 4.9 Schedules of the lower no-show first heuristic

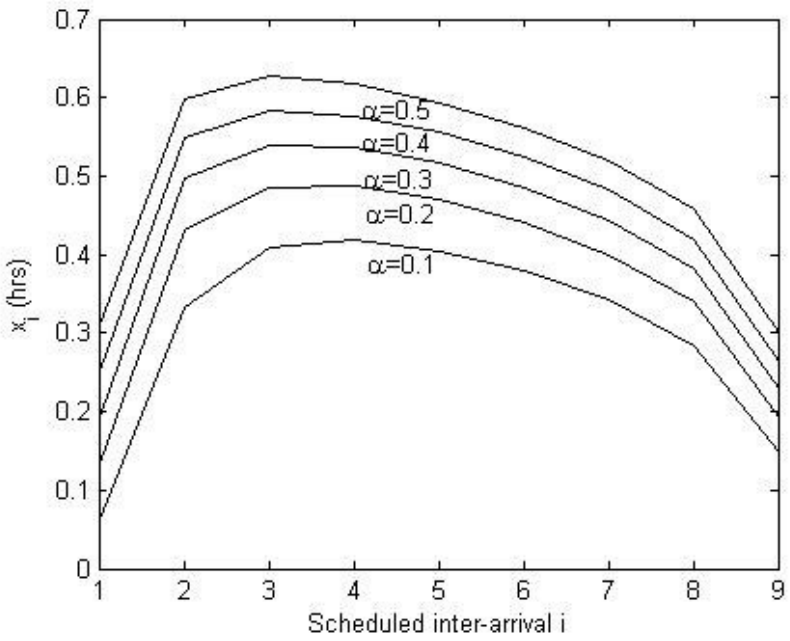


Table 4.10 Schedules of the lower no-show first heuristic

α	X								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9

0.1	0.06	0.33	0.41	0.42	0.41	0.38	0.34	0.28	0.15
0.2	0.13	0.43	0.49	0.49	0.47	0.44	0.40	0.34	0.19
0.3	0.19	0.50	0.54	0.54	0.52	0.49	0.44	0.38	0.23
0.4	0.25	0.55	0.58	0.58	0.56	0.52	0.48	0.42	0.26
0.5	0.31	0.60	0.63	0.62	0.59	0.56	0.52	0.46	0.30

Figure 4.10 Schedules of the higher no-show first heuristic

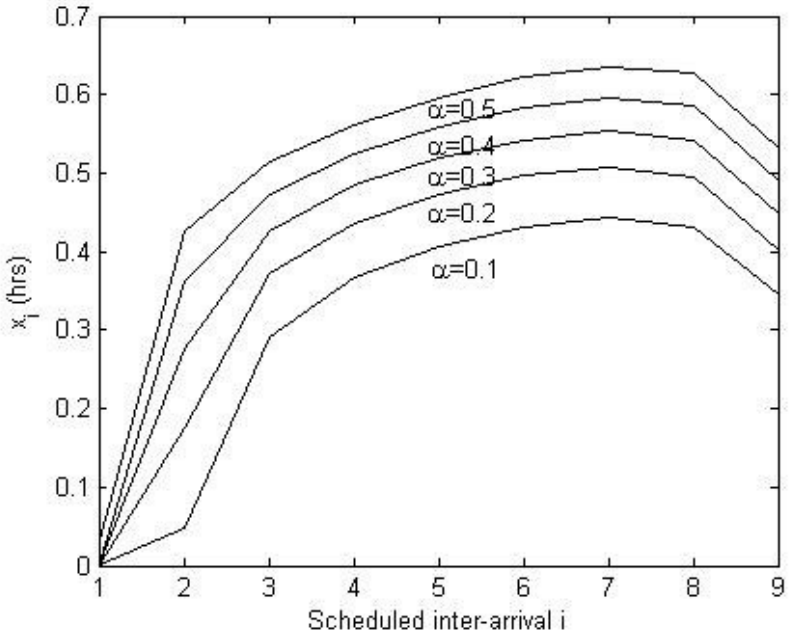


Table 4.11 Schedules of the higher no-show first heuristic

α	X								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.1	0.00	0.05	0.29	0.37	0.41	0.43	0.44	0.43	0.34
0.2	0.00	0.18	0.37	0.44	0.47	0.50	0.51	0.49	0.40
0.3	0.00	0.28	0.43	0.48	0.52	0.54	0.55	0.54	0.45
0.4	0.00	0.36	0.47	0.52	0.56	0.58	0.60	0.58	0.49
0.5	0.03	0.43	0.51	0.56	0.60	0.62	0.64	0.63	0.53

Figure 4.11 Schedules of the higher no-show in the middle heuristic

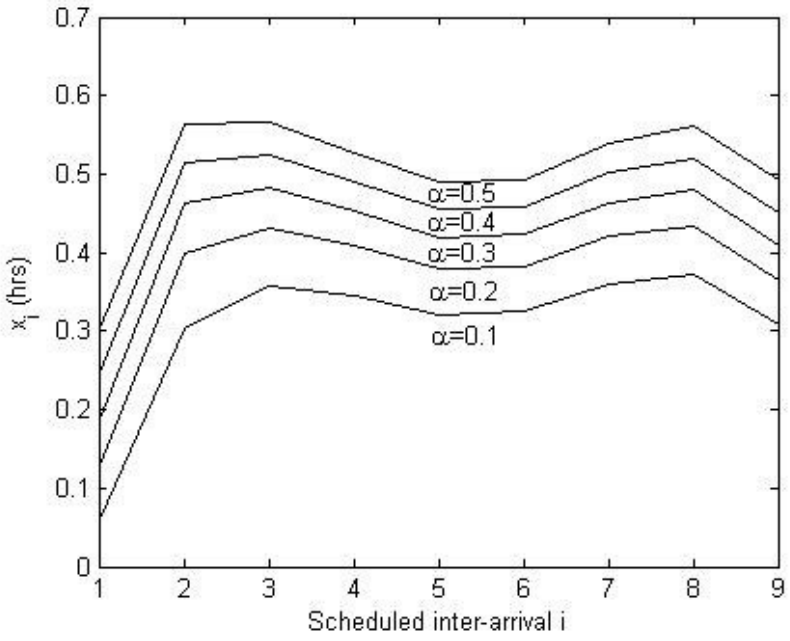


Table 4.12 Schedules of the higher no-show in the middle heuristic

α	X								
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.1	0.06	0.30	0.36	0.34	0.32	0.33	0.36	0.37	0.31
0.2	0.13	0.40	0.43	0.41	0.38	0.38	0.42	0.43	0.36
0.3	0.19	0.46	0.48	0.45	0.42	0.42	0.46	0.48	0.41
0.4	0.24	0.51	0.53	0.49	0.45	0.46	0.50	0.52	0.45
0.5	0.30	0.56	0.57	0.53	0.49	0.49	0.54	0.56	0.49

Figure 4.12, Figure 4.13 and Figure 4.14 exhibit expected waiting times of all 10 scheduled patients. i of X axis denotes patient number (i.e. i = 1 denotes 1st patient), w_i of Y axis denotes the corresponding expected waiting time in unit of hour. Comparing among the three figures, I observe some common features: starting at $w_1 = 0$, w_i increases monotonically, with a significant

boost from w_1 to w_2, and then a significant boost from w_9 to w_{10}; higher waiting cost ratio α causes longer expected waiting time w_i. At all levels of α, the higher no-show first heuristic has overall shortest w_i, and least variance of expected patient waiting times among the patient.

Figure 4.12 Expected patient waiting times of the lower no-show the first heuristic

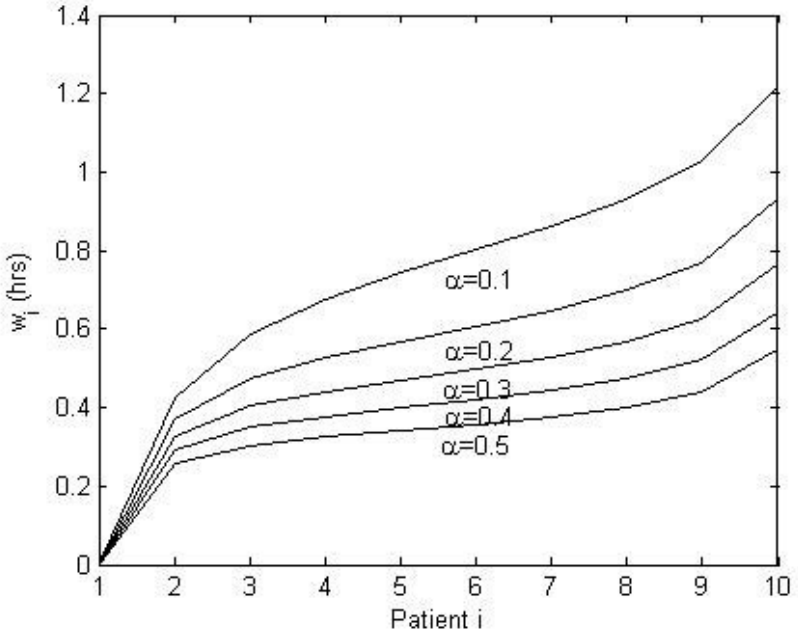


Table 4.13 Expected patient waiting times of the lower no-show the first heuristic

α	Expected patient waiting time									
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
0.1	0.00	0.43	0.59	0.68	0.74	0.80	0.86	0.93	1.03	1.22
0.2	0.00	0.37	0.47	0.53	0.57	0.61	0.65	0.70	0.77	0.93
0.3	0.00	0.33	0.40	0.44	0.47	0.50	0.53	0.57	0.62	0.76
0.4	0.00	0.29	0.35	0.38	0.40	0.42	0.44	0.47	0.52	0.64
0.5	0.00	0.26	0.30	0.33	0.34	0.36	0.38	0.40	0.44	0.55

Figure 4.13 Expected patient waiting times of the higher no-show first heuristic

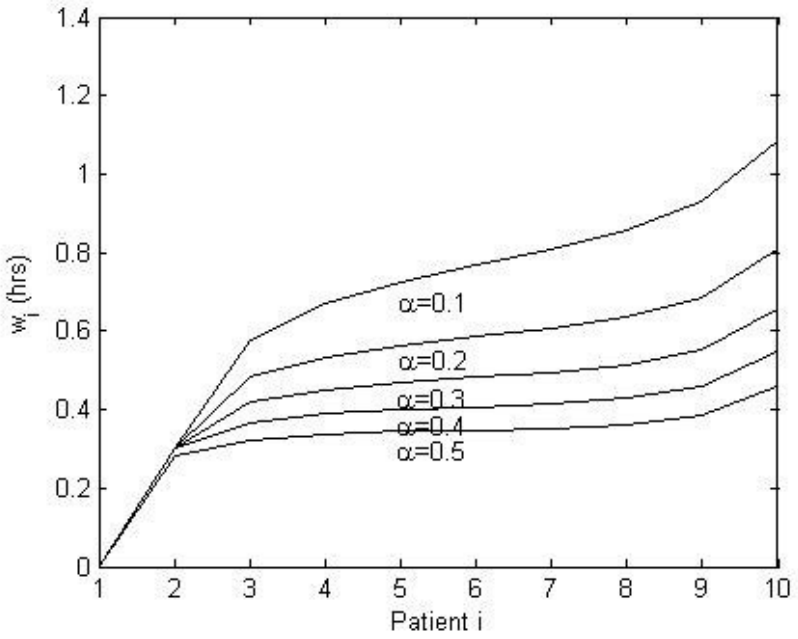


Table 4.14 Expected patient waiting times of the higher no-show first heuristic

α	Expected patient waiting time									
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
0.1	0.00	0.30	0.58	0.67	0.73	0.77	0.81	0.86	0.93	1.08
0.2	0.00	0.30	0.48	0.53	0.56	0.59	0.61	0.64	0.69	0.81
0.3	0.00	0.30	0.42	0.45	0.47	0.48	0.50	0.52	0.55	0.65
0.4	0.00	0.30	0.37	0.39	0.40	0.41	0.42	0.43	0.46	0.55
0.5	0.00	0.28	0.32	0.34	0.34	0.35	0.35	0.36	0.38	0.46

Figure 4.14 Expected waiting times of the higher no-show in the middle heuristic

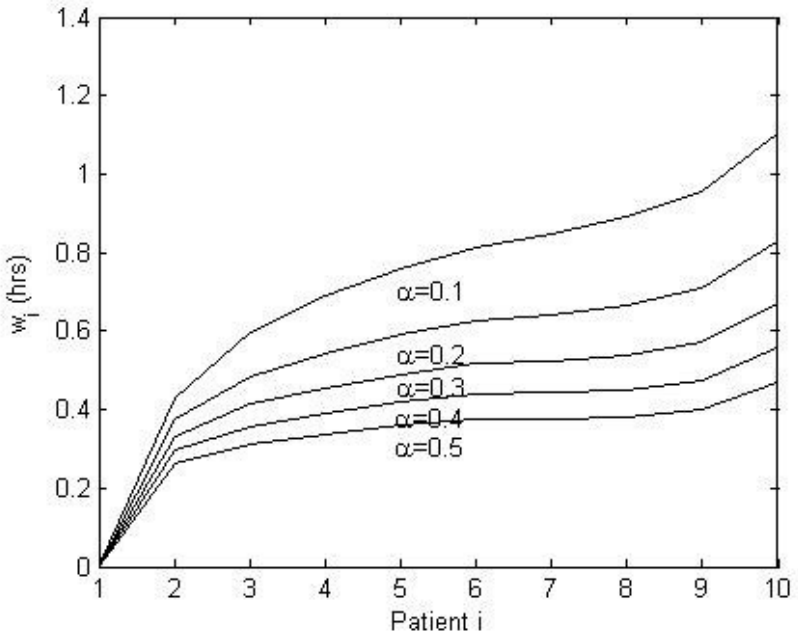


Table 4.15 Expected waiting times of the higher no-show in the middle heuristic

α	Expected patient waiting time									
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
0.1	0.00	0.43	0.60	0.69	0.76	0.81	0.85	0.89	0.96	1.10
0.2	0.00	0.37	0.48	0.54	0.59	0.62	0.64	0.67	0.71	0.83
0.3	0.00	0.33	0.41	0.46	0.49	0.52	0.53	0.54	0.57	0.67
0.4	0.00	0.30	0.36	0.39	0.42	0.44	0.44	0.45	0.47	0.56
0.5	0.00	0.26	0.31	0.34	0.36	0.38	0.38	0.38	0.40	0.47

Figure 4.15 compares the objective function values of the three booking heuristics. It shows the lower no-show first heuristic results in significantly higher system costs than the other two heuristics, among which the higher no-show in the middle heuristic is slightly better, at all levels of α (see Table 4.16). In general, total system cost decreases as α increases, with an expectation observed, that is the lower no-show first heuristic has a boost of its total cost at $\alpha = 0.3$.

Figure 4.15 Total system costs of the three heuristics at different levels of α

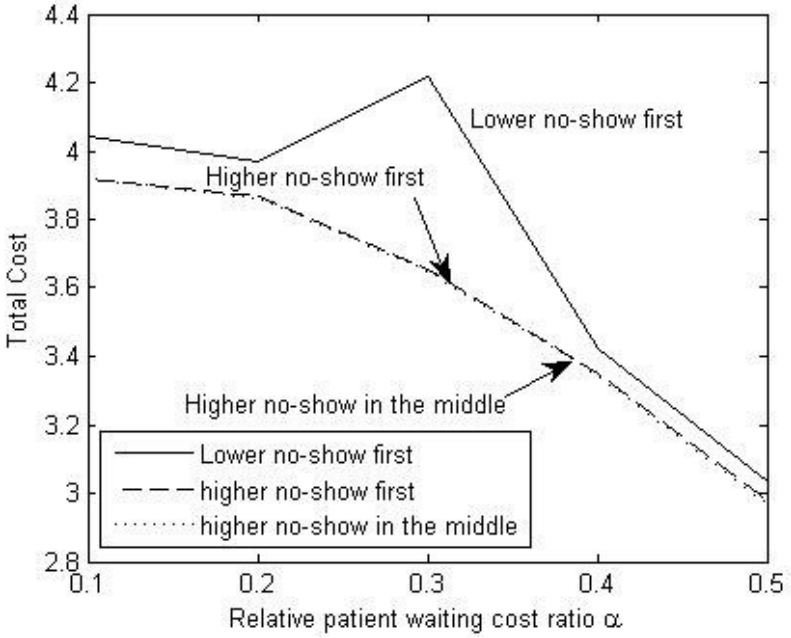


Table 4.16 Total costs of the three heuristics at different levels of α

α	Total cost		
	Lower no-show first	Higher no-show first	Higher no-show in the middle
0.1	4.05	3.92	3.92
0.2	3.97	3.87	3.86
0.3	4.22	3.65	3.65
0.4	3.42	3.35	3.34
0.5	3.03	2.98	2.97

Interestingly, Zeng et al. [75] reports a contradictory finding, in which a solution very similar to my lower no-show first heuristic achieves best system performance. I suspect it's because of the linear patient waiting cost and high patient waiting cost ratio (unit overflow cost) used in their study. To prove my suspicion, I compare my model with linear patient waiting cost model, whose objective function is described as the following:

$$\min \quad Z = \alpha \sum_{i=1}^{N} (1 - p_i) w_i + (1 - \alpha) \left(\sum_{i=1}^{N-1} x_i + w_N \right). \tag{4.11}$$

All constraints and parameter values are kept the same as my model.

I compare the performances of the lower no-show first heuristic and the higher no-show first heuristic in the linear cost model. In Figure 4.16, as α increases the performance of the two heuristics becomes very close, with the lower no-show first heuristic slightly better when $\alpha \geq 0.7$ (objective function 0.08% lower than the higher no-show first heuristic at $\alpha = 0.7$; 0.11% lower at $\alpha = 0.8$; 0.09% lower at $\alpha = 0.9$, see Table 4.17 for detailed values). Under the assumption of linear waiting cost, the lower no-show first heuristic could outperform the higher no-show first heuristic at the high end of α.

Compare Figure 4.15 with Figure 4.16, I find that under the assumption of nonhomogeneous no-show probabilities, quadratic a patient waiting cost tends to make the higher no-show first heuristic more favorable.

Figure 4.16 Total system costs of the two heuristics under linear patient waiting cost assumption

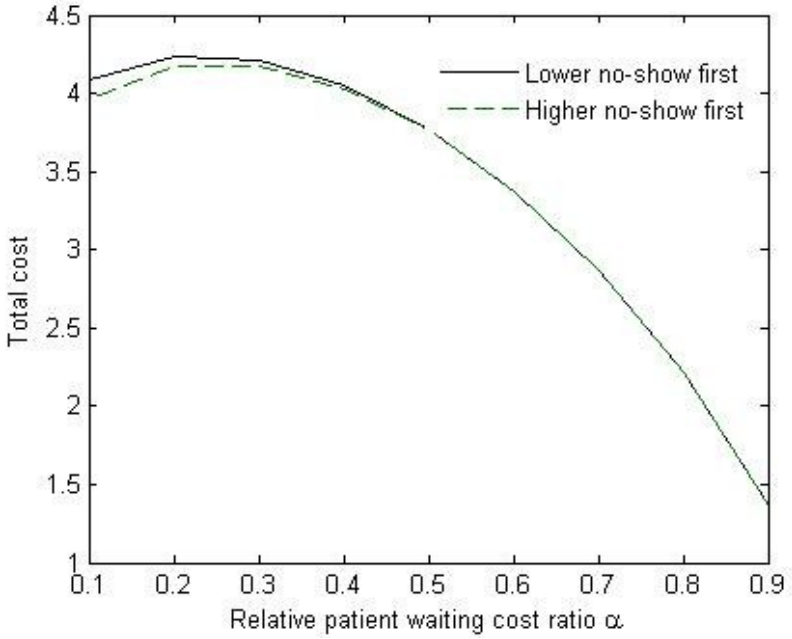


Table 4.17 Total system costs of the two heuristics under linear patient waiting cost assumption

α	Total cost	
	Lower no-show first	Higher no-show first
0.1	4.082	3.967
0.2	4.240	4.168
0.3	4.215	4.172
0.4	4.051	4.030
0.5	3.769	3.761
0.6	3.374	3.373
0.7	2.861	2.864
0.8	2.212	2.214
0.9	1.368	1.369

Among the three booking heuristics, at all levels of α, the higher no-show first heuristic results in lowest total expected patient waiting time, while the lower no-show first heuristic results in

longest total expected patient waiting time (see Figure 4.17). It implies that if most no-shows happen close to the beginning of a clinical session, the rest patients may wait less.

As shown by Figure 4.18, the expected server completion times of the three booking heuristics are close, with the greatest difference equal to 0.05 hour (see Table 4.19 for details), which indicates that expected server completion time is relatively insensitive to the arrangement of patients based on estimated no-show probability.

Figure 4.17 Total expected patient waiting time of the three heuristics

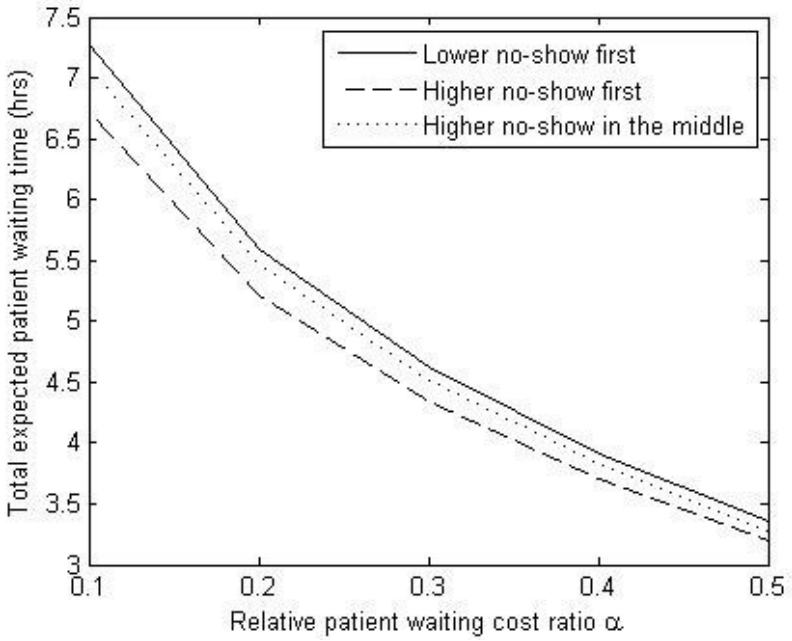


Table 4.18 Total expected patient waiting time of the three heuristics

α	Total expected patient waiting time		
	Lower no-show first	Higher no-show first	Higher no-show in the middle
0.1	7.28	6.73	7.09
0.2	5.60	5.21	5.46
0.3	4.62	4.34	4.51

0.4	3.92	3.71	3.83
0.5	3.35	3.19	3.28

Figure 4.18 Expected server completion times of the three heuristics

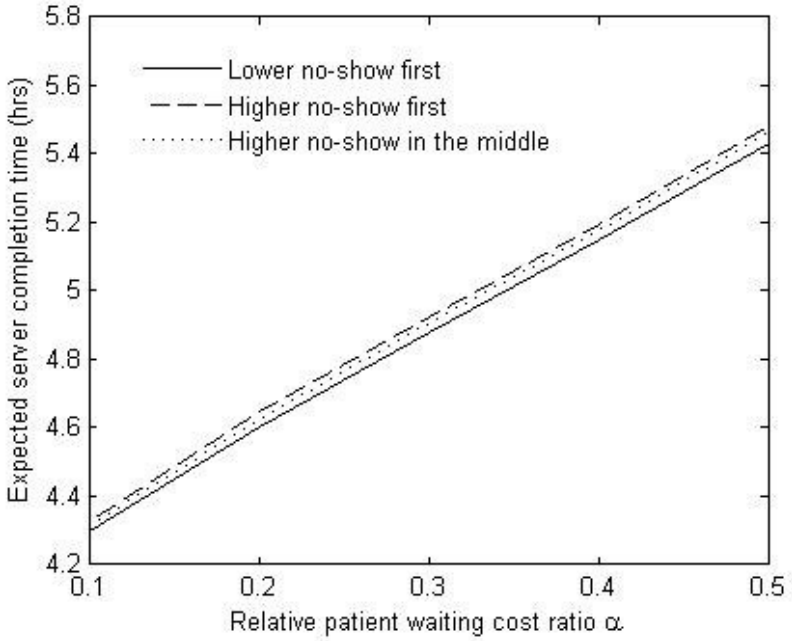


Table 4.19 Expected server completion times of the three heuristics

α	Expected server completion time		
	Lower no-show first	Higher no-show first	Higher no-show in the middle
0.1	4.30	4.32	4.31
0.2	4.61	4.65	4.63
0.3	4.88	4.93	4.91
0.4	5.15	5.20	5.18
0.5	5.43	5.48	5.46

4.3. The queuing model for patients with nonhomogeneous no-show probabilities and waiting cost ratios

Primarily motivated by different waiting costs between inpatients and outpatients, I decide to extend my study to a model which includes generalized nonhomogeneous patient waiting cost ratios.

4.3.1. Assumptions and model description

In this model, I relax the assumption of constant and identical waiting cost ratio among patients. In addition to no-show probability, waiting cost ratio varies from one patient to another. To represent varied waiting cost ratios, I replace c_W with a waiting cost ratio vector $C_W = [c_{W1},\ldots, c_{Wi},\ldots, c_{WN}]$, where c_{Wi} denotes the hourly waiting cost of the ith scheduled patient.

With the changes, the objective function becomes to be

$$\min \quad Z = \sum_{i=1}^{N} (1-p_i) c_{wi} w_i^2 + c_I \left(\sum_{i=1}^{N-1} x_i + w_N \right). \tag{4.12}$$

Since c_W is not a constant anymore, I am not able to use α to represent cost ratios (of patient waiting time, server idle time and server overtime). Instead, I choose to use the absolute monetary value representation. Constraints stay the same as in Section 4.2.

I assume patient waiting cost is positively correlated with no-show probability. This is reasonable, given the fact that a patient type with higher no-show probability usually has higher waiting cost. Specifically, in appointment scheduling practices, inpatients, outpatients and emergency patients are featured with ascending no-show probability and waiting cost ratio. A

walk-in emergency patient could be considered as being scheduled to a certain slot, but with very high no-show probability and waiting cost ratio. Based on this assumption, I decide to pair waiting cost ratio with no-show probability, by creating a coefficient β, which is defined as $c_{wi} = \beta p_i$.

After applying β to equation 4.12, the objective function can also be expressed as

$$\min \quad Z = \sum_{i=1}^{N} p_i (1 - p_i) \beta w_i^2 + c_I \left(\sum_{i=1}^{N-1} x_i + w_N \right). \tag{4.13}$$

4.3.2. Numerical results

I evaluate performances of the three no-show probability based booking heuristics defined by Table 4.9, at five levels of β (50, 100, 150, 200, and 250). β = 50 translates into an average waiting cost ratio of \$11/hr, while β = 250 translates into an average waiting cost ratio of \$55/hr. The range of β (\$2/hr to \$100/hr) enables c_{wi} to cover average hourly wages of most occupations. For the server, I estimate its idle cost as the average hourly operating cost by taking the ratio of total annual operating cost and dividing by the number of weeks in a year (52) times the number of working days per week (5) and times the number of hours per day (5). The average purchase cost for a new MRI machine is approximately \$2 million; it costs about \$870,000 to install and approximately \$1 million per year to run. The life span of a state of the art MRI machine is about 10 to 15 years. I assume it to be 10 years in this study. Therefore, the total annual operating cost is \$1,287,000, which translates into hourly idle cost c_I = \$990/hr.

It reinforces my observation in Section 4.2 that the higher no-show first heuristic performs better than lower no-show first heuristic in terms of overall system cost. It appears to be a very robust

heuristic against a wide range of parameters. Note that the higher no-show first heuristic becomes significantly better than higher no-show in the middle heuristic, which differs from what I observed in Figure 4.15. The monotonic increase of expected patient waiting time determines that later patients always have longer expected waiting times. Since patients with higher no-show probabilities are assigned higher waiting cost ratios, if I schedule patients with high waiting cost ratios at the beginning, the total patient waiting cost would be minimized. For the same sake, patient waiting times tend to be longer, and therefore, total expect patient waiting time would be maximized. Figure 4.20 compares the patient waiting performance of the three heuristics at all five levels of β. It presents a result contradictory to Figure 4.17, with the higher no-show first heuristic resulting in highest patient waiting times.

The server completion time performance is illustrated by Figure 4.21, in which the higher no-show first dominates the other two heuristics. Compared with Section 4.2, by assuming patients with higher no-show probabilities have higher waiting cost ratios, the higher no-show first heuristic changes from the one with lowest patient waiting but highest server time to one with highest patient waiting but lowest server time, that is, from a more patient friendly heuristic to a more server friendly heuristic.

Figure 4.19 Total system cost of the three heuristics at different levels of β

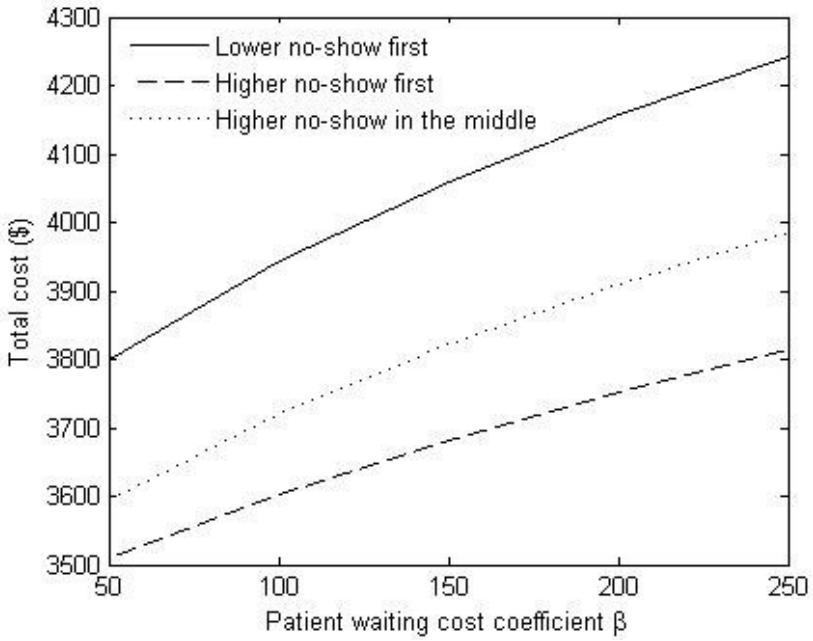


Table 4.20 Total system cost of the three heuristics at different levels of β

β	Total cost		
	Lower no-show first	Higher no-show first	Higher no-show in the middle
50	3798	3510	3595
100	3945	3604	3721
150	4060	3683	3823
200	4157	3752	3910
250	4241	3815	3986

Figure 4.20 Total expected patient waiting time of the three heuristics at different levels of β

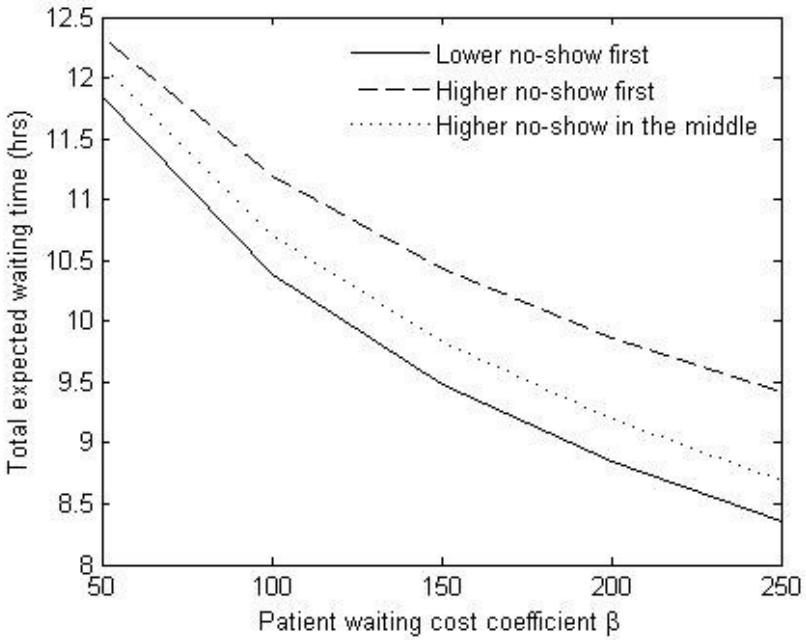


Table 4.21 Total expected patient waiting time of the three heuristics at different levels of β

β	Total cost		
	Lower no-show first	Higher no-show first	Higher no-show in the middle
50	11.85	12.34	12.10
100	10.38	11.19	10.70
150	9.48	10.43	9.82
200	8.85	9.86	9.19
250	8.35	9.41	8.70

Figure 4.21 Expected server completion time of the three heuristics at different levels of β

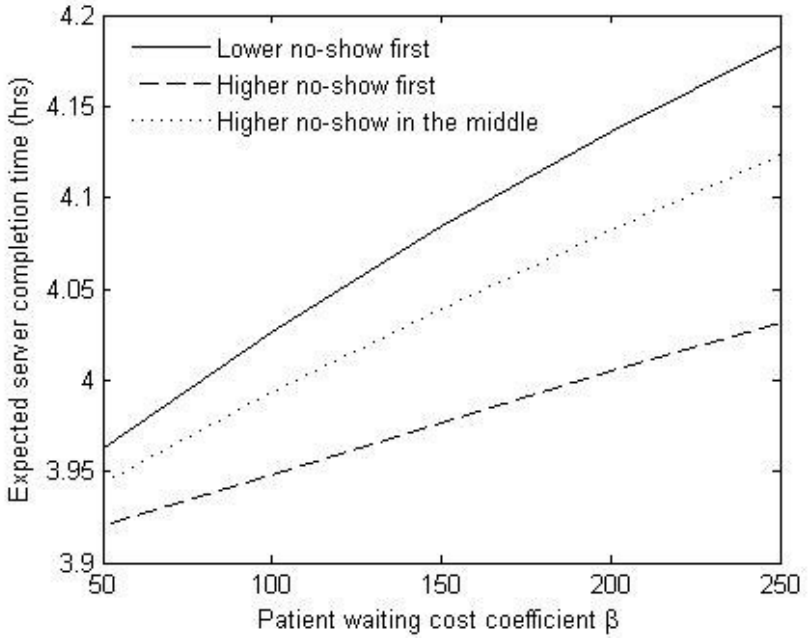


Table 4.22 Expected server completion time of the three heuristics at different levels of β

β	Total cost		
	Lower no-show first	Higher no-show first	Higher no-show in the middle
50	3.96	3.92	3.94
100	4.03	3.95	3.99
150	4.08	3.98	4.04
200	4.14	4.01	4.08
250	4.18	4.03	4.12

Chapter 5 A HYBRID OVERBOOKING MODEL FOR MULTI-CATEGORY PATIENTS WITH NO-SHOW

In this chapter, I address a daily scheduling problem of allocating available diagnostic capacity, in the form of equal-length appointment slots, among multiple categories of patients, to maximize system net revenue consisting of service revenue, equipment idle cost, patient waiting cost, and patient deny penalty cost. I analyze the decision variable (the number of patients to be scheduled, and a combination of schedule patient inter-arrival times) while three environmental factors (outpatient no-show probability, server hourly idle cost, and inpatient service fee) are varied. Three types of patients are considered in this study: inpatients, who have low level of no-show probability and waiting cost; outpatients, who have medium level of no-show probability and waiting cost; and emergency patients, who usually show up as walk-in, with extremely high waiting cost. A hybrid overbooking strategy is employed to take walk-in emergency patients into consideration of this model.

5.1. Assumptions

I consider an unknown emergency patient as a called-in outpatient, with much higher no-show probability. When an emergency patient arrives, he or she needs to be served with the next available slot, which is equivalent to a case that the patient is scheduled to next slot. Thus, I assume each appointment slot is booked to 1 virtual emergency patient with higher priority than the outpatients or inpatients. Arrival of emergency patients is random during one clinic session. Compared to scheduled outpatients and inpatients, the demand from emergency patients is

relatively low. It is unlikely for more than one emergency patient to arrive within service duration of one slot. Therefore, I assume there can be at most one emergency patient arrival during each slot, with fixed arrival probability. If the emergency patient cannot be served with the next available slot, that is, there is one emergency patient waiting in front of the queue, he/she will leave the system.

Outpatients and inpatients are scheduled, if a scheduled patient is not served by the end of a clinic session, a scheduled but denied penalty cost will be incurred. Penalty cost for inpatients is much higher than for outpatient, as denying an inpatient can result into significant cost the hospital due to another day of stay at the hospital.

Compared with the airline/hotel overbooking problems, which have fixed capacity (number of seats/rooms), the capacity of a diagnostic facility is relatively flexible, which can accommodate a certain level of overtime. In this problem, when it passes ending time of a clinic session, the system will finish the patient in service, and stopping any more patients.

Unlike previously clinic overbooking studies which associate one or more scheduled patients directly with each appointment slot, I don't assign any scheduled patient to a particular slot. Scheduled patients are served on as FCFS (First Come First Served) basis, regardless emergency patient.

For each outpatient or inpatient served, the MRI facility receives a certain amount of service revenue, in form of an insurance fee charge. In most general hospitals, scanning fee charged on outpatients is much higher than inpatients.

5.2.Model description

In this model, I am interested in determining how many patients I should schedule, and how long the scheduled inter-arrival times should be, in order to achieve the best system net revenue in one clinic session of Q equal-length appointment slots with each last θ hours. Thus, the decision variable is a combination of scheduled inter-arrival times for a variable number of patients N to be scheduled. During each slot, there is a probability of $1 - p_e$ that an emergency patient arrives.

I assume non-homogeneous scheduled patients, with hourly waiting costs represented by a cost vector $C_W = [c_{W1},\ldots, c_{Wi},\ldots, c_{WN}]$, where c_{Wi} denotes hourly waiting cost of the ith patient; no-show represented by a probability vector $P = [p_1,\ldots, p_i,\ldots, p_N]$, where p_i denotes the probability of ith patient being a no-show; service revenue represented by a revenue vector $R = [r_1,\ldots, r_i,\ldots, r_N]$, where r_i denotes service fee for the ith patient; and penalty cost represented by a cost vector $C_p = [c_{p1},\ldots, c_{pi},\ldots, c_{pN}]$, where c_{pi} is the penalty incurred to the system if ith patient is denied for service. Emergency patients are homogeneous with c_{We} and r_e as hourly waiting cost and service revenue respectively.

5.2.1. Notations

N	number of patients scheduled
n	total number of patients served
Q	number of appointment slots per clinical session
x_i	scheduled inter-arrival time between ith patient and $(i + 1)$st patient
r_i	service revenue of the ith patient
r_e	service revenue of an emergency patient
p_i	no-show probability of the ith patient
p_e	no-show probability of emergency patient

c_{Wi} waiting cost ratio of the *i*th patient

c_{We} waiting cost ratio of emergency patient

c_I server idle cost ratio

c_{Pi} penalty cost of denying the *i*th patient

θ expected service time

5.2.2. Objective function and constraints

The system net revenue is defined as total service revenue subtracts patient waiting cost, equipment idle cost and patient deny penalty cost:

$$\begin{aligned}\max \quad Z = & \sum_{i=1}^{n} r_i (1-p_i) + r_e (1-p_e) Q - \sum_{i=1}^{n} (1-p_i) c_{wi} w_i^2 - c_{we} (1-p_e) Q\theta \\ & - c_I \left[\sum_{i=1}^{n-1} x_i + w_n + (1-p_n)\theta - (1-p_e) Q\theta - \sum_{i=1}^{n} (1-p_i)\theta \right] - \sum_{i=n+1}^{N} (1-p_i) c_{Pi}\end{aligned} \quad . \tag{3.1}$$

Out of the *N* scheduled patients, the first *n* patients consume all the time of the clinic session. By intuition, in the solution, *N* equals to the threshold value *n*, greater or smaller than which would result in sub optimum. w_i denotes expected waiting time given the *i*th scheduled patient shows up.

$\sum_{i=1}^{n} r_i (1-p_i)$ and $r_e (1-p_e) Q$ represent total expected service revenue collected from scheduled patients and emergency patients, respectively;

$\sum_{i=1}^{n} (1-p_i) c_{wi} w_i^2$ and $c_{we} (1-p_e) Q\theta$ represent expected total waiting cost of scheduled patients and emergency patients, respectively. Note we assume quadratic waiting cost for scheduled patients, which is consistent with Chapter 4.

$c_I\left[\sum_{i=1}^{n-1} x_i + w_n + (1-p_n)\theta - (1-p_e)Q\theta - \sum_{i=1}^{n}(1-p_i)\theta\right]$ represents expected total server idle cost; $\sum_{i=n+1}^{N}(1-p_i)c_{Pi}$ represents expected total penalty cost.

Removing constant parts, I obtain

$$\max \quad Z = \sum_{i=1}^{n} r_i(1-p_i) - \sum_{i=1}^{n}(1-p_i)c_{wi}w_i^2 - c_I\left[\sum_{i=1}^{n-1} x_i + w_n - \sum_{i=1}^{n}(1-p_i)\theta\right] - \sum_{i=n+1}^{N}(1-p_i)c_{Pi}, \qquad (3.2)$$

Where n is defined by the following inequality constraints

$$w(n) + \sum_{i=1}^{n-1} x(i) < Q\theta, \qquad (3.3)$$

and

$$w(n+1) + \sum_{i=1}^{n} x(i) \geq Q\theta. \qquad (3.4)$$

w_i is defined as the following:

$$w_i = \begin{cases} 0 & i = 1 \\ \theta\sum_{j=0}^{i+1} j\mathrm{P}(N_i = j) & i \geq 2 \end{cases}, \qquad (3.5)$$

where $\mathrm{P}(N_i = j)$ denotes the probability of j patients in the system (including the one being served) right before the ith scheduled arrival. The recursive representation of $\mathrm{P}(N_i = j)$ is originally developed by Hassin and Mendel [25], for a case of scheduling homogenous patients with constant and identical no-show probabilities. In this problem, I relax the assumption of fixed no-

show probability by adding variable no-show probabilities for scheduled patients and fixed no-show probability for emergency patients. The modified representation is illustrated the following two equations:

When $j = 0$

$$\begin{aligned}
\mathrm{P}(N_i = 0) &= (1-p_{i-1})\sum_{k=0}^{i+1}\left[\mathrm{P}(N_{i-1}=k-1)\sum_{l=k}^{\infty}\mathrm{P}(N(x_{i-1})=l)\right] \\
&\quad + p_{i-1}\sum_{k=0}^{i}\left[\mathrm{P}(N_{i-1}=k)\sum_{l=k}^{\infty}\mathrm{P}(N(x_{i-1})=l)\right] \\
&= (1-p_{i-1})\sum_{k=0}^{i+1}\left[\mathrm{P}(N_{i-1}=k-1)\left(1-\sum_{l=0}^{k-1}e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{l}}{\theta^k l!}\right)\right] \\
&\quad + p_{i-1}\sum_{k=0}^{i}\left[\mathrm{P}(N_{i-1}=k)\left(1-\sum_{l=0}^{k-1}e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{l}}{\theta^k l!}\right)\right] \\
&= \sum_{k=0}^{i}\left[\mathrm{P}(N_{i-1}=k)\left(1-\sum_{l=0}^{k-1}e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{l}}{\theta^k l!}-(1-p_{i-1})e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{k}}{\theta^k k!}\right)\right]
\end{aligned}, \tag{3.6}$$

When $j > 0$

$$\begin{aligned}
\mathrm{P}(N_i = j) &= (1-p_{i-1})\sum_{k=0}^{i-j+1}\left[\mathrm{P}(N_{i-1}=j+k-1)\mathrm{P}(N(x_{i-1})=k)\right] \\
&\quad + p_{i-1}\sum_{k=0}^{i-j}\left[\mathrm{P}(N_{i-1}=j+k)\mathrm{P}(N(x_{i-1})=k)\right] \\
&= (1-p_{i-1})\sum_{k=0}^{i-j+1}\left[\mathrm{P}(N_{i-1}=j+k-1)e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{k}}{\theta^k k!}\right] \\
&\quad + p_{i-1}\sum_{k=0}^{i-j}\left[\mathrm{P}(N_{i-1}=j+k)e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{k}}{\theta^k k!}\right] \\
&= \sum_{k=1}^{i-j}\left[\mathrm{P}(N_{i-1}=j+k-1)e^{-\frac{x_{i-1}}{\theta}}\frac{x_{i-1}{}^{k-1}}{\theta^{k-1}(k-1)!}\left((1-p_{i-1})\frac{x_{i-1}}{\theta k}+p_{i-1}\right)\right] \\
&\quad + (1-p_{i-1})\mathrm{P}(N_{i-1}=j-1)e^{-\frac{x_{i-1}}{\theta}}
\end{aligned}. \tag{3.7}$$

5.2.3. Numerical results

Figure 5.1, Figure 5.2 and Figure 5.3 illustrate number of patients served n, system net revenue and overbook ratio, respectively, when capacity Q ranges from 2 to 10. In general, as Q increases, the growth of n and system net revenue is near linear, however, overbook rate fluctuate significantly. Due to computational constraints, I only explore a system with maximum capacity of $Q = 10$, as Q becomes greater, the fluctuation is expected to be smoother.

Figure 5.1 the threshold value for N under various capacities Q $\left(p_i = 0.2,\ r_i = \$1{,}500\ \forall i,\ c_I = \$800\right)$

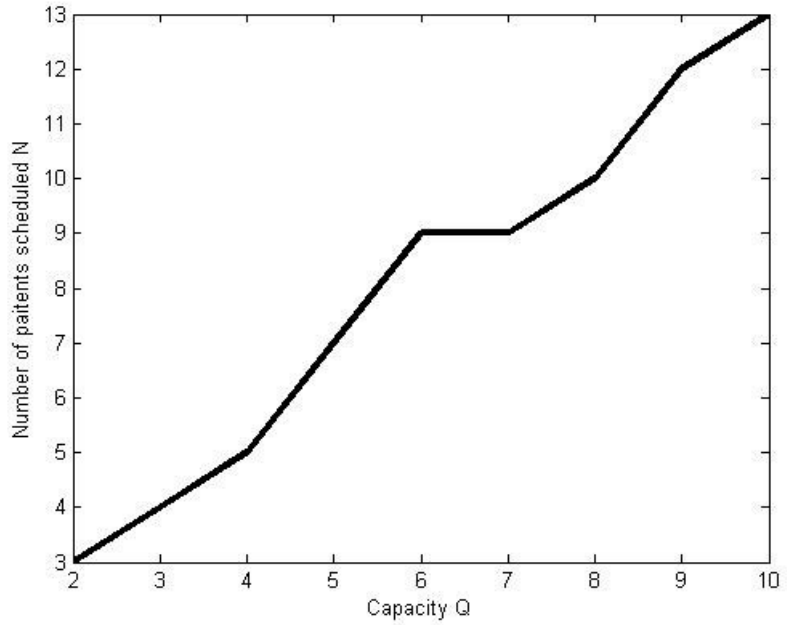


Figure 5.2 System net revenues under various capacities Q $\left(p_i = 0.2,\ r_i = \$1{,}500\ \forall i,\ c_I = \$800\right)$

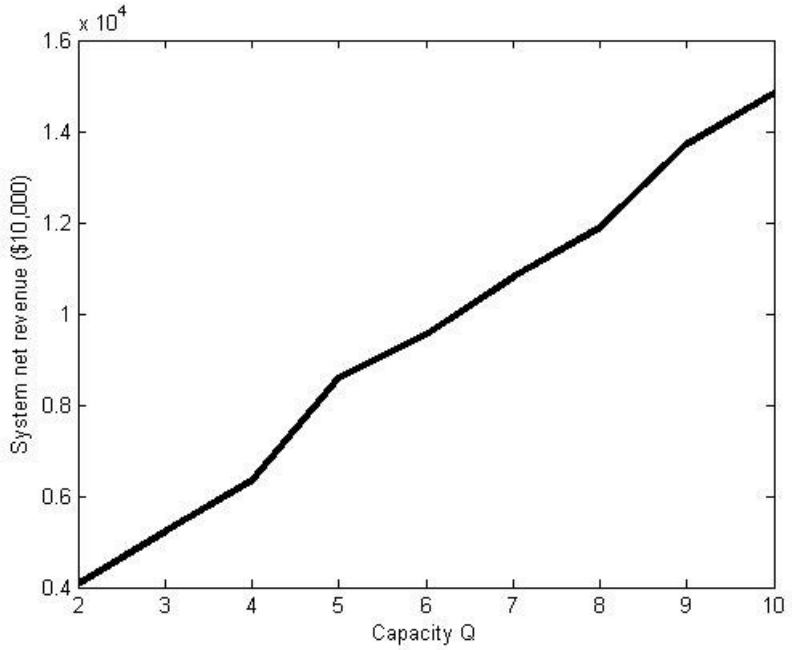
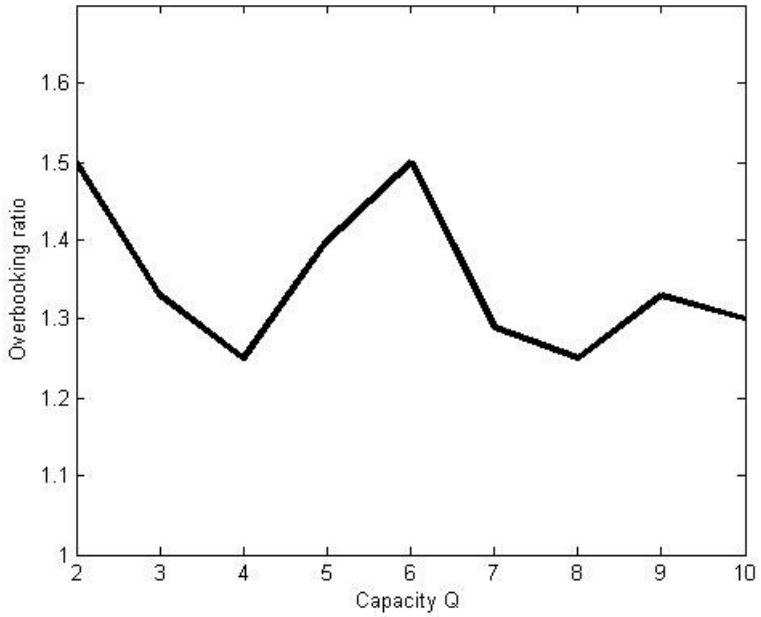


Figure 5.3 Overbook rate vs. capacity Q $\left(p_i = 0.2,\ r_i = \$1{,}500\ \forall i,\ c_I = \$800\right)$



Refer to Table 5.1 for detailed data on system net revenues, solutions, and threshold values for N at different levels of capacity Q.

Table 5.1 Threshold values, total net revenues, and schedules under various capacities Q $\left(p_i = 0.2,\ r_i = \$1{,}500\ \forall i,\ c_I = \$800\right)$

Q	N	Z	X											
			x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
2	3	\$4,051	0.00	0.01										
3	4	\$5,207	0.00	0.06	0.21									
4	5	\$6,346	0.00	0.81	0.30	0.27								
5	7	\$8,587	0.00	0.10	0.36	0.42	0.42	0.31						
6	9	\$9,532	0.00	0.11	0.38	0.45	0.46	0.44	0.31	0.00				
7	9	\$10,791	0.00	0.12	0.39	0.47	0.49	0.49	0.45	0.32				
8	10	\$11,883	0.00	0.12	0.40	0.48	0.51	0.52	0.51	0.46	0.32			
9	12	\$13,727	0.01	0.11	0.61	0.18	0.36	0.33	0.02	0.35	0.33	0.04	1.15	

10	13	$14,832	0.00	0.53	0.21	0.27	0.28	0.24	0.74	0.15	0.09	0.00	1.41	0.88

5.3. A base case of the all outpatient policy

In radiology practice, there are only two types of patients (outpatients and inpatients) can be scheduled, it's fairly reasonable to assume same attributes for same patient type. Patients of the same type are usually served with adjacent slots. A tactical solution of different waiting cost ratios to the model described in Section 5.2 may not be easily implemented in practice due to its complexity. Instead, I decide to test the model under several scheduling heuristics. In this section, I consider a scheduling policy which attempts to allocate all capacity to outpatient, with $r_i = r_O$, $p_i = p_O$, $c_{Wi} = c_{WO}$, and $c_{Pi} = c_{PO}$ $\forall i$, where r_O, p_O, c_{WO}, and c_{PO} denote service revenue, no-show probability, hourly waiting cost, and deny penalty cost for all outpatients, respectively.

Operations data used in this study was collected from the radiology department at University of Washington Medical Center (UWMC) [12], over a period of one month. The probability of outpatient no-show was originally estimated based on a sample of 1,130 MRI appointments, among which 96 appointments were not made due to one of the following reasons: "Patient did not show up", "Patient Discharged", "Rescheduled", and "Patient Cancelled". It seems to be fair to assume $p_i = 0.1$ $\forall i$, however, after comparing with no-show rates reported by various MRI related studies, I identified my estimate was on the low end, therefore, I decided to range it from 0.1 to 0.2 and use 0.2 as default value, in order to amplify the effects of overbooking. The approximation of emergency patient arrival probability $1 - p_e = 0.05$ was calculated from the monthly total number of emergency MRI scans relative to total MRI scans. Number of appointment slots per clinic session $Q = 9$ reflected a recently reduced-hour full operating

schedule for the day shift. Primarily determined by area of patient body part to be scanned and scan type (contrast or non-contrast), the length of a MRI scan can vary from 30 minutes to 90 minutes. Fixed appointment length θ = 0.75 approximates an average of 46 minutes per scan over the same sample of 678 MRI appointments.

Revenue and costs were estimated based on my literature survey on related public statistics. According to the Bureau of Statistics within US Department of Labor [65], the median and mean hourly wages of all occupations in Washington State in the year 2011 are \$19.3 and \$24.17, respectively. Hence, I used $c_{Wi} = \$20 \ \forall i$ as hourly patient waiting cost. "An average MRI machine costs approximately \$2 million to buy and install and \$800,000 per year to run" [52], thus I used c_I = \$800 as default value for hourly server idle cost and make it range from \$600 to \$1,000. The penalty cost of not denying a scheduled non-emergency patient involves cost of scheduling, potential staff overtime cost, and loss of goodwill, which is very hard to quantify. To overcome it, in my study, I use $c_{Pi} = \$200 \ \forall i$.

Table 5.2 Baseline model parameter values

Parameter	Value
Number of appointment slots per clinic session Q	9
Appointment slot length θ (in hour)	0.75
Outpatient revenue r_O	\$1,500
Outpatient hourly waiting cost c_{WO}	\$20
Hourly MRI server idle cost c_I	\$800
Penalty cost for rejecting a scheduled patient c_{PO}	\$200
Outpatient no-show probability p_O	0.2

I use Sequential Quadratic Programming (SQP) to conduct numerical search for optimal or near optimal combination of scheduled inter-arrivals for fixed N patients. Since it's obvious the objective function is uni-modal to N, the same numerical search is performed multiple times to

search for optimal number of patients to be scheduled. An example of this numerical search is presented in Figure 5.4, where there are 9 slots within a single clinic session. I start with a non-overbooking schedule where $N = 9$, and increase the value of N one by one, the system reaches highest objective function value at $N = 12$. Compared with the non-overbooking solution where $N = Q = 9$, the system performance was able to improve by 27.2%. Refer to [38] for a detailed evaluation of improvements brought by overbooking, under various parameter values. If I further look into optimal scheduled inter-arrival times for each fixed value of N (Figure 5.5), I observe that for $9 \leq N \leq 11$, the schedules follow the dome shape, with inter-arrival time increases among the first several patients, keeps relatively constant at a certain level thereafter, and then drops significantly between last two patients. The dome shaped schedule has been observed in many previous studies. When $N > 11$, the inter-arrival time appears to be very unpredictable, with multiple peaks, a surge among last several patients is observed. This observation implies the overbooking model tends to enter a probabilistically unstable state as N is greater than a threshold value (11 in this case).

Table 5.3 summarizes the numbers of patients served (n), optimal system costs (Z), and corresponding scheduled inter-arrival times (X), under different numbers of patients scheduled (N). When $N = 13$, only $n = 12$ patients are served, patient 13 is scheduled but denied, $x_{12} = 0.5$ reflects the initial value of starting point of the numerical search.

Figure 5.4 The numerical search for a problem with $Q = 9$ $\left(p_i = 0.2\ \forall i,\ c_I = \$800\right)$

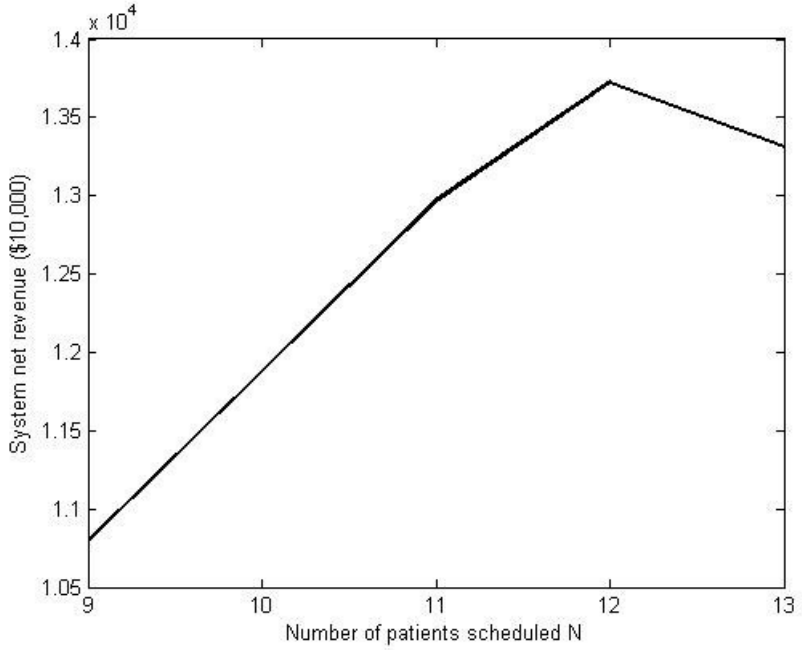


Figure 5.5 Schedules for $N = 9, 10, 11, 12, 13$ respectively $\left(Q = 9,\ p_i = 0.2\ \forall i,\ c_I = \$800\right)$

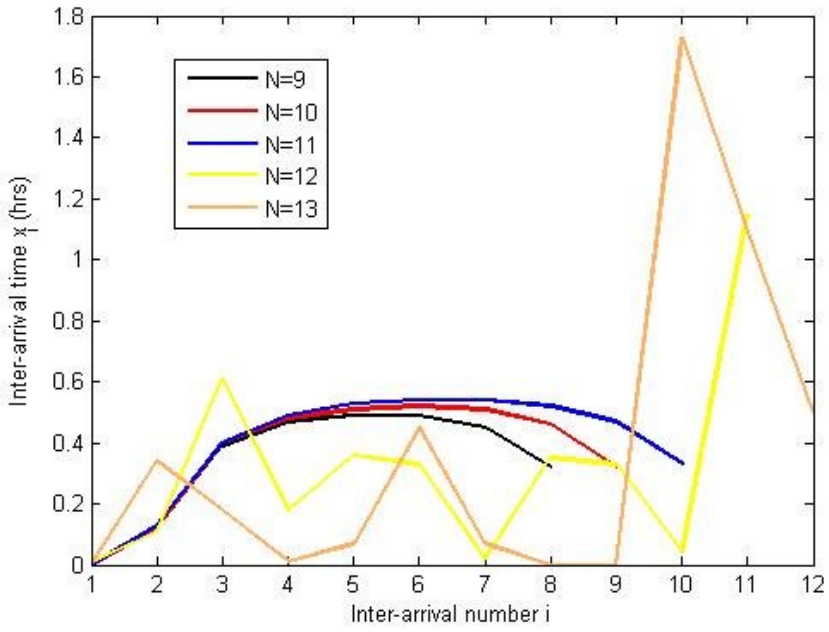


Table 5.3 Objective function values and schedules for $N = 9, 10, 11, 12, 13$ respectively $(Q = 9,\ p_i = 0.2\ \forall i,\ c_I = \$800)$

N	n	Z	X (hrs)											
			x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
9	9	\$10,791	0.00	0.12	0.39	0.47	0.49	0.49	0.45	0.32				
10	10	\$11,883	0.00	0.12	0.40	0.48	0.51	0.52	0.51	0.46	0.32			
11	11	\$12,970	0.00	0.13	0.40	0.49	0.53	0.54	0.54	0.52	0.47	0.33		
12	12	\$13,727	0.01	0.11	0.61	0.18	0.36	0.33	0.02	0.35	0.33	0.04	1.15	
13	12	\$13,312	0.00	0.34	0.18	0.01	0.07	0.45	0.07	0.00	0.00	1.73	1.09	0.50

From customer experience perspective (Figure 5.6), average of expected patient waiting times remains at a relatively low level when $9 \leq N \leq 11$, it increases dramatically at $N = 12$, where the system achieves highest objection function value. By adding one more patient to the schedule, each patient waits on average 38.6% longer, which indicates that it might not be an intelligent decision to gain 6% (growth from 12,970 to 13,727) overall revenue at the cost of a huge sacrifice on total patient waiting time. As we can see from Figure 5.7, for each single patient, when $9 \leq N \leq 11$, expected patient waiting time exhibits monotonically increasing trend following patient order; while $N > 11$, it decreases after reaching peaks at one of the last several patients. This observation is contradictory with the results by Fu and Storch [17], which studies a case without overbooking and demonstrates that expected patient waiting time is monotonic increasing with patient number.

Table 5.4 Expected patient waiting times for $N = 9, 10, 11, 12, 13$ respectively $\left(Q = 9,\ p_i = 0.2\ \forall i,\ c_I = \$800\right)$

N	n	$\bar{w}$	X												
			w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}
9	9	1.52	0.00	0.60	1.09	1.32	1.49	1.64	1.79	1.97	2.27				
10	10	1.56	0.00	0.60	1.09	1.32	1.47	1.61	1.74	1.87	2.04	2.34			
11	11	1.59	0.00	0.60	1.08	1.31	1.46	1.58	1.69	1.80	1.93	2.09	2.38		
12	12	2.21	0.00	0.59	1.09	1.14	1.57	1.83	2.12	2.70	2.95	3.23	3.79	3.26	
13	12	2.46	0.00	0.60	0.90	1.33	1.92	2.45	2.61	3.14	3.74	4.34	3.24	2.81	2.94

Figure 5.6 Average Expected patient waiting time under different values of *N*

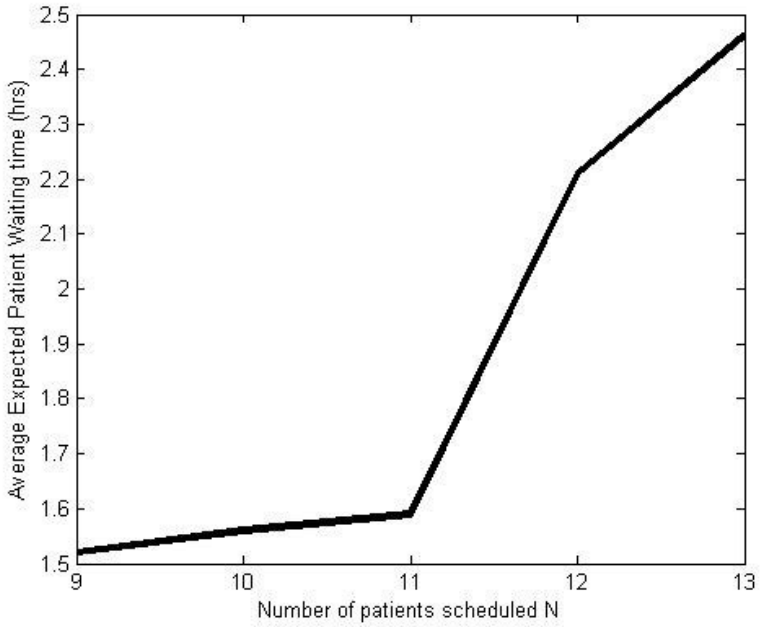
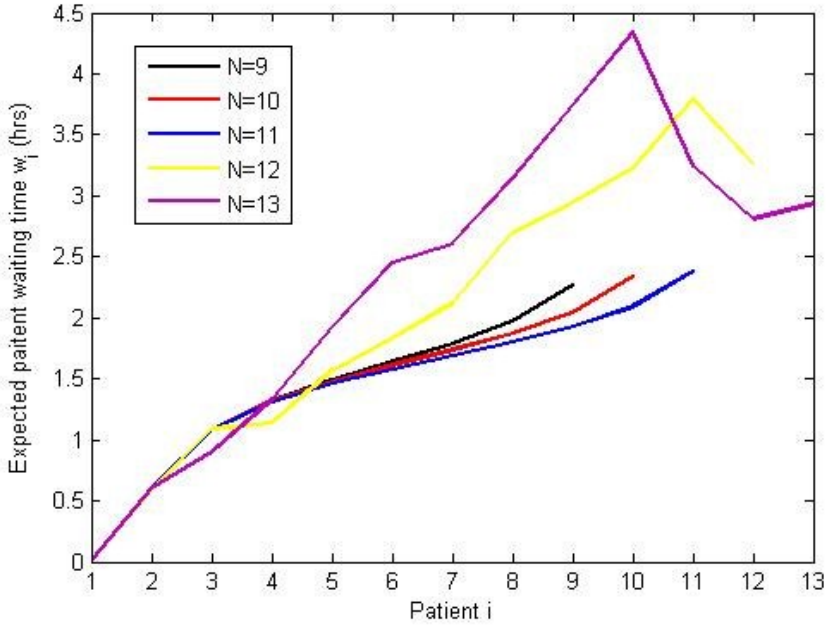


Figure 5.7 Expected patient waiting times under different values of *N*



5.4. Heuristic appointment policies

The numerical study on the based case of all outpatients helps us explore behaviors of my model, however, it restricts the potential of its application to practices due to the fact that all parameter values are fixed and inpatients are not considered. In this section, I focus on evaluating performance of several simple but commonly used booking policies, by enumerating three key parameters that I think may heavily affect system performance: outpatient no-show probability (p_O = 0.1, 0.15, 0.2), inpatient service revenue (r_I = 300, 500, 1000), and server hourly idle cost (c_I = 600, 800, 1000). Refer to Table 5.5 for a complete list of parameter values.

Table 5.5 Notation and parameter values of three heuristic booking policies

Parameter	Value
Number of appointment slots per clinic session Q	9
Appointment slot length θ (in hour)	0.75
Outpatient revenue r_O	\$1,500
Inpatient revenue r_I	\$300, \$500, \$1000
Outpatient hourly waiting cost c_{WO}	\$20
Inpatient hourly waiting cost c_{WI}	\$0
Hourly MRI server idle cost c_I	\$600, \$800, \$1000
Penalty cost for rejecting a scheduled patient c_{PO}	\$200
Penalty cost for rejecting a scheduled patient c_{PI}	\$1000
Outpatient no-show probability p_O	0.1, 0.15, 0.2
Inpatient no-show probability p_I	0

Three heuristic booking policies are considered in this section. The first one, which I name as "all outpatient", schedules only outpatients. It has been studied in Section 5.3 with fixed values of outpatient no-show probability and server hourly idle cost. It is a popular policy used by many clinics to maximize their revenues. The second heuristic booking policy is called "inpatient first", which attempts to serve inpatients in front of outpatients. Under this policy, first three patients scheduled are inpatients, and the rest scheduled patients are outpatients. The third policy is

named as “outpatient first”, which attempts to serve outpatients in front of inpatients. Under this policy, last three patients scheduled are inpatients. This policy mimics the clinic practice which serves outpatients during day shift and inpatient during night shift. Table 5.6 illustrates patient type assignment for the three policies.

Table 5.6 Patient type assignment of the three heuristic booking policies

Scheduled patient	1	2	3	4	…	N-3	N-2	N-1	N
All outpatient	OP	OP	OP	OP	OP	OP	OP	OP	OP
Inpatient first	IP	IP	IP	OP	OP	OP	OP	OP	OP
Outpatient first	OP	OP	OP	OP	OP	OP	IP	IP	IP

Note: EMP denotes emergency patient, OP denotes outpatient, and IP denotes inpatient

Table 5.7, Table 5.8, and Table 5.9 summarize system performances under various combinations of parameter values for all outpatient policy, inpatient first policy, and outpatient first policy, respectively. Since outpatient service fee is at a level significantly higher than inpatient service fee, the all outpatient policy always outperforms the other two policies in terms of system net revenue, under any combination of parameter values. Comparison between the rest two policies shows that the outpatient first policy wins when outpatient no-show probability is higher, the inpatient first policy wins when outpatient no-show probability is lower. This observation provides a general guidance on where to place inpatients based on outpatient no-show probability that is to schedule inpatients at the beginning if outpatient no-show rates are relatively low, and schedule inpatients at the end if the no-show rates are relatively high.

Table 5.7 System net revenue (in $1,000) of all outpatient policy

Outpatient	Server		
no-show probability p_O	hourly idle cost $c_I = 600$	hourly idle cost $c_I = 800$	Hourly idle cost $c_I = 1000$
0.1	14.7	14.91	15.05
0.15	13.71	13.78	13.84
0.2	13.79	13.73	13.88

Table 5.8 System net revenue (in $1,000) of outpatient first policy

Outpatient	Inpatient	Server		
no-show probability p_O	service fee r_I	hourly idle cost $c_I = 600$	hourly idle cost $c_I = 800$	Hourly idle cost $c_I = 1000$
0.1	300	10.49	10.5	10.62
	500	10.99	11.1	11.22
	1000	12.49	12.6	12.72
0.15	300	11.14	11.28	11.4
	500	11.74	11.88	12
	1000	13.21	13.38	13.5
0.2	300	10.6	10.71	10.84
	500	11.2	11.31	11.44
	1000	12.7	12.81	12.94

Table 5.9 System net revenue (in $1,000) of inpatient first policy

Outpatient	Inpatient	Server		
no-show probability p_O	service fee r_I	hourly idle cost $c_I = 600$	hourly idle cost $c_I = 800$	Hourly idle cost $c_I = 1000$
0.1	300	11.85	12.04	12.25
	500	12.45	12.64	12.85
	1000	13.95	14.15	14.35

	300	9.66	9.72	9.8
0.15	500	10.26	10.32	10.4
	1000	11.76	11.82	11.9
	300	10.16	10.26	10.34
0.2	500	10.76	11.01	11.1
	1000	12.44	12.51	12.6

One common phenomenon shared by the three policies is that system net revenue is greater in case server hourly idle cost is higher. Looking into corresponding schedules (Table 5.10, Table 5.11, and Table 5.12), I observe that higher server hourly idle cost causes overall shorter scheduled patient inter-arrival times, which shortens server idle time while elongates total patient waiting time. Considering higher patient waiting cost to server hourly idle cost, it clearly results into lower total cost. This phenomenon is more significant when outpatients have lower no-show probabilities.

For the two policies that take inpatients into consideration, system net revenue increases as inpatient service fee increases. From Table 5.11 and Table 5.12, I find that varying inpatient service fee doesn't impact outcome schedules, thus, higher inpatient fee linearly increases system net profit.

The all outpatient policy and inpatient first policy achieve best system performances when outpatient no-show probability is at the lowest level (0.1), while the outpatient first schedule reaches highest profit when p_O = 0.5. Even with the use of overbooking, increase of no-show probability cam reduces overall system net revenue.

Table 5.10 Schedules for all outpatient policy

Outpatient no-show probability	Server hourly idle cost ($)	x										
0.1	600	0.00	0.34	0.54	0.61	0.63	0.64	0.62	0.57	0.43		
0.1	800	0.00	0.21	0.46	0.53	0.58	0.60	0.57	0.56	0.51	0.39	
0.1	1,000	0.00	0.17	0.43	0.52	0.55	0.57	0.57	0.55	0.50	0.36	
0.15	600	0.00	0.30	0.44	0.55	0.57	0.14	1.00	0.48	0.50	0.44	
0.15	800	0.00	0.20	0.45	0.54	0.57	0.58	0.58	0.56	0.51	0.37	
0.15	1,000	0.00	0.16	0.42	0.51	0.55	0.56	0.56	0.54	0.49	0.36	
0.2	600	0.01	0.12	0.28	0.19	0.50	0.37	0.73	0.37	0.00	0.54	0.25
0.2	800	0.01	0.11	0.61	0.18	0.36	0.33	0.02	0.35	0.33	0.04	1.15
0.2	1,000	0.00	0.09	0.13	0.07	0.06	0.31	1.35	0.43	0.00	0.90	0.56

Table 5.11 Schedules for outpatient first policy

Outpatient no-show probability	Inpatient service fee ($)	Server hourly idle cost ($)	x									
0.1	300	600	0.00	0.31	0.50	0.55	0.53	0.41	0.00	0.00	0.00	
0.1	300	800	0.00	0.26	0.48	0.54	0.56	0.53	0.43	0.00	0.00	
0.1	300	1,000	0.00	0.22	0.45	0.52	0.53	0.51	0.41	0.00	0.00	
0.1	500	600	0.00	0.32	0.52	0.58	0.59	0.56	0.45	0.00	0.00	
0.1	500	800	0.00	0.26	0.48	0.54	0.56	0.53	0.43	0.00	0.00	
0.1	500	1,000	0.00	0.22	0.45	0.52	0.53	0.51	0.41	0.00	0.00	
0.1	1,000	600	0.00	0.32	0.52	0.58	0.59	0.56	0.45	0.00	0.00	
0.1	1,000	800	0.00	0.26	0.48	0.54	0.56	0.53	0.43	0.00	0.00	
0.1	1,000	1,000	0.00	0.22	0.45	0.52	0.53	0.51	0.41	0.00	0.00	
0.15	300	600	0.00	0.18	0.15	0.03	1.42	0.10	0.58	0.00	0.00	0.06
0.15	300	800	0.00	0.12	0.38	0.46	0.47	0.44	0.32	0.00	0.00	0.00
0.15	300	1,000	0.00	0.15	0.40	0.47	0.49	0.46	0.34	0.00	0.00	0.01
0.15	500	600	0.00	0.18	0.15	0.03	1.42	0.10	0.58	0.00	0.00	0.06
0.15	500	800	0.00	0.12	0.38	0.46	0.47	0.44	0.32	0.00	0.00	0.00
0.15	500	1,000	0.00	0.15	0.40	0.47	0.49	0.46	0.34	0.00	0.00	0.00
0.15	1,000	600	0.00	0.18	0.15	0.03	1.42	0.10	0.58	0.00	0.00	0.06
0.15	1,000	800	0.00	0.12	0.38	0.46	0.47	0.44	0.32	0.00	0.00	0.00
0.15	1,000	1,000	0.00	0.15	0.40	0.47	0.49	0.46	0.34	0.00	0.00	0.01
0.2	300	600	0.00	0.16	0.42	0.48	0.49	0.46	0.33	0.00	0.00	0.00
0.2	300	800	0.00	0.11	0.38	0.45	0.46	0.44	0.31	0.00	0.00	0.00
0.2	300	1,000	0.00	0.08	0.34	0.42	0.44	0.42	0.30	0.00	0.00	0.00
0.2	500	600	0.00	0.16	0.42	0.48	0.49	0.46	0.33	0.00	0.00	0.00
0.2	500	800	0.00	0.11	0.38	0.45	0.46	0.44	0.31	0.00	0.00	0.00
0.2	500	1,000	0.00	0.08	0.34	0.42	0.44	0.42	0.30	0.00	0.00	0.00
0.2	1,000	600	0.00	0.16	0.42	0.48	0.49	0.46	0.33	0.00	0.00	0.00
0.2	1,000	800	0.00	0.11	0.38	0.45	0.46	0.44	0.31	0.00	0.00	0.00
0.2	1,000	1,000	0.00	0.08	0.34	0.42	0.44	0.42	0.30	0.00	0.00	0.00

Table 5.12 Schedules for inpatient first policy

Outpatient no-show probability	Inpatient service fee ($)	Server hourly idle cost ($)	x									
0.1	300	600	0.03	0.03	0.56	0.58	0.48	0.68	0.49	0.52	0.71	0.47
0.1	300	800	0.00	0.00	0.55	0.51	0.54	0.55	0.55	0.52	0.62	0.49
0.1	300	1,000	0.00	0.00	0.49	0.48	0.52	0.53	0.53	0.50	0.61	0.47
0.1	500	600	0.03	0.03	0.56	0.58	0.48	0.68	0.49	0.52	0.71	0.47
0.1	500	800	0.00	0.00	0.55	0.51	0.54	0.55	0.55	0.52	0.62	0.49
0.1	500	1,000	0.00	0.00	0.49	0.48	0.52	0.53	0.53	0.50	0.61	0.47
0.1	1,000	600	0.03	0.03	0.56	0.58	0.48	0.68	0.49	0.52	0.71	0.47
0.1	1,000	800	0.00	0.00	0.55	0.51	0.54	0.55	0.55	0.52	0.62	0.49
0.1	1,000	1,000	0.00	0.00	0.49	0.48	0.52	0.53	0.53	0.50	0.61	0.47
0.15	300	600	0.00	0.00	0.62	0.53	0.55	0.55	0.54	0.48	0.50	
0.15	300	800	0.00	0.00	0.53	0.49	0.52	0.53	0.51	0.46	0.48	
0.15	300	1,000	0.00	0.00	0.46	0.46	0.50	0.50	0.49	0.44	0.47	
0.15	500	600	0.00	0.00	0.62	0.53	0.55	0.55	0.54	0.48	0.50	
0.15	500	800	0.00	0.00	0.53	0.49	0.52	0.53	0.51	0.46	0.48	
0.15	500	1,000	0.00	0.00	0.46	0.46	0.50	0.50	0.49	0.44	0.47	
0.15	1,000	600	0.00	0.00	0.62	0.53	0.55	0.55	0.54	0.48	0.50	
0.15	1,000	800	0.00	0.00	0.53	0.49	0.52	0.53	0.51	0.46	0.48	
0.15	1,000	1,000	0.00	0.00	0.46	0.46	0.50	0.50	0.49	0.44	0.47	
0.2	300	600	0.00	0.00	0.61	0.53	0.56	0.57	0.56	0.53	0.64	0.50
0.2	300	800	0.00	0.00	0.52	0.49	0.53	0.54	0.54	0.51	0.62	0.48
0.2	300	1,000	0.00	0.00	0.45	0.46	0.50	0.52	0.52	0.49	0.60	0.47
0.2	500	600	0.00	0.00	0.61	0.53	0.56	0.57	0.56	0.53	0.64	0.50
0.2	500	800	0.00	0.00	0.52	0.49	0.53	0.54	0.54	0.51	0.62	0.48
0.2	500	1,000	0.00	0.00	0.45	0.46	0.50	0.52	0.52	0.49	0.60	0.47
0.2	1,000	600	0.00	0.00	0.61	0.53	0.56	0.57	0.56	0.53	0.64	0.50
0.2	1,000	800	0.00	0.00	0.52	0.49	0.53	0.54	0.54	0.51	0.62	0.48
0.2	1,000	1,000	0.00	0.00	0.45	0.46	0.50	0.52	0.52	0.49	0.60	0.47

Chapter 6 **CONCLUSIONS**

This dissertation has presented a new perspective on the problems of scheduling arrivals of patients with no-show to queuing systems with exponential service times. As such, it is the first to take quadratic waiting cost, nonhomogeneous patients, and overbooking together into consideration in a static scheduling environment. This was intended to investigate the impact of no-show and to explore methods to alleviate the disruptive effects brought by no-shows, from a scheduling perspective.

My research was mainly motivated by three correlated streams of literature: i) a set of problems of scheduling arrivals to queuing systems, ii) general appointment scheduling problems with patient no-shows including overbooking studies, and iii) scheduling multiple categories of patients. As extensions of the first literature stream, I relaxed a series of key assumptions stepwise.

In Chapter 4, I first demonstrated that there is a need for nonlinear representation of patient waiting cost by waiting time, for the sake of which, I employed the concept of Taguchi's loss function to model it as a quadratic function of patient waiting time. I then relaxed the assumption of constant and identical patient no-show probabilities, and evaluated the model performance under three no-show probability based patient sequencing heuristics. I finally relaxed the assumption of constant and identical patient waiting cost ratios, and reevaluate the same set of heuristics. Major findings from this chapter are summarized as the following:

(1) The higher no-show first is the best among the three patient sequencing heuristics. In a case with nonhomogeneous no-show probabilities, it achieves the lowest total system cost, as well as lowest total expected patient waiting time. For the case with nonhomogeneous patient no-show probabilities and waiting cost ratios, it still outperforms the other two heuristics in terms of total system cost. Even though resulting in highest waiting time, it achieves the shortest server completion time. Therefore, I conclude that it is very robust against a wide range of parameter settings.

With its simplicity and effectiveness, the higher no-show first heuristic can be easily applied to practice. Several key assumptions need to be validated before its application:

- Patient waiting cost is evaluated with quadratic or quadratic like function of patient waiting time;
- Patient no-show probability is determined at the time of booking (based on certain attributes such as appointment delay, age, sex, marital status, and income);
- Patients receive same type of service, meaning the same expected service time for all scheduled patients.

It takes the following steps to implement this heuristic:

- Breakdown patient no-show probability into three to five adjacent intervals;
- Summarize operations data on no-show rates into a probability mass of the intervals;
- Apply total number of appointment slots per clinical session to the probability mass function to obtain the number of slots reserved for each interval;
- In case some interval is full of booked patients, the scheduler can have the flexibility overflow the next patient belonging to this interval to adjacent intervals.

(2) When patients are homogeneous with quadratic waiting costs, the inter-arrival times among the first several patients are more sensitive to no-show probability value than the inter-arrival times among the last several patients. Compared to a model with linear waiting costs, it tends to schedule patients closer when relative waiting cost ratio is above a threshold value, and vice versa, regardless the no-show probability value.

(3) The well-known dome shape for optimal schedules hold if patients are homogenous from three perspectives: non-deterministic independent and identical service times, constant and identical no-show probabilities, and constant and identical waiting cost ratios.

Chapter 5 further relaxed the assumption of scheduling a fixed number of patients, by applying a nonconventional overbooking strategy, which allows a certain level of server overtime, to a daily scheduling problem of allocating relatively flexible scanning capacities to inpatients, outpatients, and emergency patients. Instead of overbooking by specific appointment slot, I overbooked the entire clinic session. My conclusions are drawn based the following key assumptions: i) low patient hourly waiting cost ($20/hr for outpatients and $0/hr for inpatients) and high equipment idle cost (between $600/hr and $1,000/hr), ii) high outpatient service fee ($1,500/patient) and low inpatient service fee (between $300 and $1,000), and iii) low level of no-show probability (between 0.1 and 0.2 for outpatients, and 0 for inpatients). The following insights were obtained from this chapter:

(4) Schedule outpatients in front when outpatient no-show probabilities are relatively high (0.15 to 0.2); schedule inpatients first in case outpatient no-show probabilities are relatively low (0.1).

Implementation of this policy is straightforward. It requires that the estimated no-show rate of the target clinical session falls in a range between 0.1 and 0.2. Additionally, the number of slots reserved for inpatients needs to be determined with consideration of various factors such as demand, service revenue, and goodwill.

(5) The application of overbooking changes the behaviors of my static model from two perspectives:

- Arrival pattern. The solution schedules do not follow the well-known dome shape, instead, they tend to be very unpredictable;
- Patient waiting time. Expected patient waiting time w_i becomes non-monotonically increasing with patient order, which is contradictory to the results obtained from my generalize queuing model without overbooking.

(6) The optimal solution may not be the best choice in practice. It results in a huge sacrifice of patient waiting for very limited gain on net revenue improvement. The suboptimal solution in dome shape results in overall good and more balanced performance, thus, could be considered in radiology scheduling practices.

(7) System performance of the hybrid overbooking model is positively correlated with server hourly idle cost, and this relationship is more significant when outpatient no-show probabilities are relatively low.

Chapter 7 **FUTURE RESEARCH**

It would be very meaningful for future studies to extend my hybrid overbooking model to a generalized multi-objective analytical model which aims to find optimal number, sequence and inter-arrival times of scheduled patients with distinct no-show probabilities. One of my major contributions lies in the interesting findings of sequencing patients based on their no-show probabilities, the numerical studies were conducted based on several pre-defined heuristic booking policies, since it is not practical to compute optimal inter-arrival times for all possible permutations. Therefore, the biggest challenge for future research will be to take computational complexity into the consideration of developing the analytical model.

Another area for future study is to explore the hybrid overbooking model in a broader range of parameter settings. For example, a greater span of clinical session capacities might lead to different behaviors on net revenues and more stable overbooking ratios among different capacities. Throughout this dissertation, all the problems are coded in Matlab and solved by Sequential Quadratic Programming (SQP), the computation time grows exponentially as clinic session capacity increases, which limits my ability to explore the model with larger sizes of clinic capacities. A breakdown of the penalty cost for not being able to serve a scheduled patient may be important, since it is very difficult to estimate absolute monetary value for the penalty cost. Another potential way to address this issue is to use a penalty to service revenue ratio, which could be determined by healthcare decision administrators based on their sense or experience.

It may also be helpful to take a more in-depth look into the relationship between overbooking ratio and scheduled inter-arrival times. My findings in this dissertation suggest that the schedule may not follow the classic dome shape when overbooking ratio exceeds a threshold value. It would be very interesting for future research to explore variations of schedules at higher levels of overbooking ratios.

Finally, there is still a lack of emphasis on more realistic representation of appointment scheduling systems. Most analytical studies treat radiology imaging services as single phase processes, multi-phase models are desired in terms of depicting the work flows and the interactions of radiology facility with other functional department in the same hospital. From a patient arrival perspective, besides no-shows and walk-ins, it would be important for future studies to incorporate unpunctuality into arrival pattern. Even though the single server assumption holds well for many cases, as MRI equipment is generally very expensive to purchase, install and operate, there is a need for multi-server models if the future studies would like to focus the relatively inexpensive modalities (X-ray, CT, etc.).

BIBLIOGRAPHY

[1] A Alaeddini, K Yang, C Reddy, and S Yu. A probabilistic model for predicting the probability of no-show in hospital appointments. *Health Care Management Science*, 14(2):146-57, 2011.

[2] N Bailey. A study of queues and appointment systems in hospitals Outpatients departments with special reference to waiting times. *Journal of the Royal Statistic Society*, 14: 185–199, 1952.

[3] BBC News. GPs back "no show" fines. http://news.bbc.co.uk/1/hi/health/3187101.stm. 2003.

[4] A G Bean, and J Talaga. Predicting appointment breaking. *Journal of Health Care Marketing*, 15(1): 29–34, 1995.

[5] W Blanco, M J White, and M C Pike. Appointment Systems in Outpatients' Clinics and the Effect on Patients' Unpunctuality. *Medical Care*, 2: 133–145, 1964.

[6] P T Boggs, and J W Tolle. Sequential quadratic programming. *Acta Numerica*, 199–242, 1996.

[7] M Brahimi, and D J Worthington. Queuing Models for Out-patient Appointment Systems: A Case Study. *Journal of the Operational Research Society*, 42(9): 733–746, 1991.

[8] T Cayirli, and E Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4): 519–549, 2003.

[9] T Cayirli, K K Yang, and S A Quek. A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management*, 21(4): 682–697, 2012.

[10] Centers for Medicare and Medicaid Services (CMS). National Health Expenditure Data: Overview. *Historical*. 2010.

[11] Cisco Validated Design. Connected Imaging—Medical Image Architecture 2.0 Application Design Guide. 2009. http://www.cisco.com/en/US/docs/solutions/Verticals/Prfrmnagd.html#wp269187

[12] Department of Radiology Information Technology. Radiology Information System. University of Washington Medical Center, 2011

[13] T F Cox, J F Birchall, and H Wong. Optimizing the Queuing System for an Ear, Nose and Throat Outpatient Clinic. *Journal of Applied Statistics*, 12: 113–126, 1985.

[14] B Denton, D Gupta. Sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35: 1003–1016, 2003.

[15] B Denton, J Viapiano, and A Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, 10(1): 13-24, 2007.

[16] B Fries, and V Marathe. Determination of Optimal Variable-Sized Multiple-Block Appointment Systems. *Operations Research*, 29(2): 324–345, 1981.

[17] M Fu and R Storch. Static Appointment Scheduling for Patients with Nonhomogeneous No-show Probabilities and Waiting Cost Ratios. *IIE Transactions on Healthcare System Engineering*, under review.

[18] G Gallucci, W Swartz, and F Hackerman. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*, 56(3) 344 - 346, 2005.

[19] Y Gerchak, D Gupta, and M Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42: 321–334, 1996.

[20] Y Gocgun. *Approximate Dynamic Programming for Dynamic Stochastic*. Ph.D. dissertation, University of Washington, Washington, Seattle, United States, 2010.

[21] J Goldsmith. A Radical Prescription for Hospitals. *Harvard Business Review*, 67(3): 104–111, 1989.

[22] L V Green, and S Savin. Reducing delays for medical appointments: A queuing approach. *Operations Research*, 56(6): 1526-1538, 2008.

[23] L V Green, S Savin, and B Wang. Managing Patient Service in a Diagnostic Medical Facility. *Operations Research*, 54(1): 11-25, 2006.

[24] D Gupta, and B Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9): 800-819, 2008.

[25] R Hassin, and S Mendel. Scheduling Arrivals to Queues: A Single-Server Model with No-Shows. *Management Science*, 54(3): 565–572, 2008.

[26] A L Hixon, R W Chapman, and J Nuovo. Failure to keep clinic appointments: Implications for residency education and productivity. *Family Medicine*, 31(9): 627–630, 1999.

[27] C J Ho, and H S Lau. Minimizing Total Cost in Scheduling Outpatient Appointments. *Management Science*, 38(12): 1750–1764, 1992.

[28] F Huarng, and M H Lee. Using Simulation in Out-Patient Queues: A Case Study. *International Journal of Health Care Quality Assurance*, 9(6): 21–25, 1996.

[29] A Hutzschenreuter. *Queuing models for outpatient appointment scheduling*. M.Sc Thesis, University of Ulm, Ulm, Baden-Württemberg , Germany, 2005.

[30] Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. *National Academy Press*, Washington, D.C., 2001.

[31] B Jansson. Choosing a Good Appointment System: A Study of Queues of the Type (D/M/1). *Operations Research*, 14: 292–312, 1966.

[32] S Kim and R Giachetti. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans*, 36: 1211–1219, 2006.

[33] S Kim and I Horowitz. Scheduling hospital services: the efficacy of elective-surgery quotas. *Omega*, 30: 335–346, 2002.

[34] K J Klassen, and T R Rohleder. Scheduling Outpatient Appointments in a Dynamic Environment. *Journal of Operations Management*, 14(2): 83–101, 1996.

[35] R Kolisch and S Sickinger. Providing radiology health care services to stochastic demand of different customer classes. *OR Spectrum*, 30(2):375 - 395, 2008.

[36] G Koole, G Kaandorp. Optimal outpatient appointment scheduling. *Health Care Management Science*, 10: 217–229, 2007.

[37] L R LaGanga and S R Lawrence. An appointment overbooking model to improve client access and provider productivity. *POMS College of Service Operations Meeting*, 2007.

[38] L R Laganga and S R Lawrence. Clinic overbooking to improve patient access and provider productivity. *Decision Sciences*, 38: 251–276, 2007.

[39] A Lamb. Why advanced access is a retrograde step. *British Journal of General Practice*, 52(485): 1035, 2002.

[40] H Lau and A H Lau. A Fast Procedure for Computing the Total System Cost of an Appointment Schedule for Medical and Kindred Facilities. *IIE Transactions*, 32, 9, 833–839, 2000..

[41] C Liao, C D Pegden, and M. Rosenshine. Planning Timely Arrivals to a Stochastic Production or Service System. *IIE Transactions*, 25(5) 63–73, 1993.

[42] D V Lindley. The Theory of Queues with a Single Server. *Proceedings Cambridge Philosophy Society*, 48: 277–289, 1952.

[43] L Liu, and X Liu. Dynamic and Static Job Allocation for Multi-Server Systems. *IIE Transactions*, 30(9): 845–854, 1998.

[44] N Liu, S Ziya, and V G Kulkarni. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing and Service Operations Management*, 12(2): 347–364, 2010.

[45] S Mendel. *Scheduling arrivals to queues: A model with no-shows*. M.Sc. thesis, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel, 2006.

[46] A Mercer. A Queuing Problem in Which Arrival Times of The Customers are Scheduled. *Journal of the Royal Statistical Society Series B*, 22: 108–113, 1960.

[47] A Mercer. Queues with Scheduled Arrivals: A Correction, Simplification and Extension. *Journal of the Royal Statistical Society Series B*, 35(1): 104–116, 1973.

[48] S V Mondschein, and G Y Weintraub. Appointment policies in service operations: a critical analysis of the economic framework. *Production and Operations Management*, 12(2): 266–286, 2003.

[49] C G Moore, P Wilson-Witherspoon, and J C Probst. Time and money: Effects of no-shows at a family practice residency clinic. *Family Medicine*, 33(7): 522–527, 2001.

[50] K Muthuraman, and M Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40(9): 820-837, 2008.

[51] M Murray, and C Tantau. Same-day appointments: Exploding the access paradigm. *Family Practice Management*. 2000.

[52] PA Health Care Cost Containment Counci. The Growth in Diagnostic Imaging Utilization , No. 27, 2004.

[53] J Patrick, M L Puterman. Improving resource utilization for diagnostic services through flexible inpatient scheduling. *Journal of the Operations Research Society*. 58: 235–245, 2007.

[54] J Patrick, M L Puterman, and M Queyranne. Dynamic Multipriority Patient Scheduling for a Diagnostic Resource. *Operations Research*, 56(6): 1507–1525, 2008.

[55] C D Pegden, and M. Rosenshine. Scheduling Arrivals to Queues. *Computers & Operations Research*, 17(4): 343–348, 1990.

[56] V. Pesata, G Palliga, and A A Webb. A descriptive study of missed appointments: Families' perceptions of barriers to care. *Journal of Pediatric Health Care*, 13: 178–182, 1999.

[57] L W Robinson and R R Chen. Scheduling Doctors' Appointments: Optimal and Empirically-Based Heuristic Policies. Unpublished working paper. Johnson Graduate School of Management, Cornell University, Ithaca, New York, 2001.

[58] L W Robinson and R R Chen. Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management*, 13(1): 53-57, 2011.

[59] A Soriano. Comparison of Two Scheduling Systems. *Operations Research*, 14: 388–397, 1966.

[60] W Stein, and M Cote. Scheduling Arrivals to a Queue. *Computers and Operations Research*, 21(6): 607–614, 1994.

[61] D Strum, L Vargas, and J May. Surgical subspeciality block utilization and capacity planning: a minimal cost analysis model. *Anesthesiology*, 90: 1176–1185, 1999.

[62] B C Strunk, and P J Cunningham. Treading water: Americans' access to needed medical care, Report, 1997–2001. Center for Studying Health System Change, Washington, D.C., 2002.

[63] P J Tuso, K Murtishaw, and W Tadros. The easy access program: A way to reduce patient no-show rate, decrease add-ons to primary care schedules, and improve patient satisfaction. *Permanente Journal*, 3(3): 68–71, 1999.

[64] T. Ulmer, and C. Troxler. The economic cost of missed appointments and the open access system. 2006.

[65] United States Department of Labor, Occupational Employment Statistics. State Occupational Employment and Wage Estimates Washington, 2011 http://www.bls.gov/oes/current/oes_wa.htm#11-0000

[66] P M Vanden Bosch, D C Dietz, and J R Simeoni. Scheduling Client Arrivals to a Stochastic Service System. *Naval Research Logistics*, 46(3): 549–559, 1999.

[67] J Vissers, and J Wijngaard. The Outpatient Appointment System: Design of a Simulation Study. *European Journal of Operational Research*, 3(6): 459–463, 1979.

[68] S D Walter. A comparison of appointment schedules in a hospital radiology department. *British Journal of Preventive and Social Medicine*, 27:160–167, 1973.

[69] P P Wang. Static and Dynamic Scheduling of Customer Arrivals to a Single-Server System. *Naval Research Logistics*, 40: 345–360, 1993.

[70] P P Wang. Optimally Scheduling N Customer Arrival Times for a Single-Server System. *Computers & Operations Research*, 24(8): 703–716, 1997.

[71] P P Wang. Sequencing and Scheduling N Customers for a Stochastic Server. *European Journal of Operational Research*, 119(3): 729–738, 1999.

[72] E N Weiss. Models for Determining Estimated Start Times and Case Ordering in Hospital Operating Rooms. *IIE Transactions*, 22(2): 143–150, 1990.

[73] J D Welch, and N Bailey. Appointment Systems in Hospital Outpatient Departments. *The Lancet*, 1105– 1108, 1952.

[74] G C Xakellis, A Bennett. Improving clinic efficiency of a family medicine teaching clinic. *Family Medicine*, 33(7): 533–538, 2001.

[75] B Zeng, A Turkcan, J Lin, and M Lawley. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research*, 178(1): 121–144, 2010.

VITA

Mingang Fu received his B.S. degree in Industrial Engineering in 2007 from Huazhong University of Science and Technology in Wuhan CHINA. He received a M.S. degree in Engineering Management in 2008 from Duke University in Durham NC. He received another M.S. degree in Industrial & System Engineering in 2011 from the University of Washington in Seattle WA. Mingang Completed his Ph.D. studies in 2013 in the same department at UW. His research interests include Radiology Patient Scheduling, Clinic Overbooking, Queuing Problems, and Healthcare Policies.