APPOINTMENT SCHEDULING IN HEALTH CARE

Nan Liu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research (Operations Research).

Chapel Hill
2009

Approved by,

Vidyadhar G. Kulkarni, Co-Advisor

Serhan Ziya, Co-Advisor

Nilay T. Argon, Committee Member

Wendell G. Gilland, Committee Member

Haipeng Shen, Committee Member

ABSTRACT

NAN LIU: Appointment Scheduling in Health Care
(Under the direction of Professor Vidyadhar G. Kulkarni and Professor Serhan Ziya)

We propose two complementary approaches to the scheduling of outpatient appointments. The first approach is to dynamically assign appointment times depending on the continuously updated patient schedule. The second approach is to statically design the system by either limiting the appointment backlog or regulating the demand rate through controlling the panel size, i.e. the population receiving the medical service.

For the first approach - dynamic scheduling, we start with the assumption that patients come from a single class with homogeneous no-show and cancellation behaviors. We develop a Markov decision process model and propose easily implementable heuristic dynamic policies. In a simulation study that considers a "model clinic," which is created using data from practice, we find that the proposed heuristics outperform all the other benchmark policies, particularly when the patient load is high compared with the regular capacity. We then extend our model to consider the scheduling of patients from multiple classes. In this model, different classes of patients are assumed to have different probability distributions for their no-show and cancellation behaviors. As in the single-class case, we develop heuristic dynamic policies. Simulation results suggest that our proposed heuristics perform well when the regular capacity is small.

For the second approach - static design, we model the appointment backlog as a single-server queue where new appointments join the backlog from the back of the queue. Motivated by empirical findings, we assume that patients do not show up for their appointments with probabilities that increase with their waiting times before receiving service. We first study the model under the assumption of exponential service times. We characterize the optimal appointment backlog size and the optimal demand rate that maximize the system throughput and investigate how they change with other system parameters. Then we consider a special case where patients' no-show probabilities follow a specific parametric form. Under this special case, we obtain a simple closed-form expression for the optimal demand rate if we do not put a limit on the appointment backlog. Finally we conduct extensive numerical studies to investigate the situation where the service times are deterministic. The numerical results suggest that the insights generated in our analytical study by assuming exponential service times also hold for the situation with deterministic service times.

ACKNOWLEDGEMENTS

I am sincerely grateful to my advisors, Professor Vidyadhar G. Kulkarni and Professor Serhan Ziya, for introducing me to the area of stochastic modeling and for their direction and guidance over the past three years. Words are not enough to express my gratitude for their dedication and encouragement, which made my dissertation research a truly joyful and rewarding experience. They provided me with a self-motivated, deadline-free environment that allowed me to pursue research following my own interests while not under pressure. Their friendly offices always opened doors for me and welcomed discussions on problems of any kind. Besides knowledge, one important thing they taught and showed me is the persistence to strive for perfection, which inspired me to be a scholar and shall influence and benefit me for the rest of my life.

I would also like to thank my committee members, Professor Nilay T. Argon, Professor Wendell G. Gilland and Professor Haipeng Shen, for their inspiring suggestions on my research. In particular, I am thankful to Professor Argon for her courses on Discrete Event Simulation and Design and Control of Queueing Systems, and Professor Shen for his hands-on training on my statistics skills.

I am grateful to Professor Samuel S. Weir for his help in building cost and reward functions used in the simulation study, and providing us with the data which we used to estimate patient cancellation and no-show probabilities. I also thank Mr. Aaron Ratcliffe for his help in collecting and organizing the data.

Special thanks to Professor Jeannette Song from Duke University, Professor Jayashankar M. Swaminathan from UNC-Chapel Hill and Professor Paul H. Zipkin from Duke University for their enlightening courses on various topics, which greatly broadened my horizons in the area of stochastic modeling. I would also like to thank Professor E. Michael Foster, Professor Noah F. Gans from the University of Pennsylvania, Professor Eda Kemahlioglu-Ziya, Professor Gabor Pataki, and Professor David S. Rubin for their heartful help and priceless advice on my career development. In particular, Professor Rubin gave me many valuable suggestions on teaching.

Thanks to the faculty and staff of our lovely department, my sincere fellow graduate students and those who have helped me throughout the years. They made my experience in Chapel Hill a beautiful and unforgettable one.

Last, but not the last, I would like to thank my parents, Ruiwen Liu and Jianmin Gong, and my wife, Xiaohong Pan, for their love and support over the years and for their bearing of my neglect while I was busy with my research. Without them, this work would have been impossible.

This dissertation is partially supported by NSF grant CMMI-0620736.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1. Overview

As stated in the 2005 report by the National Academy of Engineering (NAE) and the Institute of Medicine (IOM), the $1.6 trillion health care sector of the United States “is now mired in deep crises related to safety, quality, cost and access that pose serious threats to the health and welfare of many Americans.” On the other hand, other major industry sectors, including manufacturing, transportation, warehousing and service-providing, have been relying on systems-engineering tools to achieve operational excellence in all aspects - raise revenues, reduce costs and gain customer satisfaction and maintain their loyalty. These engineering tools which can generate useful methodologies and yield valuable insights for the health care sector, however, are overlooked in the health care community. In order to build a better health care delivery system, NAE and IOM call for a new partnership between systems engineering and health care.

Operations Research (OR), the science of better, provides theoretical support for systems engineering. It can help build system-level models and predict unanticipated consequences, study the dynamics of complex systems, generate insights into among subsystems and processes, and understand and manage the tradeoffs among different parties. To better describe and understand the structure and dynamics of the complicated health care systems, NAE and IOM (2005) adapted a four-level model by Ferlie and Shortell (2001), which lays a framework for a systems approach to health care delivery. These four levels are (1) the individual patient, (2) the care team, including the professional care providers, e.g. physicians, nurses, surgeons etc., (3) the organization, e.g.

clinics and hospitals, and (4) the political and economic environment. OR tools can be used in all these four levels to help improve operational effectiveness and efficiency. Table 1.1 lists a collection of OR tools and presents the four-level model of Ferlie and Shortell (2001). An application of OR tools to solve problems in health care systems can fit into one or more cells in the table accordingly.

Table 1.1: OR Tools and the Four-level Model

	A Four-level Model (Ferlie and Shortell (2001))			
OR tools	Patient	Care Team	Organization	Environment
Dynamic Programming Techniques			✓	
Enterprise-management Tools				
Game-theory Tools				
Optimization Tools			✓	
Queueing Theory			✓	
Simulation			✓	

✓: Appointment Scheduling

This dissertation answers the call of the 2005 report by NAE and IOM. In particular, we focus on the design of (outpatient) appointment scheduling systems used in clinics/hospitals. This topic fits into the organization level of the four-level model and we use a number of OR techniques to model and tackle the problem (represented by the check mark ✓ in Table 1.1).

Appointment scheduling is used in health care systems to regulate demand (patient appointment requests) and supply (service capacity of health care providers). Appointment systems have two objectives: (i) provide a better service to patients by assigning them very short time window during which they are guaranteed to get a service, and (ii) protect the system from daily fluctuations in demand which can lead to an inefficient system with low utilization levels in some days and overloads in others. However, appointment schedules do not resolve all the uncertainties in daily demand. One important reason is that health care systems suffer from high no-show and cancellation rates. As reported in Gallucci et al. (2005), patient no-show rates can be as high as 12% even when patients are given same-day appointments. Our study reveals that this rate can attain an even higher level of 17.99%. (see Chapter 3 for more details.) High no-show and cancellation rates introduce additional layers of uncertainty and can cause severe inefficiencies if not dealt with appropriately. The objective of this dissertation is to develop methods for designing appointment scheduling systems in health care recognizing the possibility that patients can cancel their appointments or simply not show up for their appointments. The methods proposed here also apply to other service operations, e.g. repair shops and hair salons, where customers exhibit no-show and cancellation behaviors. Therefore, in this dissertation, we use “patients” and “customers” interchangeably.

Past research shows that the longer the *appointment delay*, defined as the time between the day a patient calls for an appointment and the day her appointment is scheduled for, the higher the chances that she will cancel or not show up (Gallucci et al. (2005)). This suggests an obvious way to to reduce no-shows and cancellations: give patients same-day appointments on the day they call for appointments. This is called Open Access (OA) (see Murray and Tantau (2000)). It is now a popular scheduling paradigm and a subject of active research and discussion. However, as we will discuss in Chapter 2, OA can not be a universal solution to solve appointment scheduling problems for all clinics. In particular, OA will not work well in clinics having small capacities but facing relatively large volumes of patient appointment requests. In this dissertation, we propose two complementary approaches to solve patient scheduling problems in health care, especially for those which can not implement OA. One approach is to dynamically schedule patients based on the current appointment schedule; the other one is to statically design the system by managing the appointment backlog or regulating the demand rate through controlling the panel size, i.e. the population receiving the medical service.

1.1.1 Dynamic Scheduling

The first approach we propose is the method of dynamic scheduling. Consider scheduling patients on a daily basis. There can be two extremes over the spectrum of scheduling policies. On one side it is the OA, which leads to a minimal no-show rate at the expense of frequent/severe daily overloads; on the other side it is a policy that smooths the workload among days so that daily overloads are kept at a minimum, which however causes clinics suffer from the unavoidable no-shows. This is the basic trade-off we will deal with here.

More specifically, the objective of this research is to develop dynamic methods that help assign an appointment date to each patient depending on the clinic’s appointment schedule at the time of the patient's call. In the base setup we consider, we assume that all patients behave identically and independently, i.e their no-show and cancellation behaviors are statistically the same. We first formulate the problem as a Markov decision process (MDP). We have chosen to use a model that makes it possible to estimate various parameters using data that are typically available for most clinics. The model takes the following as an input: the "reward" of serving each patient, the "cost" of providing a certain number of appointment slots each day (both parameters to be set by the clinic), and the cancellation and no-show probability distributions. The objective of the MDP model is to maximize the long-run average net "reward" for the clinic.

In theory, one can solve this MDP to optimality using one of the standard solution methods. However, this is not practical since the system state space is very large even for relatively small, toy systems. Thus, we propose heuristic methods instead. The heuristic methods are developed using a known technique that employs a single step of the policy improvement algorithm on a static (state-independent) policy. The static policy is ideally a "good" policy if not the optimal among the class of state-independent policies so that the dynamic policy to be obtained after the policy improvement is even better. We call this dynamic policy policy improvement heuristics.

We have obtained data from the Department of Family Medicine of the School of Medicine at the University of North Carolina at Chapel Hill (UNC). Using these data we populate a "model clinic." Then, we carry out an extensive simulation study to evaluate the system performances under our heuristics and other "intuitively" good policies in this "model clinic." The simulation results show that our proposed heuristic is overall the best and most robust policy among all polices compared.

We also extend our work to a more general setting where patients are categorized into different classes according to their no-show and cancellation behaviors. We formulate the problem as an MDP and develop heuristics by applying the same idea used in the base setup. To evaluate the performance of the proposed policy, we follow a similar procedure to that in the base setup. Since we do not have data that enable us to study the heterogeneity in patients' no-show and cancellation behaviors, we first simulate the data as if we collected them from practice. Then we use these data to create a "model clinic" as we did for the base case, and finally conduct a simulation study. The simulation results reveal that our proposed policy performs well when the regular daily capacity is small.

1.1.2 Static Design

The second approach we consider is to design the system statically. We model the appointment backlog as a single-server queue where new appointments join the backlog from the back of the queue. Motivated by empirical findings (see, e.g. Gallucci et al. (2005)), we assume that customers do not show up for their appointments with probabilities that increase with their appointment delays. To be more specific, each patient has a tolerance time and if her appointment delay exceeds her tolerance time, she will be a no-show. Thus patients' show-up probabilities depend on the backlog at the time they join the backlog queue and the service rate of the server. We assume that patients behave identically and independently. The objective is to maximize the system throughput rate, the rate at which patients indeed show up and get served.

We make two different assumptions for service times. We assume that they are deterministic or exponentially distributed. We further assume that customers who do show up, show up on time. Though deterministic service time assumption appears to be a more realistic assumption for an appointment scheduling system, this assumption does not lead to an analytically tractable model. Thus we will first assume that service times are exponentially distributed to prove theoretical results that provide us insights on the optimal design questions. Then we will numerically verify that these insights hold when service times are deterministic.

We consider the following two models which use different controls. In the first model, we assume that there is a limit on the backlog size which is under our control, say, appointment requests will not be accepted if the appointment backlog is too long. For this model, we characterize the optimal backlog size that maximizes the system throughput rate and investigate how the optimal backlog size changes with other system parameters.

In the second model, we control the arrival rate, i.e. demands. We show that the system throughput rate is either a strictly increasing or a unimodal function of the arrival rate. Thus there exists a unique optimal arrival rate which maximizes the system throughput rate. Furthermore, we consider a special case under which patients' tolerance times are exponentially distributed. In particular, if we assume no limits on the size of the backlog queue, we obtain a simple closed-form expression for the optimal arrival rate, using which the clinic can get a quick estimate of the corresponding optimal panel size.

1.2. Organization of the Dissertation

The dissertation is organized as follows. We consider the method of dynamic scheduling in Chapters 2, 3, and 4. In Chapter 2, we develop the MDP model for the dynamic scheduling of outpatient appointments and derive the policy improvement heuristics. In Chapter 3, we analyze the data obtained from the Department of Family Medicine at UNC, describe the simulation model and discuss the simulation results. In Chapter 4, we extend the MDP model developed in Chapter 2 to a more general one which considers the scheduling of multi-class patients. The simulation study and its results for this general model will also be discussed in the same chapter. We consider the static design of appointment scheduling systems in Chapters 5 and 6. We introduce the queueing model and prove the analytical results under the assumption of exponential service times in Chapter 5. In Chapter 6, we study a special case where customers' tolerance times are exponentially distributed and numerically investigate the model with deterministic service times. Finally, we provide our concluding remarks and discuss some of the future research directions in Chapter 7.

CHAPTER 2

Dynamic Scheduling of Outpatient Appointments: The Model

2.1. Introduction

Many firms strive to match demand and supply in the presence of uncertainty. Production systems deal with randomness in demand by keeping inventories. However, that is not an option for service systems since service capacity cannot be stored. Instead, service systems like patient clinics, repair shops, and hair salons, regulate demand through appointments. Appointment systems have two objectives: (i) provide a better service to customers by assigning them a very short time window during which they are guaranteed to get a service, and (ii) protect the system from daily fluctuations in demand which can lead to an inefficient system with low utilization levels in some days and overloads in others. However, appointment schedules do not resolve all the uncertainties in daily demand. Some service systems such as those in health care suffer from high no-show and cancellation rates, which introduce additional layers of uncertainty and can cause severe inefficiencies if not dealt with appropriately. The objective of this chapter is to develop a framework and introduce a method for the dynamic scheduling of patient appointments recognizing the possibility that patients can cancel their appointments or simply not show up for their appointments.

Past research shows that the longer the *appointment delay*, defined as the time between the day a patient requests an appointment and her actual appointment date, the higher the chances that she will cancel or not show up (Gallucci et al. (2005)). This suggests an obvious way of minimizing no-shows and cancellations: ask patients to come right away or make appointment requests on the day they want to be seen. This is called an *open access* (OA) or *advanced access* policy (see, e.g. Murray and Tantau (2000)) and of late it has become a popular paradigm in practice and the subject of active research. Several authors report on their experiences in implementing OA, both positive and negative. See Murray et al. (2003), Solberg et al. (2004), Belardi et al. (2004), and Dixon et al. (2006). Some practitioners strongly advocate OA (e.g., Murray and Tantau (2000)) while there are some who are strongly against it (e.g., Lamb (2002)). However, there seems to be an agreement on the fact that for OA to have at least a chance to work, demand and supply (capacity) need to be "in balance." Simply ensuring that average demand is less than supply is not sufficient for OA to work. Because of the stochastic nature of the daily demand, if average patient demand is not sufficiently low relative to the capacity, this will result in a high frequency of days in which daily demand exceeds the regular daily capacity. The clinic has to deal with this excess demand in some way (e.g., by working overtime, squeezing in additional appointments during the day, delegating some work to nurses) that will cause the clinic to incur overtime costs and/or reduce the quality of the service provided to the patients. Therefore OA is unlikely to be sustainable for a clinic observing demand levels that are close to its capacity. Although it is in general difficult to define exactly what it means for supply and demand to be "in balance," Green and Savin (2008) use a queueing model to provide some answers by developing a method to help determine the largest panel size sustainable for a physician using OA. They also find that panel sizes typically need to be much smaller than what is required for the queueing-theoretic stability of the system, i.e., long-run demand rate to be less than the long-run supply rate.

Even though the critics of OA have many concerns other than possible demand/supply imbalance, reported success stories strongly support the case in favor of OA since most clinics that implement OA report significant improvements in various measures of service quality when demand and supply are in balance. The problem, however, is that it appears that keeping the demand at a level that is necessary for an effective OA implementation may not be possible for many clinics and the number of such clinics may even increase in the future. Some of the recent articles published in academic journals and in the news media discuss the physician shortage problem that is currently being felt especially in some of the rural areas within the United States (e.g., upstate New York), and this is expected to grow significantly worse in the next 10-15 years unless action is taken to increase physician supply (see, e.g. Blumenthal (2004), Cauchon (2005), York (2007), Arvantes (2007)). One of the more problematic states is Massachusetts, which, interestingly, is the state with the highest doctors per capita according to U.S. Census Bureau (2008). Yet, according to the Massachusetts Medical Society (2007), the state has severe or critical shortages in several specialty areas including family practice and internal medicine. The society's reports from the last five years have already indicated increasing levels of physician shortage, but since the middle of 2007, patient demand has started to increase at an even faster rate as a result of the state's mandate on its residents to have health insurance, which has been effective since July 1, 2007. A New York Times article published in April 2008 reports that since the law took effect, about 340,000 of the approximately 600,000 uninsured have gained health coverage, and as a result clinics around the state have started to admit more patients by stretching their regular capacities (Sack (2008)). One of the family physicians the article mentions has a panel size of 3,000, which is well above the number suggested by Green and Savin for a potentially successful OA implementation. It is no surprise that patients of this particular physician have to wait for more than a year for a physical at the time of the article's publication.

These reports and articles do not discredit OA as an efficient solution for clinics that can keep their demand and supply in balance. However, they clearly indicate that a strict implementation of OA cannot be a universal solution to the appointment scheduling problem since some of the clinics will be overloaded out of obligation and some by choice. For such systems, scheduling some of the appointments to a not-so-distant future can relieve some of the stress of having to keep up with the demand on a daily basis while not causing no-show rates to increase in any significant way. Distributing the demand over several days has the obvious benefit of having a more regular daily load on the system, thereby reducing the possibility and/or the severity of daily overloads. However, distributing demand over days will have the unavoidable consequence of increasing no-shows and cancellations. This is the basic trade-off we are dealing with in this chapter. There are two policies at the two ends of the policy spectrum. On the one side is the OA, which leads to a minimal no-show rate at the expense of frequent/severe daily overloads; on the other side is a policy that schedules appointments so that daily overloads are kept at a minimum, which however causes clinics to suffer from unavoidable no-shows. A clinic's choice of a scheduling policy would depend on a variety of factors that determine its sensitivity towards no-shows, flexibility in adjusting its physician capacity, willingness to work overtime, and/or willingness to overbook and possibly cram in more patients depending on the daily load. For example, a physician working for his/her own private practice might not be bothered too much by working overtime and thus the overtime "cost" for such a clinic might be lower than what other clinics would typically face. Similarly, some clinics might prefer to see as many patients as possible and thus choose to overbook, while some others might refrain from that and choose to provide their patients with shorter waiting times and keep them more satisfied. The question is how to determine the scheduling policy in response to such diverse preferences. This research provides some answers to this question.

More specifically, the objective of this chapter is to develop dynamic methods that help assign an appointment date to each patient depending on the clinic's appointment schedule at the time of the patient's call. We first formulate the problem as a Markov decision process (MDP). We have chosen to use a model that makes it possible to estimate various parameters using data that are typically available for most clinics. (We will discuss how one can estimate model parameters using data from practice in Chapter 3.) The model takes the following as an input: the expected "net reward" of serving some x number of patients on a day with some z scheduled appointments at the beginning of the day, and the cancellation and no-show probability distributions. The objective of the MDP model is to maximize the long-run average net "reward" for the clinic.

In theory, one can solve this MDP to optimality using one of the standard solution methods. However, as we demonstrate in Section 2.3, that is not practical since the system state space is very large even for relatively small, toy systems. Thus, we propose heuristic methods instead. The heuristic methods are developed using a known technique that employs a single step of the policy improvement algorithm on a static (state-independent) policy. The static policy is ideally a "good" policy if not the optimal among the class of state-independent policies so that the dynamic policy to be obtained after the policy improvement is even better. Two static policies that we use are the OA policy and the two-day probabilistic scheduling policy. For either one of these two static policies, the policy improvement step gives an index policy that operates as follows: when an appointment request comes in, an index value is computed for each day in the scheduling horizon and the appointment is scheduled for the day with the highest index value. We show in Section 2.4 that the indices can easily be computed.

The remainder of this chapter is organized as follows. Section 2.2 reviews the relevant literature. In Section 2.3, we introduce our MDP model for appointment scheduling, and in Section 2.4 we describe the general form of the proposed heuristics. Section 2.5 gives detailed descriptions of the heuristic policies that we propose. In Section 2.6, we give more precise descriptions of the heuristic methods for two special cases and prove that they are optimal under certain conditions.

2.2. Literature Review

The Operations Research (OR) literature on outpatient appointment scheduling is extensive. A recent survey paper by Gupta and Denton (2008) discusses the main practical issues related to appointment scheduling, provides a review of the state of the art in modeling and optimization, and points to future directions. One classification that Gupta and Denton make regarding research on appointment scheduling is with respect to the type of waiting modeled: *direct* and *indirect*.

As Gupta and Denton indicate, most of the existing research has concentrated on *direct* waiting times, the times that patients spend waiting in the clinic on the day of their appointments from their arrival until their service. This work typically aims to minimize the expected "cost" for a day, which is a function of patients' direct waiting times, and the physician's idle time and overtime. In this body of work, typical decision variables are the number of appointment intervals, the length of each interval, and the number of patients assigned to each interval etc. We refer the reader to Cayirli and Veral (2003) for a comprehensive review of this literature pre-2003. More recent work in this line of research includes Denton and Gupta (2003), Robinson and Chen (2003), and Klassen and Rohleder (2004). These articles focus on the determination of appointment times for a sequence of punctual patients (jobs) with random service times, and their objective is to balance server idling, customer waiting and tardiness (overtime) costs. On the other hand, LaGanga and Lawrence (2007) and Muthuraman and Lawley (2008) study how to use overbooking to compensate for patient no-shows in an appointment system so as to improve the overall performance of the clinic. Our formulation and that of Muthuraman and Lawley (2008) are similar in that they both consider sequential scheduling of the patients as they call for appointments. However, while we are interested in determining the appointment day for each incoming patient, Muthuraman and Lawley are interested in determining in which particular time slot in a service session the appointment should be scheduled. Furthermore, Muthuraman and Lawley assume that using past data the clinic has identified the correlation between various patient attributes and their no-show probabilities and use these patient attributes in making scheduling decisions. In our formulation, the clinic does not differentiate among the patients in terms of their attributes, but unlike the model of Muthuraman and Lawley (2008), it does take into account the fact that no-show probabilities depend on the appointment delays.

Few articles deal with *indirect* waiting times, which correspond to appointment delays in this chapter, and which refer to the times between the days patients call for an appointment and the actual appointment dates. Gupta and Denton (2008) point to several difficulties of modeling indirect waiting, which might be part of the reason why work has been rather limited. Existing articles typically focus on the question of how many patients to admit or whether or not to admit a given patient for a particular day given the system state (i.e., current patient backlog). Patrick et al. (2008) study a dynamic multi-priority patient scheduling problem and develop cost-effective booking policies that meet waiting time targets for the patients. Gupta and Wang (2008) study a clinic capacity management problem using a model where patients' physician and time preferences are explicitly formulated, the decision is whether or not an appointment request should be accepted upon its arrival, and the objective is to maximize the revenue obtained on a given day. Our paper also belongs to this stream of research since we also deal with patients' indirect waiting times. However, unlike the articles above, our model takes into account the possibility that patients might cancel or not show-up for their appointments. As Gupta and Denton (2008) discuss, "indirect patient waiting" and "late cancelations and no-shows" largely remain as open research challenges, and in particular no prior work has explicitly studied appointment scheduling decisions in a model that relates overbooking decisions to no-show and cancellation rates. To the best of our knowledge, our paper is the first such work that proposes dynamic appointment scheduling policies using a model that explicitly takes into account patient no-shows and cancellations.

Several recent articles have investigated the open access policy. Kopach et al. (2007) use discrete event simulation to investigate the performance of OA under various settings. One key finding is that for clinics which predominantly use open access, offering provider care groups and overbooking appointments can help maintain the continuity of care provided to the patients, which is one of the important performance measures. Qu et al. (2007) identify the optimal percentage of appointment slots that a clinic should keep open within a session so as to maximize the throughput using a model where patients may not show up for their appointments. They also investigate the sensitivity of this optimal percentage to the provider capacity, patient no-show rates, and demand distributions. Green and Savin (2008) use a single-server queueing system to carry out capacity analysis for a clinic that uses the open access policy. In their model, each patient can be a no-show with a probability that depends on the patient's waiting time for the appointment. The authors provide a method to calculate the largest panel size that the clinic can handle; in a way, they "define" what it means for demand and supply to be in balance for a clinic using open access. Robinson and Chen (2008) compare the performance of OA with that of a traditional appointment scheduling system. Assuming deterministic service times and fixed and homogeneous patient no-show probabilities, the authors identify some of the structural properties of the optimal traditional scheduling policies and develop bounds for the system performance. Through numerical analysis, they find that in most cases - that is unless patient waiting times have marginal weights in the objective function and patient no-show rates are too small - OA is more preferable to traditional scheduling systems. In short, these four articles deal with the design of an OA system and comparison of OA with traditional scheduling policies. In this research, even though our main objective is not to investigate OA specifically, we also provide some support to some of the findings of this earlier work by identifying conditions under which OA performs reasonably well compared with the heuristic policies we propose.

Finally, a number of articles outside the OR literature investigate patient no-shows empirically. For example, see Oppenheim et al. (1979), Pesata et al. (1999), Moore et al. (2001), and Gallucci et al. (2005). All of these articles point to patient no-shows as being a significant problem in appointment scheduling and find that no-show rates depend on a variety of factors including race, gender, socioeconomic status etc. In particular, Gallucci et al. (2005) find that no-show and cancellation rates increase with the appointment delay. As we discuss in Section 3.1.1, we also find a similar relationship using data from the UNC clinic.

2.3. The Model

We consider a clinic where patients call to make appointments for a visit in the future or some time during the day of the call. Given the current appointment schedule, the administrative staff schedules each incoming request for an appropriate day and updates the schedule accordingly. We assume that patients do not have a strong preference for the date they want to be seen and thus accept the first appointment date offered by the staff. It can be a reasonable assumption for many clinics since patients generally do not want to be seen by physicians other than their own and as a result they are not inclined to go anywhere else as long as their situation does not constitute an emergency. In Section 7.2, we discuss how the heuristics we propose can be used in cases where patients do have time preferences and also the challenges associated with modeling patient preferences in general.

Let A^t denote the number of appointment requests that arrive on day t. We assume that $\{A^t, t = 1, 2, \dots\}$ is a sequence of independent and identically distributed (i.i.d.) random variables. As defined earlier, the *appointment delay* for a patient is the time between the day the patient requests an appointment and her actual appointment date. The appointment delay for a patient is zero if the appointment is scheduled on the same day the patient calls. The clinic uses a scheduling horizon of length T so that no patient has an appointment delay that is larger than T. Hence, a patient requesting an appointment on day t will be scheduled on one of the days $t, t+1, \dots, t+T$. For modeling convenience, we assume that all appointment requests are received at the beginning of the day so that all scheduling decisions on a given day t are made given the realization a^t of A^t. We shall see below that this assumption is needed to derive the heuristic policy, but is not needed when implementing the heuristic.

For any given day t, we define type (i, j) patients as those who called on day $t - i$, were given appointments for day $t + j$, and have not canceled their appointments by the beginning of day t. All patients of type (i, j) have an appointment delay of $i + j$ days and thus we must have $i + j \leq T$. Note that a given patient's type changes with time. For example, today's type (i, j) patient is tomorrow's type $(i + 1, j - 1)$ patient assuming she does not cancel.

Let X^t_{ij} denote the number of type (i, j) patients at the beginning of day t and define $\mathbf{X}^t = \{X^t_{ij} : 1 \leq i + j \leq T, i = 1, 2, \ldots, T\}$ as the vector of the number of patients of each type in the schedule at the beginning of day t. In the rest of the paper, we refer to $\mathbf{X}^t$ as the *backlog* or the *schedule* on day t, based on which the clinic schedules the incoming appointment requests.

There are three possible outcomes for each appointment made. The patient may show up for her appointment, cancel her appointment on or before the day of the appointment, or she may not cancel but simply not show up for her appointment. We assume that each patient's behavior is independent of that of the other patients and the arrival process $\{A^t, t = 1, 2, \ldots\}$. Past research (e.g., Gallucci et al. (2005)) as well as our own analysis clearly show that the longer the appointment delay for a patient, the higher the chances that she will cancel or that she will be a no-show if she does not cancel. In order to capture this relationship, we formulate the cancellation/no-show behavior as follows: We assume that each patient cancels her appointment at some random time in the future, which can possibly be beyond the patient's appointment day. We use T_c to denote the generic random variable that represents the time between the day a patient calls for an appointment and the day she decides to cancel it. A patient who has not canceled on or before her appointment day (which happens if T_c for this particular patient is larger than the patient's appointment delay) may or may not show up for her appointment. Let "S" and "NS" represent the "show" and "no-show" events for a particular patient, respectively and define

$$\alpha_{ij} \quad = \quad \mathbf{P}(T_c \geq i + j + 1, \mathrm{S} | T_c \geq i), \tag{2.1}$$

$$\beta_{ij} \quad = \quad \mathbf{P}(T_c \geq i+j | T_c \geq i). \tag{2.2}$$

Thus, α_{ij} is the probability that a patient who is currently of type (i, j) will show up for her appointment and β_{ij} is the probability that a patient who is currently of type (i, j) will not cancel her appointment before her appointment day. If there are n type (i, j) patients today, out of these n patients, the number of those who will not have canceled by the morning of their appointment day is a Binomial random variable with parameters n and β_{ij}. Similarly, the number of these patients who will show up for their appointment is a Binomial random variable with parameters n and α_{ij}.

We assume that events occur in the following order on each day. First, new patients call in and are given appointments. During the day, some patients cancel their appointments, and some do not show up for their appointments. At the end of the day, the clinic makes an "expected net reward" of $r(x, z)$ if z patients were scheduled on that day and the clinic ended up serving x of these patients during the day. For generality, we do not specify any particular form for $r(\cdot, \cdot)$ and thus each clinic can choose the function that best captures the circumstances it is in and its own valuations of different situations (i.e., the cost of reduced quality of service given to the patients if they are served in overtime slots). One special case of $r(\cdot, \cdot)$ is

$$r(x, z) = \eta(x) - w(z) \tag{2.3}$$

where $\eta(x)$ can be seen as the "expected reward" and is possibly linear or more generally concave in x and $w(z)$ is the total "expected cost" for a day if there are z patients scheduled at the beginning of the day. Here, the idea of making the expected cost a function of z as opposed to x makes it possible to capture the possibility that the clinic will plan according to the number of scheduled appointments (e.g., staffing cost). Any "cost" that depends on the number of patients who show up can be included in the $\eta(\cdot)$ function. These reward and cost functions can be estimated using models that were developed for scheduling appointments within a day. For example, Denton and Gupta (2003) developed a model to schedule patients over a day so as to minimize the total cost of patient waiting, staff idling and overtime. Their work provides a way to estimate the cost based on the number of scheduled appointments.

Note that the general form of the function $r(\cdot,\cdot)$ allows for more sophisticated choices than the special case given in (2.3). For instance, as reported by Moore et al. (2001), not all no-show slots are wasted since walk-in patients fill in some of these empty slots. According to the authors' study, only 12.2% of the slots are left unfilled by the end of the day, and on the average walk-in or triage patients help recover 89.5% of the costs of no-show patients. Thus, in some cases it might be reasonable to assume that the "reward" not only depends on x, the number of patients who show up, but also $z-x$, the number of no-show patients on a given day. This can be easily captured with our reward/cost formulation.

The objective of the clinic is to schedule arriving appointment requests so that the long-run average expected net reward is maximized. This problem can be modeled as an MDP, where the decision epochs are the times right after the appointment requests arrive every day and the system state at decision epoch t is given by $(A^t, \mathbf{X}^t)$. Let Y_j^t represent the number of patients who make their requests on day t and are given appointments for day $t+j$. (For example, Y_0^t is the number of patients who are given same-day appointments on day t.) Thus $\mathbf{Y}^t = \{Y_j^t : j = 0, 1, \ldots, T\}$ is the scheduling "action" taken on day t. Given that there are A^t appointment requests, the set of actions available on day t is $\{\mathbf{Y}^t : \sum_{i=0}^{T} Y_i^t = A^t, Y_i^t \in \mathbb{Z}_+, i = 0, 1, \ldots, T\}$ where $\mathbb{Z}_+$ represents the set of all nonnegative integers. Note that it is straightforward to include "rejection" of the appointment request as another action. However, to keep the presentation simpler, we assume that rejection is not an option. Later, we discuss how our heuristic policies would change if rejection were an option.

Let $B(n,p)$ represent a Binomial random variable with parameters n and p. Given $(A^t, \mathbf{X}^t)$ and the scheduling action $\mathbf{Y}^t$, $\mathbf{X}^{t+1}$ can be characterized as follows:

$$X_{ij}^{t+1} \overset{d}{=} \begin{cases} B(Y_{j+1}^t, \beta_{01}), & i = 1, j = 0, \ldots, T-1, \\ B(X_{i-1,j+1}^t, \beta_{i-1,1}), & i \geq 2, 0 \leq j \leq T-i, \end{cases} \tag{2.4}$$

where $\overset{d}{=}$ denotes the equality in distribution. Let

$$b(n,k,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

denote the probability mass function for a $B(n,p)$ random variable. Then, the transition probabilities given the action $\mathbf{Y}^t = \mathbf{y}^t$ can be expressed as follows:

$$\begin{aligned} &\mathbf{P}[(A^{t+1}, \mathbf{X}^{t+1}) = (a^{t+1}, \mathbf{x}^{t+1}) | (A^t, \mathbf{X}^t) = (a^t, \mathbf{x}^t), \mathbf{Y}^t = \mathbf{y}^t] \\ &= \mathbf{P}(A^{t+1} = a^{t+1}) \textstyle\prod_{i=1,\ldots,T, j=0,\ldots,T-i} P_{ij}(\mathbf{x}^{t+1}, \mathbf{x}^t, \mathbf{y}^t), \end{aligned}$$

where

$$P_{ij}(\mathbf{x}^{t+1}, \mathbf{x}^t, \mathbf{y}^t) = \begin{cases} b(y_{j+1}^t, x_{ij}^{t+1}, \beta_{01}), & i = 1, j = 0, \ldots, T-1, 0 \leq x_{ij}^{t+1} \leq y_{j+1}^t \\ b(x_{i-1,j+1}^t, x_{ij}^{t+1}, \beta_{i-1,1}), & i \geq 2, 0 \leq j \leq T-i, 0 \leq x_{ij}^{t+1} \leq x_{i-1,j+1}^t. \end{cases}$$

Let $U_i^t, i = 0, 1, \ldots, T$ denote the number of patients who call on day $t - i$ and show up for their appointments on day t. Then we have

$$U_i^t \overset{d}{=} \begin{cases} B(Y_0^t, \alpha_{00}), & i = 0, \\ B(X_{i0}^t, \alpha_{i0}), & i = 1, 2, \ldots, T. \end{cases}$$

We define $c_0((A^t, \mathbf{X}^t), \mathbf{Y}^t)$ to be the net reward obtained on day t given A^t, $\mathbf{X}^t$, and $\mathbf{Y}^t$. Then,

$$c_0((A^t, \mathbf{X}^t), \mathbf{Y}^t) = r(\sum_{i=0}^{T} U_i^t, Y_0^t + \sum_{i=1}^{T} X_{i0}^t).$$

Now, consider a scheduling policy f, and let $\phi_f(a, \mathbf{x})$ be the long-run expected average net reward under policy f given the initial state $A^1 = a$ and $\mathbf{X}^1 = \mathbf{x}$, i.e.

$$\phi_f(a, \mathbf{x}) = \lim_{k \to \infty} \frac{\mathbf{E}_f[\sum_{t=1}^{k} c_0((A^t, \mathbf{X}^t), \mathbf{Y}^t) | (A^1, \mathbf{X}^1) = (a, \mathbf{x})]}{k}.$$

A scheduling policy f^* is said to be optimal if

$$\phi_{f^*}(a, \mathbf{x}) = \sup_f \phi_f(a, \mathbf{x}), \quad \forall a, \mathbf{x}.$$

In theory, one can solve this MDP problem using one of the standard procedures such as the policy improvement or value iteration algorithms. However, the formulation suffers significantly from the curse of dimensionality. To see that, suppose that the maximum number of appointment requests that can possibly be received on a single day is $N < \infty$. Then one can show that the number of states for the MDP formulation we have given above would be $(N+1)\prod_{i=1}^{T}\sum_{k=0}^{N}\binom{k+i-1}{i-1}$. Note that even when $N = T = 5$, this number equals 1.34×10^9 and thus determining the optimal policy is not practically feasible for any realistically sized problem. Therefore it is of interest to develop heuristic scheduling methods that are efficient and perform well. We do this in the following section.

2.4. Policy Improvement Heuristics

In this section, we develop a heuristic dynamic appointment scheduling policy based on the idea of applying a single step of the policy improvement algorithm starting with a "good" initial policy. The idea is that since the policy improvement algorithm is generally believed to converge fast, applying a single step on an already "good" policy could give a policy that might be a reasonable substitute for the optimal dynamic policy. As we will see in this section, using a static policy as the initial policy makes it possible to fully characterize the policy obtained after the application of the policy improvement step so that the heuristic policy we propose will be easily implementable. In particular, it will not require the application of the policy improvement step for each instance of the appointment scheduling problem. This heuristic development technique has previously been used in such diverse areas as routing for parallel queues (see, e.g., Krishnan (1990), Opp et al. (2005) and Argon et al. (2009)) and dynamic kidney allocation (Zenios et al. (2000)), but to our knowledge, not within the context of appointment scheduling.

The procedure of developing policy improvement heuristics is as follows. Consider a state-independent policy that schedules each patient according to a stationary probability distribution $\mathbf{p} = (p_0, p_1, \ldots, p_T)$ where p_i is the probability that the patient is given an appointment for the ith day from today. We call this policy *Probabilistic Static Policy* (PSP). Consider a policy $\pi_{\mathbf{y}}$ that takes action $\mathbf{y} \equiv \{y_j\}_{j=0}^{T}$ at the beginning of today, i.e., schedules y_j patients on the jth day from today for $j \in \{0, 1, \ldots, T\}$, but from tomorrow on uses PSP. Define $\Delta((a, \mathbf{x}), \mathbf{y}, \mathbf{p})$ to be the difference in total expected net rewards over an infinitely long period of time by following policy $\pi_{\mathbf{y}}$ rather than using PSP all along given the initial state $(a, \mathbf{x})$. To conduct a one-step policy improvement, we maximize $\Delta((a, \mathbf{x}), \mathbf{y}, \mathbf{p})$ with respect to $\mathbf{y}$ for each state $(a, \mathbf{x})$, and the resulting optimal $\mathbf{y}$, denoted as $\mathbf{y}^*(a, \mathbf{x})$, specifies the heuristic scheduling policy for state $(a, \mathbf{x})$. We call this policy *Heuristic Dynamic Policy* (HDP). In the following, we first give an expression for $\Delta((a, \mathbf{x}), \mathbf{y}, \mathbf{p})$. Then, we use this expression to determine HDP. For a simple example of how to develop policy improvement heuristics, see Example 3.6.2 of Tijms (1994).

2.4.1 Improving the Probabilistic Static Policy

Since patients can be given appointments at most T days in advance, the action $\mathbf{y}$ taken today (say day 0) will only affect the schedule on days $0, 1, \ldots, T$. Thus $\pi_{\mathbf{y}}$ and PSP are guaranteed to give stochastically the same appointment schedule from day $T + 1$ onwards. In order to compute $\Delta((a, \mathbf{x}), \mathbf{y}, \mathbf{p})$, we only need to find the difference between the total expected net rewards for days $0, 1, \ldots, T$ under these two policies. Let $R_{\pi_{\mathbf{y}}}((a, \mathbf{x}), \mathbf{y}, \mathbf{p})$ and $R_{PSP}((a, \mathbf{x}), \mathbf{p})$ denote the total expected net reward accumulated over days $0, 1, \ldots, T$ under policy π_y and PSP, respectively, given the initial state $(a, \mathbf{x})$. Then,

$$\Delta((a, \mathbf{x}), \mathbf{y}, \mathbf{p}) = R_{\pi_{\mathbf{y}}}((a, \mathbf{x}), \mathbf{y}, \mathbf{p}) - R_{PSP}((a, \mathbf{x}), \mathbf{p}). \tag{2.5}$$

We first compute $R_{\pi_{\mathbf{y}}}((a, \mathbf{x}), \mathbf{y}, \mathbf{p})$. For $1 \le i \le T$ and $0 \le j \le T - i$, let $V_{ij}(\mathbf{x})$ denote the number of patients who called for appointments i days before day 0 and will not cancel by the morning of their appointment on day j, and $W_{ij}(\mathbf{x})$ denote the number of these patients who show up for their appointments. Similarly, for $0 \le j \le T$, let $\bar{V}_j(\mathbf{y})$ denote the number of patients who call for appointments on day 0, are scheduled for day j and will not cancel by the morning of their appointment, and $\bar{W}_j(\mathbf{y})$ denote the number of these patients who show up for their appointments. Finally, for $1 \leq k \leq T$ and $k \leq j \leq T$, we define $\hat{V}_{kj}(\mathbf{p})$ to be the number of patients who will call for appointments on day k and will not have canceled their appointment by the morning of their appointment on day j, and we define $\hat{W}_{kj}(\mathbf{p})$ to be the number of these patients who will show up for their appointments. Then, since the cancellation/no-show behaviors of the patients are independent of each other, we know that each one of these random variables has a Binomial distribution. More precisely,

$$V_{ij}(\mathbf{x}) \stackrel{d}{=} B(x_{ij}, \beta_{ij}), \bar{V}_j(\mathbf{y}) \stackrel{d}{=} B(y_j, \beta_{0j}), \hat{V}_{kj}(\mathbf{p}) \stackrel{d}{=} B(A_{(k)}, p_{j-k}\beta_{0,j-k}),$$

and

$$W_{ij}(\mathbf{x}) \stackrel{d}{=} B(x_{ij}, \alpha_{ij}), \bar{W}_j(\mathbf{y}) \stackrel{d}{=} B(y_j, \alpha_{0j}), \hat{W}_{kj}(\mathbf{p}) \stackrel{d}{=} B(A_{(k)}, p_{j-k}\alpha_{0,j-k}),$$

where $A_{(k)}$ denotes the number of appointment requests that will be made on day k, $1 \leq k \leq T$.

To see why the distribution of $\hat{V}_{kj}(\mathbf{p})$ and $\hat{W}_{kj}(\mathbf{p})$ are as given above, note that starting tomorrow, the policy $\pi_{\mathbf{y}}$ switches to implementing PSP, and therefore a patient requesting appointment on day k will be given an appointment on day j (where $k \leq j$) with probability p_{j-k}. From (4.1) and (4.2), this particular patient will not cancel the appointment by the morning of the appointment with probability $\beta_{0,j-k}$, and will show up for the appointment with probability $\alpha_{0,j-k}$.

We can then write

$$R_{\pi_{\mathbf{y}}}((a, \mathbf{x}), \mathbf{y}, \mathbf{p}) = \sum_{j=0}^{T} f_j(y_j, \mathbf{x}, \mathbf{p}),$$

where $f_j(y_j, \mathbf{x}, \mathbf{p})$ is the expected net reward accumulated on day j ($j = 0, 1, \ldots, T$), and is given by

$$f_j(y_j, \mathbf{x}, \mathbf{p}) = \mathbf{E}\left[r\left(\sum_{i=1}^{T-j} W_{ij}(\mathbf{x}) + \bar{W}_j(\mathbf{y}) + \sum_{k=1}^{j} \hat{W}_{kj}(\mathbf{p}), \sum_{i=1}^{T-j} V_{ij}(\mathbf{x}) + \bar{V}_j(\mathbf{y}) + \sum_{k=1}^{j} \hat{V}_{kj}(\mathbf{p}) \right) \right] \quad (2.6)$$

where we let $\sum_{k=1}^{j} \hat{W}_{kj}(\mathbf{p}) = \sum_{k=1}^{j} \hat{V}_{kj}(\mathbf{p}) = 0$ for $j = 0$.

Now, we need to determine $\mathbf{y}$ that maximizes (2.5) while ensuring that $\sum_{i=0}^{T} y_i = a$. But then since $R_{PSP}((a, \mathbf{x}), \mathbf{p})$ does not depend on $\mathbf{y}$, it can be dropped from the optimization problem altogether. Thus, HDP can be determined by solving the following resource allocation problem for given $(a, \mathbf{x})$ and $\mathbf{p}$:

$$\begin{aligned} \max \quad & R_{\pi_{\mathbf{y}}}((a, \mathbf{x}), \mathbf{y}, \mathbf{p}) = \textstyle\sum_{j=0}^{T} f_j(y_j, \mathbf{x}, \mathbf{p}) \\ \text{s.t.} \quad & \textstyle\sum_{j=0}^{T} y_j = a, \\ & y_j \in \mathbb{Z}_+, \quad j = 0, 1, \ldots, T. \end{aligned} \tag{2.7}$$

2.4.2 Description of the Heuristic Dynamic Policy

In its most general form, the heuristic policy we propose can be described as follows: first, pick a distribution $\mathbf{p}$, and if the system state is $(a, \mathbf{x})$ at the beginning of the day, solve Problem (2.7). Clearly, solving (2.7) is significantly simpler than solving the original MDP problem in general, but under certain conditions this is especially the case. In particular, we make the following assumption:

Assumption 1. The net reward function $r(x, z)$ is (i) increasing in x for fixed z and decreasing in z for fixed x, (ii) submodular, and (iii) jointly concave in x and z.

This assumption enforces reasonable conditions on the net reward function. In particular, if $r(\cdot, \cdot)$ is in the form of (2.3), Assumption 1 holds if the function $\eta(\cdot)$ is increasing and concave, capturing the (possibly) diminishing returns for additional patients and the function $w(\cdot)$ is increasing and convex, apprehending the (possibly) increasing marginal cost of each additional appointment.

As the following proposition states, Assumption 1 ensures that the objective function of Problem (2.7) is well-behaved. The proof is given in the appendix.

Proposition 1. Under Assumption 1, for given $\mathbf{x}$ and $\mathbf{p}$, the function $f_j(y_j, \mathbf{x}, \mathbf{p})$ is concave in y_j.

From Proposition 1 it follows that when Assumption 1 holds, Problem (2.7) is a resource allocation problem with a separable concave objective function, which has been well-studied in the literature. In particular, its optimal solution can be found by a simple algorithm given below (see, e.g. Section 4.2 of Ibaraki and Katoh (1988)). To avoid trivialities, we assume $a > 0$.

1. **Initialization**: Set $n := 1$ and $y_j := 0$, $j = 0, 1, \ldots, T$.

2. **Scheduling the nth patient of the day**: For each $j \in \{0, 1, \ldots, T\}$, compute

$$I_j = I_j(y_j, \mathbf{x}, \mathbf{p}) = f_j(y_j + 1, \mathbf{x}, \mathbf{p}) - f_j(y_j, \mathbf{x}, \mathbf{p}), \tag{2.8}$$

   and determine

$$j^* = \arg\max_{j \in 0,1,\ldots,T} I_j.$$

   Set

$$y_{j^*} := y_{j^*} + 1.$$

3. **Termination test**:

   - if $n < a$, let $n := n + 1$ and return to step 2;
   - otherwise, terminate the algorithm with $\mathbf{y}$ being the optimal solution.

At each iteration, the algorithm simply assigns a patient to the day j with the largest index value I_j, i.e., the day that will bring the largest improvement in the objective function of (2.7). Note that since patients are scheduled one by one and the index values I_j's do not depend on a, i.e., the total number of appointment requests on that day, we can relax our assumption that the clinic knows the total number of requests when scheduling patients. As new appointments are made, the clinic updates the appointment schedule, and then when a request comes in, the indices are calculated based on the updated information. Thus the HDP works as follows.

Heuristic Dynamic Policy

When an appointment request comes in, first for each day $j \in \{0, 1, 2, \ldots, T\}$ calculate the corresponding index I_j using (2.8), where y_j is the number of patients who called so far today and were scheduled for the jth day from today and $\mathbf{x}$ is the appointment schedule upon the arrival of this request. Then, determine $j^* = \arg\max_{j \in \{0,1,\ldots,T\}} I_j$, schedule the new appointment for day j^* and set $y_{j^*} = y_{j^*} + 1$.

If the clinic also had the option of rejecting appointment requests, the resulting policy would have an additional index for rejection, say the index I_R. It is then straightforward to show that this index would simply be equal to zero regardless of the system state. Therefore, with the rejection option available, the description of the Heuristic Dynamic Policy would need to be updated so that the appointment is rejected if all indices I_j where $j \in \{0, 1, 2, \ldots, T\}$ are negative. Otherwise, the appointment is scheduled on day j^* as described above.

2.5. Picking the Distribution $\mathbf{p}$ for HDPs

In Section 2.4.2, we described the heuristic policy that we propose for a given distribution $\mathbf{p}$, but we did not specify how this distribution should be picked. One can come up with different heuristic policies by choosing different distributions, but here we describe and propose two particular choices, which are both easy to determine and yield easily implementable HDPs.

Since the policy improvement heuristic is guaranteed to improve upon the initial policy, ideally, one would like to pick the optimal $\mathbf{p}$, i.e., the one that maximizes the long-run average reward among all static policies. Although there is no guarantee that a better static policy will lead to a better dynamic policy, as long as there is no reason to believe otherwise, it appears to be the most reasonable choice. However, finding the optimal $\mathbf{p}$ is a significant challenge by itself, requiring the solution of a maximization problem with $T + 1$ variables, and with an objective function that becomes increasingly more difficult to calculate as T increases. Therefore, in this chapter, we propose two simple alternatives.

Policy I: Open Access Policy (OAP): Open access policy is a static policy where $p_0 = 1$ and $p_1 = p_2 = \cdots = p_T = 0$.

Policy II: Optimal Two-day Probabilistic Static Policy (OTPSP): This is the optimal probabilistic static policy with the restriction that $p_2 = p_3 = \cdots = p_T = 0$ (p_0 and p_1 are picked optimally).

OAP is readily available, but finding OTPSP needs some explanation. We first derive the long-run average net reward under a general two-day probabilistic static policy (TPSP) as a function of p_0 (since $p_1 = 1 - p_0$), which is to be maximized with respect to p_0 to determine OTPSP. The TPSP is state-independent. Thus we pick a random day and derive an expression for the expected net reward for that day. Now, since $p_i = 0$ for $i \geq 2$, patients seen on a given day either made appointment on that day or the day before. Let $\tilde{A}_0$ and $\tilde{A}_1$ denote the number of appointment requests that arrived today and yesterday, respectively. Let $\tilde{V}_0(p_0)$ and $\tilde{V}_1(p_0)$ denote the number of patients who called in today and yesterday, respectively, to make appointments for today have not canceled their appointments by the morning of today, and $\tilde{W}_0(p_0)$ and $\tilde{W}_1(p_0)$ respectively denote the number of these patients who will show up for their appointments. Then,

$$\tilde{V}_0(p_0) \overset{d}{=} B(\tilde{A}_0, p_0), \tilde{V}_1(p_0) \overset{d}{=} B(\tilde{A}_1, (1 - p_0)\beta_{01}),$$

and

$$\tilde{W}_0(p_0) \overset{d}{=} B(\tilde{A}_0, p_0\alpha_{00}), \tilde{W}_1(p_0) \overset{d}{=} B(\tilde{A}_1, (1 - p_0)\alpha_{01}).$$

It follows that the long-run average net reward under TPSP with a same-day scheduling probability p_0 can be written as

$$R_{TPSP}(p_0) = \mathbf{E}\left[r\left(\tilde{W}_0(p_0) + \tilde{W}_1(p_0), \tilde{V}_0(p_0) + \tilde{V}_1(p_0)\right)\right], \tag{2.9}$$

and the OTPSP is obtained by solving the following optimization problem

$$\max_{0 \leq p_0 \leq 1} R_{TPSP}(p_0), \tag{2.10}$$

where $R_{TPSP}(p_0)$ is given in (2.9). This is an optimization problem with a single decision variable, and thus determining the optimal solution is relatively straightforward given the probability distribution for the daily appointment requests, $\tilde{A}_0$ and $\tilde{A}_1$. Furthermore, under certain conditions on this distribution and the reward/cost parameters, the objective function is well-behaved making the determination of the optimal solution to the problem easier. (See the appendix for the proof.)

Proposition 2. Suppose that the number of appointment requests that arrive on a single day has a Poisson distribution and the reward function $r(\cdot,\cdot)$ is as given in (2.3). Then, under Assumption 1, the objective function of Problem (2.10) is concave.

Although there is no strong evidence in support of the number of daily appointment requests having a Poisson distribution, it is a common assumption in the literature. As argued in Robinson and Chen (2008), if the panel size of a clinic is N and each patient independently makes an appointment for any given day with a small probability p, then the total number of appointment requests arriving on any given day has a Binomial distribution with parameters N and p, which converges to a Poisson distribution (with mean Np) as N gets large. For the heuristics we propose, although Poisson assumption is by no means necessary, it is nevertheless convenient.

Based on these two static policies, i.e., OAP and OTPSP, we propose two dynamic policies:

Policy III: Improved Open Access Policy (Imp-OAP): This policy works as described in Section 2.4.2 with distribution $\mathbf{p}$ picked as specified for OAP, i.e., $p_0 = 1$ and $p_1 = p_2 = \cdots = p_T = 0$.

Policy IV: Improved Optimal Two-day Probabilistic Static Policy (Imp-OTPSP): This policy works as described in Section 2.4.2 with distribution $\mathbf{p}$ picked as in the description of OTPSP, i.e. $\mathbf{p} = [p_0, 1-p_0, 0, \ldots, 0]$ where p_0 is the optimal solution to (2.10).

2.6. Two Examples and the Optimality of Imp-OAP

To give the reader a better idea about how the policies OTPSP, Imp-OTPSP and Imp-OAP look, we study two simple examples that lead to indices that can be expressed explicitly. For both examples, we assume that the function $r(\cdot,\cdot)$ is in the form of (2.3) with $\eta(x) = \tau x$ where τ is a positive constant, $\mathbf{E}[A^t] = \mu$ and $\mathbf{E}[(A^t)^2] = \xi$ for $t = 1, 2, \ldots$. In the first example, the function $w(z)$ is increasing linearly with z, which would be a reasonable assumption in cases where the clinic does not have to deal with very high patient loads. In the second example, the function $w(z)$ is assumed to be quadratic, which means the "marginal cost" of a patient is increasing with each additional scheduled patient, which is reasonable in cases where the clinic is relatively understaffed so that additional patients bring increasingly more burden on the clinic.

2.6.1 Linear Rewards and Costs

Suppose that $w(z) = \nu_1 z$ where $\nu_1 \geq 0$ is the cost per scheduled appointment at the beginning of a day. Then, the optimal p_0 (denoted as p_0^*) which solves problem (2.10) is given by

$$p_0^* = \begin{cases} 0 & \text{if } \tau(\alpha_{00} - \alpha_{01}) \leq \nu_1(1 - \beta_{01}), \\ 1 & \text{otherwise.} \end{cases} \tag{2.11}$$

We can then show that for Imp-OTPSP, the indices (2.8) simplify to

$$I_j = I_j(y_j, \mathbf{x}, \mathbf{p}) = \tau\alpha_{0j} - \nu_1\beta_{0j}, \tag{2.12}$$

which corresponds to the expected net reward of scheduling one more patient for day j. Notice that in this case the index I_j does not depend on $\mathbf{x}$ or y_j, and thus the HDP becomes a deterministic static policy which schedules all appointment requests received today for the j^*th day from today where $j^* = \arg\max_{j\in 0,1,\ldots,T} I_j$. For this example, Imp-OAP is the same as Imp-OTPSP, and as we prove in Theorem 1, they are in fact optimal. (See the appendix for the proof.)

Theorem 1. *If the function $\eta(\cdot)$ and $w(\cdot)$ are both linear, then the dynamic policy Imp-OAP, described by the indices given by* (2.12)*, is optimal among all policies.*

2.6.2 Linear Rewards and Quadratic Costs

Suppose that $w(z) = \nu_2 z^2$ where $\nu_2 \geq 0$. Define $\kappa_0 = \mu(\tau\alpha_{01} - \nu_2\beta_{01}) - \nu_2(\xi - \mu)\beta_{01}^2$, $\kappa_1 = \mu[\tau(\alpha_{00} - \alpha_{01}) - \nu_2(1 - \beta_{01})] + \nu_2[2(\xi - \mu)\beta_{01}^2 - 2\mu^2\beta_{01}]$, and $\kappa_2 = \nu_2[2\mu^2\beta_{01} - (\xi - \mu)(1 + \beta_{01}^2)]$. Then, we can show that

$$R_{TPSP}(p_0) = \kappa_2 p_0^2 + \kappa_1 p_0 + \kappa_0. \tag{2.13}$$

Let p_0^* be the maximizer of (2.13) and $\mathbf{p}^*$ be a $1 \times (T+1)$ vector defined as $\mathbf{p}^* = [p_0^*, 1-p_0^*, 0, 0, \ldots]$. Then, index $I_j = I_j(y_j, \mathbf{x}, \mathbf{p}^*)$ for Imp-OTPSP can be shown to be

$$I_j = \begin{cases} \tau\alpha_{00} - \nu_2(1 + 2y_0 + 2\sum_{i=1}^{T} x_{i0}\alpha_{i0}) & \text{if } j = 0, \\ \tau\alpha_{01} - \nu_2\beta_{01}[1 + 2\beta_{01}y_1 + 2(\sum_{i=1}^{T-1} x_{i1}\alpha_{i1} + \mu p_0^*)] & \text{if } j = 1, \\ \tau\alpha_{0j} - \nu_2\beta_{0j}[1 + 2\beta_{0j}y_j + 2(\sum_{i=1}^{T-j} x_{ij}\alpha_{ij} + \mu((1 - p_0^*)\beta_{01} + p_0^*))] & \text{if } j = 2, 3, \ldots, T. \end{cases} \tag{2.14}$$

Note that I_j decreases with y_j and x_{ij}. Thus under Imp-OTPSP, the more patients scheduled in one day, the smaller the chances that an additional patient will be scheduled on the same day, as we would expect for an intuitively "good" policy. This is markedly different from the linear-cost example since when costs are linear the marginal cost of an appointment is fixed while in the quadratic-cost case, the marginal cost of an additional appointment for a given day increases with the total number of appointments already scheduled for that day.

Similarly, we can show that the index I_j for Imp-OAP can be shown to be

$$I_j = \begin{cases} \tau\alpha_{00} - \nu_2\beta_{00}(1 + 2\beta_{00}y_0 + 2\sum_{i=1}^{T} x_{i0}\alpha_{i0}) & \text{if } j = 0, \\ \tau\alpha_{0j} - \nu_2\beta_{0j}[1 + 2\beta_{0j}y_j + 2(\sum_{i=1}^{T-j} x_{ij}\alpha_{ij} + \mu)] & \text{if } j = 1, 2, \ldots, T, \end{cases} \tag{2.15}$$

which is also decreasing in y_j and x_{ij} and thus carries similar characteristics as the index for Imp-OTPSP.

CHAPTER 3

Dynamic Scheduling of Outpatient Appointments: A Simulation Study

In this chapter we evaluate the performances of the heuristic policies through an extensive simulation study, by comparing them with other benchmark policies. These benchmark polices either mimic those used in practice or are sensibly good policies under some situations if not all. The "model clinic" used in the simulation study is created by using data provided to us by the Department of Family Medicine at the University of North Carolina (UNC). These data help us estimate the no-show and cancellation probability distributions. The simulation results indicate that the proposed heuristics significantly outperform other benchmark heuristics especially when the system is highly loaded. As expected, the OA policy performs reasonably well when average demand is below daily regular capacity but it performs very poorly under high demand.

This chapter is organized as follows. In Section 3.1, we describe our "model clinic," which we used in our simulation study and discuss how we estimated various model parameters. We introduce the benchmark polices, outline the simulation setup and report the results in Section 3.2.

3.1. Estimating Model Parameters

In this section, we discuss how we estimated the parameters for the "model clinic" - which we used for our simulation study described in the next section - using data from an actual clinic. The data were obtained from the Department of Family Medicine of the School of Medicine at the University of North Carolina at Chapel Hill, and in consultation with Professor Samuel Weir, M.D. of the same organization. More specifically, the data came from the outpatient clinic of the Department of Family Medicine and consist of logs from 7/1/2005 to 5/31/2007. From these data, we extracted the following information, which has been mainly used to estimate the probability distributions associated with cancellations and no-shows:

C_i: Number of patients who had an appointment delay of i days but canceled their appointments on or before their appointment days.

S_i: Number of patients who had an appointment delay of i days and showed up for their appointments.

M_i: Number of patients who had an appointment delay of i days, did not cancel in advance but missed their appointments.

Notice that $C_i + S_i + M_i$ is the total number of patients who had an appointment delay of i days.

3.1.1 Cancellation and No-show Distributions

Implementation of the heuristics we propose requires the estimation of cancellation and no-show probabilities (α_{ij} and β_{ij} for integer i and j such that $i + j \leq T$). According to the data, for more than 99% of the patients, the appointment delay is less than 90 days. Therefore, estimates are determined for α_{ij} and β_{ij} for which $i + j \leq 90$. We determined the Maximum Likelihood Estimators (MLEs) for these probabilities.

First, define the following:

$$\begin{aligned} q_i &= \mathbf{P}(\mathrm{NS} | T_c \geq i + 1), \\ r_i &= \mathbf{P}(\mathrm{S} | T_c \geq i + 1), \\ u_i &= \mathbf{P}(T_c \leq i). \end{aligned} \tag{3.1}$$

Note that q_i and r_i are respectively the probabilities of the "patient no-show" and "patient show" events given that the patient does not cancel in the first $i + 1$ days after the day she calls for an appointment; u_i is the probability that a patient will cancel no later than i days after she calls for an appointment.

Clearly, we must have $q_i + r_i = 1$ and $u_i \leq u_{i+1}$, $i \in \{0, 1, \dots, T\}$. Recall that each patient's cancellation and no-show behaviors are independent of those of other patients and for any appointment made, there are three possible outcomes: the patient cancels any time on or before the appointment day, the patient misses the appointment without cancellation, and the patient shows up for the appointment. Let $\mathbf{q} = \{q_i\}_{i=0}^T$, $\mathbf{r} = \{r_i\}_{i=0}^T$ $\mathbf{u} = \{u_i\}_{i=0}^T$. Then the MLEs for q_i, r_i, and u_i can be obtained by solving the following maximization problem

$$\begin{aligned} \max \quad & L(\mathbf{q}, \mathbf{r}, \mathbf{u}) = \textstyle\prod_{i=0}^T u_i^{C_i} [(1-u_i)q_i]^{M_i} [(1-u_i)r_i]^{S_i} \\ \text{s.t.} \quad & q_i + r_i = 1, \quad i \in \{0, 1, \dots, T\}, \\ & u_i \leq u_{i+1}, \quad i \in \{0, 1, \dots, T-1\}, \\ & u_i \geq 0, q_i \geq 0, r_i \geq 0, \quad i \in \{0, 1, \dots, T\}. \end{aligned} \tag{3.2}$$

Suppose that we solve the optimization problem (3.2) and obtain the MLEs $\hat{q}_i$, $\hat{r}_i$, and $\hat{u}_i$ for q_i, r_i, and u_i, respectively. Then, we can get the MLEs $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$ for α_{ij} and β_{ij} as follows:

$$\begin{aligned} \hat{\alpha}_{ij} &= \tfrac{\hat{r}_{i+j}(1-\hat{u}_{i+j})}{1-\hat{u}_{i-1}}, \\ \hat{\beta}_{ij} &= \tfrac{1-\hat{u}_{i+j-1}}{1-\hat{u}_{i-1}}. \end{aligned} \tag{3.3}$$

where $\hat{u}_{-1} = 0$ by definition.

However, solving problem (3.2) is difficult especially since it has too many decision variables, i.e., there are too many parameters to estimate. Hence, we propose a parsimonious parametric model, which requires estimating only four parameters. Specifically, assume that q_i, r_i, and u_i take the following form:

$$\begin{aligned} q_i &= 1 - \theta b^{i+1}, \quad i \geq -1, \\ r_i &= \theta b^{i+1}, \quad i \geq -1, \\ u_i &= \begin{cases} 0, & i = -1, \\ 1 - \gamma a^i, & i \geq 0. \end{cases} \end{aligned} \tag{3.4}$$

This model is appealing not only because it is sufficiently simple but also because it has an interpretation that is quite fitting in the appointment scheduling context. First, note that u_i is the cumulative distribution function for the random variable T_c, i.e. the time between the patient's call and the day she cancels (or the day she would cancel if her appointment was not earlier).

Under the parametric form we describe in (3.4), the probability mass function of T_c is a mixture of two distributions, one being a constant and the other being a geometric distribution. To be more precise, we have

$$\mathbf{P}(T_c = i) = (1-\gamma)\mathbf{1}_{\{i=0\}} + \gamma \mathbf{P}(Y_c = i)\mathbf{1}_{\{i \geq 1\}},$$

where $\mathbf{1}_A$ is the indicator function and Y_c is a geometric random variable with parameter $1-a$. One way of interpreting this mixture structure is that there are two different types of patients: those who cancel on the same day they make their appointments, which constitute $1-\gamma$ fraction of the whole patient population, and those who cancel later, which constitute γ fraction of the whole patient population. Furthermore, the model also implies that for those who make appointments at least one day before their appointment day, the probability of cancelling on each day is $1-a$ independently of everything else. (Note that it is possible to make similar interpretations for q_i and r_i as well.)

Notice that once the probabilities are restricted to be in the form given above, the only additional condition needed for all the constraints of problem (3.2) to hold is that $0 \leq \gamma, a, \theta, b \leq 1$. Let $\hat{\gamma}$, $\hat{a}$, $\hat{\theta}$, and $\hat{b}$ denote the MLEs for γ, a, θ, and b, respectively. Then, the first order optimality condition yields

$$\begin{aligned}
\sum_{i=0}^{T}(C_i + M_i + S_i) &= \sum_{i=0}^{T} \frac{C_i}{1-\hat{\gamma}\hat{a}^i}, \\
\sum_{i=0}^{T} i(C_i + M_i + S_i) &= \sum_{i=0}^{T} \frac{iC_i}{1-\hat{\gamma}\hat{a}^i}, \\
\sum_{i=0}^{T}(M_i + S_i) &= \sum_{i=0}^{T} \frac{M_i}{1-\hat{\theta}\hat{b}^{i+1}}, \\
\sum_{i=0}^{T}(i+1)(M_i + S_i) &= \sum_{i=0}^{T} \frac{(i+1)M_i}{1-\hat{\theta}\hat{b}^{i+1}}.
\end{aligned}$$

This system of nonlinear equations can be solved by one of the standard algorithms such as Gauss-Newton method or Trust-Region method. Once this solution is found, the estimates for α_{ij} and β_{ij} can be determined using the following equations, which are obtained by simply substituting

(3.4) into (3.3):

$$\hat{\alpha}_{ij} = \begin{cases} \hat{\theta}\hat{b}^{j+1}\hat{\gamma}\hat{a}^{j} & \text{if } i = 0, \\ \hat{\theta}\hat{b}^{i+j+1}a^{j+1} & \text{if } i \geq 1, \end{cases}$$

$$\hat{\beta}_{ij} = \begin{cases} 1 & \text{if } i = 0, j = 0, \\ \hat{\gamma}\hat{a}^{j-1} & \text{if } i = 0, j \geq 1, \\ \hat{a}^{j} & \text{if } i \geq 1, j \geq 1. \end{cases}$$

Using our data, we find that $\hat{\gamma} = 0.9297$, $\hat{a} = 0.9987$, $\hat{\theta} = 0.8863$, and $\hat{b} = 0.9953$. Then, we numerically verify that this solution is indeed a maximizer. In order to build confidence for our statistical model, we have also conducted a Chi-square goodness-of-fit test, and we find that the distribution we proposed can not be rejected at the significance level of 0.01 (the p-value is 0.042).

Figure 3.1 compares the empirical results reported in Gallucci et al. (2005) with those obtained numerically from our statistical model. Gallucci et al. (2005) study a sample of around 6000 patients who were appointed to a psychiatry outpatient program, and estimate that the total rate of no-shows and cancellations is 12%, 23%, 42%, and 44% corresponding to 0, 1, 7, and 13 days of appointment delay respectively. With our data and using our model, we find the same rates to be 17.99%, 18.48%, 21.37%, and 24.15%, respectively. Not surprisingly, the numbers we find are different from those found by Gallucci et al. Nevertheless, they both point to the same relationship: the longer the appointment delay, the higher the chances of a no-show or cancellation.

3.1.2 Daily Requests for Appointments

We estimate the average number of daily appointment requests to be approximately 50. However, there are no data that would make it possible to estimate its probability distribution. Therefore, in the simulation study, we assume this distribution to be Poisson following other work in the literature, e.g. Patrick et al. (2008), Gupta and Wang (2008) and Robinson and Chen (2008).

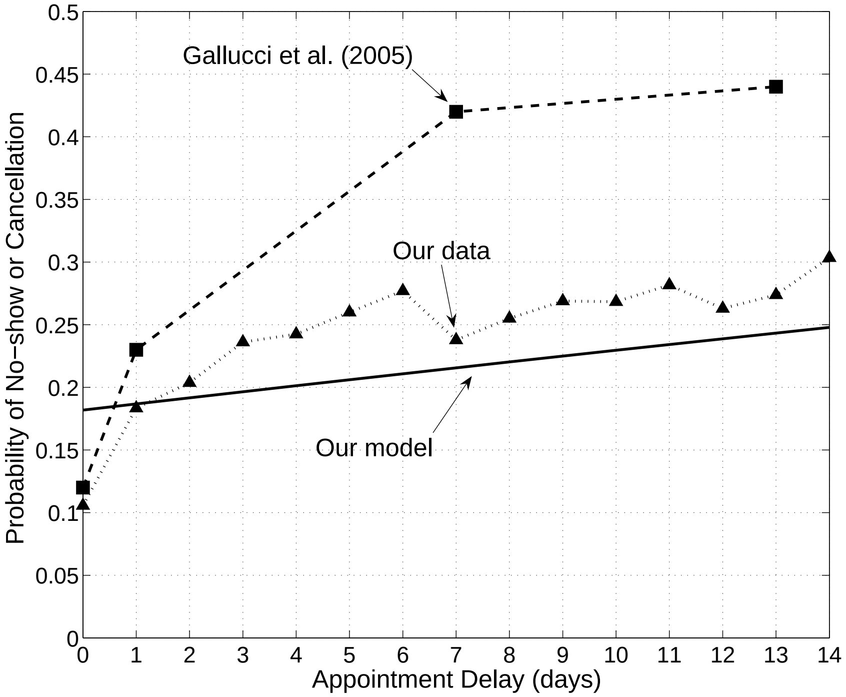

Figure 3.1: Probability of No-show or Cancellation vs. Appointment Delay

3.1.3 Cost and Reward Functions

The reward function $r(\cdot, \cdot)$ for a clinic can be estimated in dollar terms given the relevant data. In most cases, however, the function will be used to reflect the preferences of a clinic regarding various trade-offs that are in play, which will no doubt be highly influenced by financial concerns. While profitability would be a major concern for many private clinics, this might be less of an issue at university hospitals because their missions mainly lie elsewhere: research, education, and public service. For such hospitals, it is difficult to quantify the cost of asking a professor to work overtime because the "cost" is not just the money paid to the professor, but also the "cost" of time spent in the clinic as opposed to time spent for research or teaching. In fact, regardless of the type of the clinic even if the only concern was maximizing profits, not everything could be easily quantified. For example, what is the cost of the reduced quality of service given to the patients if these patients are seen on a day when the clinic exceeded its regular capacity by 15%? On an overloaded day patients will experience longer waits and they will be seen by physicians who are under pressure to clear up the high load. Therefore, the reward function should mostly be seen as user inputs that will differ depending on the circumstances different clinics are operating in as well as their preferences.

In our simulation study, we used the reward function that we determined in consultation with Samuel Weir, Professor and Co-Director of the Family Medicine Center at the University of North Carolina. Although values of specific parameters might be different for different clinics, we believe that the structure of these functions capture the basic trade-off that will be faced by many clinics. Specifically, we assumed that the reward function is in the form of (2.3) with $\eta(x) = x$ (implying one nominal reward for each patient served) and

$$w(z) = \begin{cases} K + h_1 z, & z \leq M, \\ K + h_1 M + h_2(z - M), & z > M, \end{cases} \tag{3.5}$$

In (3.5), $K \geq 0$ can be seen as the daily fixed cost, $M \geq 0$ as the regular daily capacity of the clinic, $h_1 \geq 0$ as the regular time cost of one scheduled patient, $h_2 \geq h_1$ as the overtime cost of one scheduled patient. This cost formulation makes it possible for the clinic to make its preference regarding daily patient overloads. If h_2 is set sufficiently large, the policy will not schedule more than M patients on a single day unless all days are full. On the other hand, if h_1 and h_2 are set equal to each other, that implies the clinic does not mind overloading the clinic. Most clinics will prefer being somewhere in between, which they can determine by setting h_1 and h_2 accordingly.

3.2. Performance Comparison of the Heuristics: A Simulation Study

This section summarizes the results of the simulation study we carried out in order to investigate the performances of the scheduling policies proposed. As we have discussed in the previous section, the "model clinic" we used in our simulation study was created using data from the Department of Family Medicine at UNC. We first derive OTPSP, and indices for Imp-OTPSP, and Imp-OAP for the "model clinic," which has a linear reward function with one nominal reward for each patient served and a cost function given by (3.5). Then we introduce the benchmark policies to be compared with the heuristics, describe the simulation setup and finally we report and discuss the findings of the simulation study.

3.2.1 Derivation of the Heuristics for the Simulation Study

Suppose that λ denotes the mean number of daily appointment requests. Note that for our "model clinic," λ is estimated to be 50. Let $\tilde{\lambda}_1(p_0) = \lambda p_0(\alpha_{00} - \alpha_{01}) + \lambda\alpha_{01}$ and $\tilde{\lambda}_2(p_0) = \lambda p_0(1 - \beta_{01}) + \lambda\beta_{01}$. Then, we can show that

$$R_{TPSP}(p_0) = \tilde{\lambda}_1(p_0) - \left[K + h_1\tilde{\lambda}_2(p_0) + (h_2 - h_1)\left(\sum_{i=M}^{\infty}(i-M)e^{-\tilde{\lambda}_2(p_0)}\frac{[\tilde{\lambda}_2(p_0)]^i}{i!}\right)\right]. \tag{3.6}$$

Recall that OTPSP is obtained by finding p_0 that maximizes (3.6). Let p_0^* be the maximizer and $\mathbf{p}^* = [p_0^*, 1 - p_0^*, 0, 0, \ldots, 0]$, which is a $(T+1)$-dimensional vector. Then, OTPSP is the policy of scheduling appointments independently of the system state using probability vector $\mathbf{p}^*$.

Now, let $\Gamma_j \overset{d}{=} \sum_{i=1}^{T-j} V_{ij}(\mathbf{x}) + \bar{V}_j(\mathbf{y}) + \sum_{k=\max\{j-1,1\}}^{j} \hat{V}_{kj}(\mathbf{p}^*)$ where $\sum_{k=\max\{j-1,1\}}^{j} \hat{V}_{kj}(\mathbf{p}^*) = 0$ if $j = 0$. Then, using (2.6), (2.8), and (3.5) we can show that the indices for Imp-OTPSP are given by

$$I_j = \alpha_{0j} - \beta_{0j}\{\mathbf{E}[w(\Gamma_j + 1)] - \mathbf{E}[w(\Gamma_j)]\} = \alpha_{0j} - \beta_{0j}[h_1 + \mathbf{P}(\Gamma_j \geq M)(h_2 - h_1)]. \tag{3.7}$$

The indices for Imp-OAP are also given by (3.7) except that for Imp-OAP, $\mathbf{p} = [1, 0, 0, \ldots]$, $\Gamma_j \overset{d}{=} \sum_{i=1}^{T-j} V_{ij}(\mathbf{x}) + \bar{V}_j(\mathbf{y}) + \hat{V}_{jj}(\mathbf{p})$, and $\hat{V}_{00}(\mathbf{p}) = 0$.

3.2.2 Other Benchmark Policies

Since determining the optimal policy is not practically feasible because of the large state space, we compare the performances of the policies we propose with those of other heuristic policies that mimic some of the scheduling principles that are followed in practice and can be intuitively expected to perform well at least under certain conditions. Here, we describe these policies which will serve as benchmark policies.

Policy V: Threshold Policy (TP): This policy schedules each new appointment for the earliest day with less than M patients already scheduled. If there are no days within the scheduling horizon that has less than M already scheduled, the new appointment is scheduled for the day with the fewest appointments, and ties are broken in favor of the earliest day. Threshold M is a policy parameter. One reasonable choice for M is the regular daily capacity of the clinic.

Policy VI: Balanced Scheduling Policy (BSP): This policy schedules each new appointment for the day with the fewest appointments, and ties are broken in favor of the earliest day.

Policy VII: Random Scheduling Policy (RSP): This policy schedules each new appointment randomly for one of the days in the scheduling horizon. More precisely, it is a probabilistic static policy with $p_i = 1/(T+1), i = 0, 1, \ldots, T$.

3.2.3 Findings of the Simulation Study

The objective of the simulation study is to compare the performances of the heuristic policies we propose (OTPSP, Imp-OAP, and Imp-OTPSP) with those of the benchmark heuristics (OAP, TP, BSP, RSP) introduced in Section 2.5. For TP, we set the threshold to be the regular daily capacity of the clinic, which is denoted by M as defined in Section 3.2.2.

Recall that we picked $\eta(x) = x$ and $w(\cdot)$ as in (3.5). As for the values of the parameters, we set $K = 0$ without loss of generality and $h_2 = 0.95$. We simulated various scenarios by considering different combinations of values for M and h_1. Specifically, M took values in $\{40, 45, 50, 55\}$ and h_1 took values in $\{0.0, 0.2, 0.5\}$. Since the mean daily arrivals were set to 50, picking different values for M allowed us to test the performances of the policies under various conditions of system load. Clinics in scenarios where $M = 55$ can be seen as underloaded while those in scenarios where $M = 40$ or $M = 45$ are overloaded. On the other hand, assigning different values for h_1 helped us capture different preferences that a clinic might have for admitting more patients than the regular daily capacity. Our preliminary analysis indicated that, the policies we propose very rarely scheduled appointments more than 15 days in advance, and the times it took the simulation runs to complete quickly increased with the value of T, the maximum value for the appointment delay. Therefore, we set $T = 15$ in all scenarios.

We coded a simulation program in Matlab and used the batch-means method (see, e.g. Section 9.5 in Law and Kelton (2000)). For each scenario, we ran 11 batches, each batch consisting of 200 consecutive workdays. The first batch was used as the warm-up period. As we will see later, this simulation setup is adequate to allow us to make informative conclusions. A total of 12 different scenarios were considered each with a different pair of values for M and h_1. Each scenario was simulated under each one of the seven scheduling policies, i.e. OAP, Imp-OAP, OTPSP, Imp-OTPSP, TP, BSP and RSP, and the long-run average net reward was computed. For all scenarios considered, OTPSP turns out to be a deterministic static policy which always assigns the arriving appointment request to the next day, i.e. $p_0 = 0$ and $p_1 = 1$. It is difficult to give a clear description of the dynamic heuristics Imp-OAP and Imp-OTPSP, but our numerical observations suggest that these policies generally resemble BSP, but unlike BSP, which schedules patients to the day with the fewest scheduled appointments, these two heuristics "balance" the load according to the indices. In order to facilitate comparison, we chose OAP (Open Access Policy) as the benchmark policy and for every other policy computed the percentage improvement that would be obtained by using the policy as opposed to OAP. Finally, we determined the 95% confidence interval for the mean percentage improvement. The results are given in Table 3.1.

In Table 3.1, the first number for each scenario-policy pair is the mean percentage improvement while the second number is the half width of the 95% confidence interval for the mean. Therefore, if the first number is larger than the second number, that indicates the corresponding policy is superior than OAP at the 5% significance level. On the other hand, if the first number is negative and it is larger than the second number in absolute value, that implies the superiority of OAP. The cases where the numbers indicate superiority of one policy over the other (in either direction) at the 5% significance level are shown in bold face. Note that the comparison is inconclusive in only two cases.

A quick look at Table 3.1 reveals that Imp-OTPSP, OTPSP, Imp-OAP, and TP are the "best" policies under a variety of conditions. Although the Open Access Policy (OAP) does not perform as well as these policies, it does perform better as the regular capacity M gets larger. This is not

		Imp-OTPSP	OTPSP	Imp-OAP
$M = 55$	$h_1 = 0$	**$2.11\% \pm 0.46\%$**	**$0.78\% \pm 0.32\%$**	**$2.18\% \pm 0.49\%$**
	$h_1 = 0.2$	**$4.10\% \pm 0.95\%$**	**$3.23\% \pm 0.75\%$**	**$3.08\% \pm 0.63\%$**
	$h_1 = 0.5$	**$12.74\% \pm 1.05\%$**	**$12.14\% \pm 1.10\%$**	**$3.72\% \pm 1.34\%$**
$M = 50$	$h_1 = 0$	**$6.77\% \pm 0.76\%$**	**$2.75\% \pm 0.37\%$**	**$5.42\% \pm 0.70\%$**
	$h_1 = 0.2$	**$8.28\% \pm 0.97\%$**	**$5.48\% \pm 0.86\%$**	**$6.96\% \pm 0.41\%$**
	$h_1 = 0.5$	**$18.56\% \pm 1.30\%$**	**$15.29\% \pm 1.31\%$**	**$9.25\% \pm 1.26\%$**
$M = 45$	$h_1 = 0$	**$10.63\% \pm 0.52\%$**	**$6.23\% \pm 0.60\%$**	**$9.25\% \pm 0.51\%$**
	$h_1 = 0.2$	**$13.35\% \pm 0.77\%$**	**$9.13\% \pm 0.76\%$**	**$11.53\% \pm 0.72\%$**
	$h_1 = 0.5$	**$25.01\% \pm 2.10\%$**	**$20.32\% \pm 1.41\%$**	**$21.78\% \pm 1.57\%$**
$M = 40$	$h_1 = 0$	**$9.84\% \pm 0.67\%$**	**$9.12\% \pm 0.44\%$**	**$10.21\% \pm 0.39\%$**
	$h_1 = 0.2$	**$13.03\% \pm 0.66\%$**	**$12.48\% \pm 0.68\%$**	**$13.69\% \pm 0.90\%$**
	$h_1 = 0.5$	**$27.41\% \pm 1.87\%$**	**$26.82\% \pm 1.49\%$**	**$28.13\% \pm 1.59\%$**

		TP	BSP	RSP
$M = 55$	$h_1 = 0$	**$2.11\% \pm 0.46\%$**	**$-6.30\% \pm 0.53\%$**	**$-3.28\% \pm 0.41\%$**
	$h_1 = 0.2$	**$3.25\% \pm 0.61\%$**	**$-5.48\% \pm 0.72\%$**	**$-1.53\% \pm 0.49\%$**
	$h_1 = 0.5$	**$4.39\% \pm 1.08\%$**	**$-4.53\% \pm 1.21\%$**	**$2.68\% \pm 1.02\%$**
$M = 50$	$h_1 = 0$	**$6.45\% \pm 0.73\%$**	**$-2.22\% \pm 0.73\%$**	**$-1.20\% \pm 0.51\%$**
	$h_1 = 0.2$	**$8.21\% \pm 0.92\%$**	**$-1.09\% \pm 0.89\%$**	$0.50\% \pm 0.66\%$
	$h_1 = 0.5$	**$12.11\% \pm 1.68\%$**	$0.72\% \pm 1.47\%$	**$5.31\% \pm 1.28\%$**
$M = 45$	$h_1 = 0$	**$5.24\% \pm 0.80\%$**	**$4.11\% \pm 0.54\%$**	**$1.81\% \pm 0.70\%$**
	$h_1 = 0.2$	**$6.28\% \pm 0.10\%$**	**$4.91\% \pm 0.69\%$**	**$3.28\% \pm 0.88\%$**
	$h_1 = 0.5$	**$10.40\% \pm 1.97\%$**	**$8.10\% \pm 1.38\%$**	**$9.16\% \pm 1.65\%$**
$M = 40$	$h_1 = 0$	**$2.79\% \pm 0.70\%$**	**$2.99\% \pm 0.55\%$**	**$4.23\% \pm 0.73\%$**
	$h_1 = 0.2$	**$3.57\% \pm 0.97\%$**	**$3.83\% \pm 0.75\%$**	**$6.05\% \pm 0.95\%$**
	$h_1 = 0.5$	**$6.79\% \pm 2.12\%$**	**$7.32\% \pm 1.62\%$**	**$13.58\% \pm 1.96\%$**

Table 3.1: Results of the Simulation Study - (The first number indicates the mean percentage improvement of the corresponding policy over OAP, and the second number indicates the half width for the 95% confidence interval.)

surprising since as we also discussed in Section 2.1, OAP is an ideal policy when the system is not overloaded. It would thus be reasonable to expect that OAP would be among the best policies if M were larger.

In order to better compare the four best policies Imp-OTPSP, OTPSP, Imp-OAP, and TP among each other we also determined the best policies for each scenario separately, which are listed in Table 3.2. More specifically, for each scenario we conducted paired t-tests for every possible pair of policies and determined whether or not there is a statistical difference between their performances at a significance level of 0.05. For every scenario, we list the policies whose performances are better than those of the others. Note that for some of the scenarios, there is more than one policy. That is because in those cases, paired t-test is inconclusive meaning that the performances of the policies are not statistically different.

Scenarios		Best Policies			
$M = 55$	$h_1 = 0$	Imp-OTPSP		Imp-OAP	TP
	$h_1 = 0.2$	Imp-OTPSP		Imp-OAP	TP
	$h_1 = 0.5$	Imp-OTPSP	OTPSP		
$M = 50$	$h_1 = 0$	Imp-OTPSP			
	$h_1 = 0.2$	Imp-OTPSP			TP
	$h_1 = 0.5$	Imp-OTPSP			
$M = 45$	$h_1 = 0$	Imp-OTPSP			
	$h_1 = 0.2$	Imp-OTPSP			
	$h_1 = 0.5$	Imp-OTPSP			
$M = 40$	$h_1 = 0$	Imp-OTPSP		Imp-OAP	
	$h_1 = 0.2$	Imp-OTPSP		Imp-OAP	
	$h_1 = 0.5$	Imp-OTPSP		Imp-OAP	

Table 3.2: "Best" Policies for Each Scenario at the 5% Significance Level

Table 3.2 clearly shows that the policies we propose particularly Imp-OTPSP perform well. The superiority of Imp-OTPSP over OTPSP is actually guaranteed given that Imp-OTPSP is obtained by applying a policy improvement step over OTPSP. It is also not surprising (though not guaranteed) that overall Imp-OTPSP performs better than Imp-OAP since the static policy that Imp-OTPSP improves upon (OTPSP) is superior than the static policy Imp-OAP improves upon (OAP).

TP performs quite well when the regular capacity M is large, but as it can be observed from Table 3.1, when M is small it is no better than BSP or even RSP, which schedules appointments randomly. On the other hand, all three policies we propose perform consistently well across all scenarios, suggesting that they are more robust than TP. Note that, one can in fact show that TP is an optimal policy if the clinic ignores patient no-shows and cancellations. Therefore, the poor performance of TP when the regular capacity is small also shows how damaging such omissions in formulation can be.

The superiority of Imp-OTPSP is more pronounced for mid values of the regular capacity ($M = 45$ or $M = 50$). This might be a consequence of the fact that there is no room for Imp-

OTPSP to make a difference when the capacity is large or small. When the regular capacity is large, most policies manage to keep scheduled appointments below daily capacity without delaying appointments too long. As a result most policies perform reasonably well and there is not much to gain from more sophisticated policies. On the other hand, when the regular capacity is small, under most policies the regular capacity is exceeded, which again limits the opportunity for sophisticated policies to make a difference. For mid values of the regular capacity, there are more choices that an intelligent scheduling policy can make because unlike the low or high capacity cases, some days will be overloaded and some days will be underloaded. An intelligent policy can work to smooth the daily loads so that overloads and no-shows are avoided to the extent possible.

So, what are the implications of our findings for practice? One should seek simplicity if complexity does not bring any significant advantages. Therefore, it appears that as suggested by Green and Savin (2008), if demand and supply are in balance, Open Access is quite reasonable especially if this policy is believed to have additional benefits that were not quantified in our formulation. However, if demand and supply are not in balance, if average demand is close to regular capacity or somewhat higher, then the policies we propose appear to be much better than the Open Access policy or other alternatives.

CHAPTER 4

Dynamic Scheduling of Outpatient Appointments: Multi-class Patients

We study the dynamic scheduling of outpatient appointments in Chapter 2. One key assumption there is that all patients are statistically the same in their no-show and cancellation behaviors. To be more specific, if two patients have the same appointment delay, have been waiting for their appointments for the same amount of time and have not cancelled their appointments yet, then the probabilities that they will show up for their appointments are exactly the same, regardless of anything else. However, as shown in empirical studies (see, e.g. Barron (1980), Vikander et al. (1986) and Weingarten et al. (1997) etc.), patient no-show rates also depend on a variety of other factors including age, income, family size, social and marital status etc.

In this chapter we consider the dynamic scheduling of outpatient appointments when patients are categorized into different classes according to their no-show and cancellation behaviors. We develop an MDP formulation which generalizes the model described in Chapter 2. Due to its significantly enlarged state space, we derive policy improvement heuristics by applying the same ideas used in Chapter 2. In order to evaluate the performance of the heuristics, we follow a similar procedure to that in Chapter 2. Since we do not have data which enable us to study no-show and cancellation behaviors of patients from multiple class, we first simulate data as if we collected them from practice. Then we use these data to populate a "model clinic," and perform an extensive simulation study to compare the performances of the proposed heuristic policies and those of the other benchmark polices introduced in Section 3.2.2.

4.1. The Model and Policy Improvement Heuristics

Suppose that patients can be grouped into C different classes. Patients of the same class have the same no-show and cancellation probability distributions, which are different from those of the patients from other classes. This assumption is similar to the one used in the model of Muthuraman and Lawley (2008). To capture patients' no-show and cancellation behaviors, we choose a probability model similar to the one used in Section 2.3. The only difference is that the no-show and cancellation probabilities now depend on the class that patients belong to. To be more specific, we formulate the cancellation/no show behavior as follows: we assume that each patient cancels her appointment at some random time in the future, which can possibly be beyond the patient's appointment day. We use T_c^k to denote the generic random variable that represents the time between the day a patient of class k calls for an appointment and the day she decides to cancel it. Recall that "S" and "NS" represent the "show" and "no-show" events for a particular patient, respectively. Define

$$\alpha_{ijk} = \mathbf{P}(T_c^k \geq i+j+1, \mathrm{S} | T_c^k \geq i), \tag{4.1}$$

$$\beta_{ijk} = \mathbf{P}(T_c^k \geq i+j | T_c^k \geq i). \tag{4.2}$$

Recall that for any given day t, type (i,j) patients are referred to those who called on day $t-i$, were given appointments for day $t+j$, and have not canceled their appointments by the beginning of day t. Thus, α_{ijk} is the probability that a patient of class k who is currently of type (i,j) will show up for her appointment and β_{ijk} is the probability that a patient of class k who is currently of type (i,j) will not cancel her appointment before her appointment day.

Let A_k^t denote the total number of calls made by patients of class k in the morning of day t, $k = 1, 2, \ldots, C$. Suppose that for any k, $\{A_k^t, t = 1, 2, \ldots\}$ is a sequence of i.i.d. random variables. Assume that patients of different classes call for appointments independently. Thus A_k^t and A_j^s are independent if $t \neq s$ and $k \neq j$. Suppose that the clinic knows the class a patient belongs to upon her call. Let $\mathbf{X}^t = \{X_{ijk}^t\}$ denote the appointment backlog in the morning of day t, where X_{ijk}^t represents the number of patients from class k who are of type (i,j). The scheduling decision on day t is made based on $\mathbf{X}^t$.

We assume the same cost structure as in the single-class scheduling problem considered in Chapter 2: the clinic makes a net reward $r(x, z)$ if z patients (regardless of their classes) has been scheduled on that day and the clinic has actually served x patients (regardless of their classes) during the day. The events are assumed to occur in the same order on each day as that in the single-class scheduling problem.

The objective of the clinic is to schedule arriving appointment requests (based on the current schedule) so that the long-run average expected net reward is maximized. This problem can be formulated as an MDP. We assume that $A_k^t, k = 1, 2, \ldots, C$ is known (realized) at the beginning of day t. As we will see later, this assumption is only necessary for the formulation of the problem and the policy we will propose does not require knowing A_k^t in advance. Let $\mathbf{A}^t = \{A_k^t, k = 1, 2, \ldots, C\}$. Thus the decision epochs are the times right after the appointment requests arrive every day, and the system state on day t is $(\mathbf{A}^t, \mathbf{X}^t)$. Suppose that the scheduling horizon is T. The scheduling decision to make on day t is to decide out of A_k^t patients how many of them should be scheduled for day $t + j$ given $\mathbf{X}^t$, for $k = 1, 2, \ldots, C$ and $j = 0, 1, \ldots, T$. To be more precise, let Y_{jk}^t denote the number of patients from class k calling on day t and scheduled for day $t + j$. The clinic needs to decide the value of Y_{jk}^t's. Let $\mathbf{Y}^t = \{Y_{jk}^t, j = 0, 1, \ldots, T, k = 1, 2, \ldots, C\}$. Then the action space on day t is the set $\{\mathbf{Y}^t : \sum_{i=0}^T Y_{jk}^t = A_k^t, Y_{jk}^t \in \mathbb{Z}_+, j = 0, 1, \ldots, T, k = 1, 2, \ldots, C\}$.

This MDP problem is a generalization of the one formulated in Section 2.3. The state transition probabilities and the long-run expected average net reward under an arbitrary policy f can be defined in a similar way to those in Section 2.3. This MDP formulation has a significantly enlarged state space as compared with its counterpart in Section 2.3, and thus any realistically sized instance of it can not be solved to optimality using the standard numerical methods such as the policy improvement or value iteration algorithms. Hence we follow the same ideas in Section 2.4 to develop policy improvement heuristics to solve this MDP problem.

The whole procedure of developing the heuristics is analogous to that presented in Section 2.4. Thus we will only sketch the outline here. For the ease of presentation, we use similar notations as used in Section 2.4 wherever they do not cause ambiguities. Consider the following state independent policy that schedules each patient of class k according to a stationary probability distribution $\mathbf{p}_{\cdot k} = (p_{0k}, p_{1k}, \ldots, p_{Tk})$ where p_{jk} is the probability that a patient of class k is given an appointment for the jth day from today. We call this policy *Generalized Probabilistic Static Policy* (GPSP). Consider a policy $\pi_{\mathbf{y}}$ that takes action $\mathbf{y} \equiv \{y_{jk}, j = 0, 1, \ldots, T, k = 1, 2, \ldots, C\}$ at the beginning of today, i.e., schedules y_{jk} patients of class k on the jth day from today for $j = 0, 1, \ldots, T$ and $k = 1, 2, \ldots, C$, but from tomorrow on uses GPSP. Let $\mathbf{p} = \{\mathbf{p}_{\cdot k}, k = 1, 2, \ldots, C\}$ and define $\Delta((\mathbf{a}, \mathbf{x}), \mathbf{y}, \mathbf{p})$ to be the difference in total expected net rewards over an infinitely long period of time by following policy $\pi_{\mathbf{y}}$ rather than using GPSP all along given the initial state $(\mathbf{a}, \mathbf{x})$. To conduct a one-step policy improvement, we maximize $\Delta((\mathbf{a}, \mathbf{x}), \mathbf{y}, \mathbf{p})$ with respect to $\mathbf{y}$ for each state $(\mathbf{a}, \mathbf{x})$, and the resulting optimal $\mathbf{y}$, denoted as $\mathbf{y}^*(\mathbf{a}, \mathbf{x})$, specifies the heuristic scheduling policy for state $(\mathbf{a}, \mathbf{x})$. We call this policy *Generalized Heuristic Dynamic Policy* (GHDP).

It is straightforward to follow the similar derivation in Section 2.4.1 to make a one-step improvement over GPSP to obtain GHDP. However, it entails solving the following resource allocation problem (4.3) which is much harder than problem (2.7).

$$\begin{array}{ll} \max & \sum_{j=0}^{T} f_j(y_{j1}, y_{j2}, \ldots, y_{jC}, \mathbf{x}, \mathbf{p}) \\ \text{s.t.} & \sum_{j=0}^{T} y_{jk} = a_k, \quad k = 1, 2, \ldots, C, \\ & y_{jk} \in \mathbb{Z}_+, \quad j = 0, 1, \ldots, T, \quad k = 1, 2, \ldots, C. \end{array} \tag{4.3}$$

We next describe the function $f_j(\cdot)$. For $1 \leq i \leq T$ and $0 \leq j \leq T - i$, let $V_{ijk}(\mathbf{x})$ denote the number of class k patients who called for appointments i days before day 0 and will not cancel by the morning of their appointment on day j, and $W_{ijk}(\mathbf{x})$ denote the number of these patients who show up for their appointments. Similarly, for $0 \leq j \leq T$, let $\bar{V}_{jk}(\mathbf{y})$ denote the number of class k patients who call for appointments on day 0, are scheduled for day j and will not cancel by the morning of their appointment, and $\bar{W}_{jk}(\mathbf{y})$ denote the number of these patients who show up for their appointments. Finally, for $1 \leq s \leq T$ and $s \leq j \leq T$, we define $\hat{V}_{sjk}(\mathbf{p})$ to be the number of class k patients who will call for appointments on day s and will not have canceled their appointment by the morning of their appointment on day j, and we define $\hat{W}_{sjk}(\mathbf{p})$ to be the number of these patients who will show up for their appointments. Then, since the cancellation/no-show behaviors of the patients are independent of each other, we know that each one of these random variables has a Binomial distribution. More precisely,

$$V_{ijk}(\mathbf{x}) \overset{d}{=} B(x_{ijk}, \beta_{ijk}), \bar{V}_{jk}(\mathbf{y}) \overset{d}{=} B(y_{jk}, \beta_{0jk}), \hat{V}_{kj}(\mathbf{p}) \overset{d}{=} B(A_k^{(s)}, p_{j-s,k}\beta_{0,j-s,k}),$$

and

$$W_{ijk}(\mathbf{x}) \overset{d}{=} B(x_{ijk}, \alpha_{ijk}), \bar{W}_{jk}(\mathbf{y}) \overset{d}{=} B(y_{jk}, \alpha_{0jk}), \hat{W}_{sjk}(\mathbf{p}) \overset{d}{=} B(A_k^{(s)}, p_{j-s,k}\alpha_{0,j-s,k}),$$

where $A_k^{(s)}$ denotes the number of class k appointment requests that will be made on day s, $1 \leq s \leq T$. Then $f_j(y_{j1}, y_{j2}, \ldots, y_{jK}, \mathbf{x}, \mathbf{p})$ is is given by

$$\begin{aligned} & f_j(y_{j1}, y_{j2}, \ldots, y_{jC}, \mathbf{x}, \mathbf{p}) \\ & = \mathbf{E}\left\{ r\left[\sum_{k=1}^{C} \left(\sum_{i=1}^{T-j} W_{ijk}(\mathbf{x}) + \bar{W}_{jk}(\mathbf{y}) + \sum_{s=1}^{j} \hat{W}_{sjk}(\mathbf{p}) \right), \sum_{k=1}^{C} \left(\sum_{i=1}^{T-j} V_{ijk}(\mathbf{x}) + \bar{V}_{jk}(\mathbf{y}) + \sum_{s=1}^{j} \hat{V}_{sjk}(\mathbf{p}) \right) \right] \right\}, \end{aligned}$$

where we let $\sum_{s=1}^{j} \hat{W}_{sjk}(\mathbf{p}) = \sum_{s=1}^{j} \hat{V}_{sjk}(\mathbf{p}) = 0$ for $j = 0$ and $k = 1, 2, \ldots, C$.

The problem (4.3) can be solved to optimality by a dynamic programming (DP) procedure called DPMDR presented in Chapter 10 of Ibaraki and Katoh (1988). We illustrate this procedure as follows. For a nonnegative integer vector $\mathbf{n} = \{n_k\}_{k=1}^{C}$ and a positive integer u with $0 \leq u \leq T$, define $F_j(n)$ by

$$\begin{aligned} & F_{-1}(\mathbf{n}) = 0, \\ & F_u(\mathbf{n}) = \min\{\sum_{j=0}^{u} f_j(y_{j1}, y_{j2}, \ldots, y_{jC}, \mathbf{x}, \mathbf{p}) | \sum_{j=0}^{u} y_{jk} = n_k, \quad k = 1, 2, \ldots, C; \\ & \qquad\qquad y_{jk} \in \mathbb{Z}_+, \quad j = 0, 1, \ldots, u, \quad k = 1, 2, \ldots, C.\} \end{aligned}$$

It is clear that for $\mathbf{a} = \{a_k\}_{k=1}^{C}$, $F_T(\mathbf{a})$ solves problem (4.3). As an analog to Equation (10.5.14) in Ibaraki and Katoh (1988), the following recursion equation holds.

$$F_0(\mathbf{n}) = f_0(n_1, n_2, \ldots, n_C, \mathbf{x}, \mathbf{p}), \tag{4.4}$$

$$F_u(\mathbf{n}) = \min\{F_{u-1}(\mathbf{n} - \mathbf{y}_u) + g_u(\mathbf{y}_u) | \mathbf{0} \leq \mathbf{y}_u \leq \mathbf{n}\}, \quad u = 1, 2, \ldots, T, \tag{4.5}$$

where $\mathbf{y}_u = \{y_{uk}\}_{k=1}^C$ and $g_u(\mathbf{y}_u) = f_u(y_{u1}, y_{u2}, \ldots, y_{uC}, \mathbf{x}, \mathbf{p})$. Thus the DP scheme works as follows.

The DP Scheme

Step 1: compute $F_0(\mathbf{n})$ for $\mathbf{0} \leq \mathbf{n} \leq \mathbf{a}$ using equation (4.4);

Step 2: compute $F_u(\mathbf{n})$ for $u = 1, 2, \ldots, T$ and $\mathbf{0} \leq \mathbf{n} \leq \mathbf{a}$ using equation (4.5);

Step 3: $F_T(\mathbf{a})$ yields the optimal solution to problem (4.3).

Though this DP scheme can solve problem (4.3) and yield the GHDP, however, this problem is NP-hard in general (see Theorem 10.5.3 in Ibaraki and Katoh (1988)). This means that the computational time taken by the DP scheme to solve problem (4.3) increases exponentially in the size of the problem. To be more specific, the complexity of this DP scheme is $O(CT \prod_{k=1}^{C}(a_k + 1)^2)$, assuming the evaluation of $f_j(\cdot)$ is done in constant time (see Theorem 10.5.4 in Ibaraki and Katoh (1988)). Such a complexity makes this DP procedure impractical to solve large sized problems. If there are only two classes of patients, i.e. $C = 2$, Zeitlin (1981) develops an easier algorithm to solve problem (4.3) to optimality if the function $f_j(y_{j1}, y_{j2}, \mathbf{x}, \mathbf{p})$ is doubly, mixed and generally concave in y_{j1} and y_{j2} for all $j = 0, 1, \ldots, T$. However, this assumption is too restrictive and not usually satisfied. These suggest that it is impossible to attempt an optimal solution for problem (4.3) by using standard numerical procedures. Thus we need to develop heuristics or approximation methods for solving this problem. Considering practical implementability and computational feasibility, we propose the following greedy algorithm, call Greedy-HDP, to solve problem (4.3). It is an analog to the HDP presented in Section 2.4.2: whenever an appointment request arrives, the Greedy-HDP computes an index for each day within the scheduling horizon and schedules the appointment for the day with the largest index.

Greedy Heuristic Dynamic Policy

When an appointment request comes in, first identify which class it belongs to, say k. Then for *each day $j \in \{0, 1, 2, \ldots, T\}$ calculate the corresponding index I_j using the following formula*

$$I_j = f(y_{j1}, y_{j2}, \ldots, y_{jk} + 1, \ldots, y_{jC}, \mathbf{x}, \mathbf{p}) - f(y_{j1}, y_{j2}, \ldots, y_{jk}, \ldots, y_{jC}, \mathbf{x}, \mathbf{p}), \tag{4.6}$$

where y_{jk} is the number of class k patients who called so far today and were scheduled for the jth day from today and $\mathbf{x}$ is the appointment schedule upon the arrival of this request. Then, determine $j^* = \arg\max_{j \in \{0,1,\ldots,T\}} I_j$, schedule the new appointment for day j^* and update $y_{j^*k} = y_{j^*k} + 1$.

Though the Greedy-HDP does not guarantee to solve problem (4.3) to optimality, it does provide an implementable and computational feasible policy for practice compared to the DP scheme and Zeitlin's method. On the one hand, if the scheduling administrator wants to use the DP scheme or Zeitlin's method, she has to know all appointment requests received in a day before making scheduling decisions, i.e. she needs to wait until $\mathbf{A}^t$ has been realized before deciding $\mathbf{Y}^t$ on day t. This is not a very realistic assumption. However, the Greedy-HDP schedules appointment requests on a call-by-call basis and thus requires no information on how many appointment requests will be received in the rest of the day. On the other hand, the Greedy-HDP is much less computationally demanding than the DP scheme and does not require any assumptions on the objective function compared with Zeitlin's method. It can be used to solve realistically sized problems in practice.

4.2. Performance Comparison of the Heuristics: A Simulation Study

In Chapter 3 we obtained the appointment data from the Department of Family Medicine at UNC and used these data to create a "model clinic," based on which we carried out an extensive simulation study to evaluate the performances of different policies. In this section, we follow a similar procedure. Since we do not have data that enable us to study the heterogeneity in patients' no-show and cancellation behaviors, we choose to simulate the appointment data as if we collected them from a clinic which scheduled patients of multiple classes. Then we use these data to estimate the parameters for the statistical model of patients' no-show and cancellation behaviors. Using these parameters, we conduct a simulation study to compare the performances of the proposed heuristics and those of the other benchmark policies introduced in Section 3.2.2.

4.2.1 Data Generation

Suppose that there are two classes of patients requesting appointments at the clinic. For class k patients, define

$$\begin{aligned} q_{ik} &= \mathbf{P}(\text{NS}|T_c^k \geq i+1), \\ r_{ik} &= \mathbf{P}(\text{S}|T_c^k \geq i+1), \\ u_{ik} &= \mathbf{P}(T_c^k \leq i). \end{aligned} \tag{4.7}$$

We assume that the no-show and cancellation behaviors of patients of both classes can be characterized by a same family of probability distributions but with different parameters. To be more specific, for class k patients we have

$$\begin{aligned} q_{ik} &= 1 - \theta_k b_k^{i+1}, \quad i \geq -1, \\ r_{ik} &= \theta_k b_k^{i+1}, \quad i \geq -1, \\ u_{ik} &= \begin{cases} 0, & i = -1, \\ 1 - \gamma_k a_k^i, & i \geq 0. \end{cases} \end{aligned} \tag{4.8}$$

Table 4.1 shows the true values of parameters for patients of both classes. Note that class 2 patients have higher probabilities to cancel or not show up for their appointments. We assume that daily requests from each class of patients follow a Poisson distribution and the average total daily arrivals from both classes are 50 patients per day. We consider three mixes of these two classes of patients. Table 4.2 presents the average daily arrivals of each class under these three mixes, where λ_k is the

Parameters	$k=1$	$k=2$
γ_k	0.96	0.93
a_k	0.99	0.97
b_k	0.91	0.89
θ_k	0.99	0.97

Table 4.1: Parameters for Patient Behaviors

average daily arrivals of class k patients. We observe that class 2 patients dominate in Mix I, class 1 patients prevails in Mix III and both classes have the same arrival rate in Mix II.

Mix	λ_1	λ_2
I	10	40
II	25	25
III	40	10

Table 4.2: Patient Mixes

We assume that the clinic has a daily regular capacity 50 and follows the Threshold Policy (TP) as described in Section 3.2.2 to schedule arriving appointment requests. We simulate the clinic's operation for one year and collect the appointment data throughout the year. Depending on whether or not the clinic knows exactly which class an arriving patient belongs to, the clinic will collect the data on a disaggregate or an aggregate level.

Disaggregate Data: If the clinic knows exactly which class an arriving patient belongs to, then it can collect the following disaggregate data for both class, in a similar format to that of Section 3.1.

C_{ik}: Number of class k patients who had an appointment delay of i days but canceled their appointments on or before their appointment days.

S_{ik}: Number of class k patients who had an appointment delay of i days and showed up for their appointments.

M_{ik}: Number of class k patients who had an appointment delay of i days, did not cancel in advance but missed their appointments.

In practice, the clinic does not know the true values for γ_k, a_k, θ_k and b_k. However, it can obtain the MLEs for them (denoted by $\hat{\gamma}_k$, $\hat{a}_k$, $\hat{\theta}_k$ and $\hat{b}_k$, respectively) using the data C_{ik}, S_{ik} and M_{ik} above for $k = 1, 2$, following the procedure as discussed in Section 3.1.

Aggregate Data: However, if the clinic can not distinguish patients from different classes, then it can only collect the data at an aggregate level in the following format.

$C_i = C_{i1} + C_{i2}$: Number of patients from both classes who had an appointment delay of i days but canceled their appointments on or before their appointment days.

$S_i = S_{i1} + S_{i2}$: Number of patients from both classes who had an appointment delay of i days and showed up for their appointments.

$M_i = M_{i1} + M_{i2}$: Number of patients from both classes who had an appointment delay of i days, did not cancel in advance but missed their appointments.

In this case, the clinic treats patients of two classes as if they came from a pseudo single class since it does not have the information about their classes. Using the data C_i, S_i and M_i, the clinic can follow the statistical procedure in Section 3.1 to fit a same parametric model for patients' no-show and cancellation behaviors assuming that patients come from a single class. The clinic can not obtain the MLEs of the parameters for each class. Instead, it obtains the MLEs of the parameters for this pseudo single class of patients. We use the same notation in 3.1: let γ, a, θ, and b be the parameters for this pseudo single class and $\hat{\gamma}$, $\hat{a}$, $\hat{\theta}$, and $\hat{b}$ denote the MLEs for these parameters, respectively.

Table 4.3 summarizes the MLEs estimated using both the disaggregate and aggregate data collected under the three mixes of patients. As expected, if the clinic uses the aggregate data, the values of MLEs $\hat{\gamma}$, $\hat{a}$, $\hat{\theta}$, and $\hat{b}$ tend to increase as the average daily arrivals of class 1 patients increase and those of class 2 patients decrease since class 1 patients have lower probabilities to cancel or be a no-show, i.e. class 1 patients have higher values of parameters γ, a, θ and b. On the other hand, if the clinic uses the disaggregate data, the MLEs for class k patients tend to be more accurate, i.e. closer to their true values, when the average daily arrivals of class k patients are larger.

4.2.2 Findings of the Simulation Study

The objective of the simulation study is to compare the performances of the heuristics we propose with those of the other benchmark policies introduced in Section 3.2.2. We consider three mixes of patients, as described in Table 4.2, for the "model clinic." We assume that patients' no-show and cancellation behaviors follow parametric model (4.8) with parameters shown in Table 4.1. For the net reward, we assume the same setup used in Section 3.1.3, i.e. the net reward is in the form of (2.3) with $\eta(x) = x$ and the cost function $w(z)$ takes the form of (3.5).

If the clinic can only collect data at an aggregate level, then it operates as if patients came from

Mix	Aggregate Data		Disaggregate Data			
Mix I	$\hat{\gamma}$	0.9329	$\hat{\gamma}_1$	0.9537	$\hat{\gamma}_2$	0.9478
	$\hat{a}$	0.9761	$\hat{a}_1$	0.9899	$\hat{a}_2$	0.9694
	$\hat{\theta}$	0.8365	$\hat{\theta}_1$	0.8903	$\hat{\theta}_2$	0.8987
	$\hat{b}$	0.9816	$\hat{b}_1$	0.9902	$\hat{b}_2$	0.9684
Mix II	$\hat{\gamma}$	0.9240	$\hat{\gamma}_1$	0.9473	$\hat{\gamma}_2$	0.9435
	$\hat{a}$	0.9836	$\hat{a}_1$	0.9905	$\hat{a}_2$	0.9699
	$\hat{\theta}$	0.8653	$\hat{\theta}_1$	0.9197	$\hat{\theta}_2$	0.9100
	$\hat{b}$	0.9874	$\hat{b}_1$	0.9897	$\hat{b}_2$	0.9697
Mix III	$\hat{\gamma}$	0.9399	$\hat{\gamma}_1$	0.9597	$\hat{\gamma}_2$	0.9245
	$\hat{a}$	0.9880	$\hat{a}_1$	0.9902	$\hat{a}_2$	0.9689
	$\hat{\theta}$	0.8864	$\hat{\theta}_1$	0.9107	$\hat{\theta}_2$	0.8940
	$\hat{b}$	0.9897	$\hat{b}_1$	0.9901	$\hat{b}_2$	0.9694

Table 4.3: MLEs Obtained from the Simulated Data

a single class. The clinic can choose to implement the HDPs, e.g. Imp-OAP and Imp-OTPSP, as introduced in Chapter 3. Notice that when implementing the HDPs, the clinic need first estimate parameters for the statistical model of patients' no-show and cancellation behaviors, i.e. calculate $\hat{\gamma}$, $\hat{a}$, $\hat{\theta}$, and $\hat{b}$ as we did in Table 4.3, then plug these parameters into formula (2.8) to compute indices when making scheduling decisions.

On the other hand, if the clinic gets data at the disaggregate level, then it has two options. First, the clinic may treat patients as if they came from a single class by converting the disaggregate data into the aggregate data, then follow the policies for the single-class patient scheduling. Second, the clinic can choose to use the Greedy-HDP proposed in Section 4.1. In this case, the clinic need first find parameters $\hat{\gamma}_k$, $\hat{a}_k$, $\hat{\theta}_k$ and $\hat{b}_k$ as we did in Table 4.3, and then use these parameters and equation (4.6) to compute the indices.

Besides the heuristics we propose, the clinic can also use OAP, TP, BSP and RSP as introduced in Chapter 3. Notice that the scheduling decisions under these four policies will not be affected by whether or not the clinic collects the data, since these policies do not take patient's no-shows and cancellations into account and the model parameters estimated from the data will not be used in these policies to make decisions. In particular, we assume that the threshold for TP is set to be 50, the average total daily appointment requests from patients of both classes.

For a fair comparison, we choose OAP as the initial policy to derive the HDP and the Greedy-

HDP. Recall that the HDP derived from OAP is Imp-OAP, as described in Section 2.5. To obtain the Greedy-HDP, we set OAP as follows: $p_{0k} = 1$ and $p_{jk} = 0$ for $j = 1, \ldots, T$ and $k = 1, 2$. We call the Greedy-HDP obtained upon OAP as Imp-OAP$^{\#}$.

As for the values of the parameters, we set $K = 0$ without loss of generality and $h_2 = 0.95$. We set $T = 10$ and simulated various scenarios by considering different combinations of values for M and h_1. Specifically, M took values in $\{45, 50\}$ and h_1 took values in $\{0.0, 0.2, 0.5\}$. We used the batch-means method. For each combination of parameters, we ran 11 batches, each batch consisting of 200 consecutive workdays. The first batch was used as the warm-up period. For each mix of patients, a total of 6 different scenarios were considered each with a different pair of values for M and h_1. Each scenario was simulated under each one of the six scheduling policies, i.e. OAP, Imp-OAP$^{\#}$, Imp-OAP, TP, BSP and RSP, and the long-run average net reward was computed. In order to facilitate comparison, we chose OAP as the benchmark policy and for every other policy computed the percentage improvement that would be obtained by using the policy as opposed to OAP. Finally, we determined the 95% confidence interval for the mean percentage improvement. The results of each scenario under three patient mixes are given in Tables 4.4, 4.5 and 4.6, respectively.

In each table, the first number for each scenario-policy pair is the mean percentage improvement while the second number is the half width of the 95% confidence interval for the mean. Therefore, if the first number is larger than the second number, that indicates the corresponding policy is superior than OAP at the 5% significance level. On the other hand, if the first number is negative and it is larger than the second number in absolute value, that implies the superiority of OAP. The cases where the numbers indicate superiority of one policy over the other (in either direction) at the 5% significance level are shown in bold face. Note that the comparison is inconclusive in only five cases for all patient mixes.

From Tables 4.4, 4.5 and 4.6, we observe that BSP and RSP perform the worst among all policies for all scenarios. In order to better compare other policies we also determined the best policies for each scenario under each patient mix separately, which are listed in Table 4.7. More specifically, as we did in Section 3.2, for each scenario we conducted paired t-tests for every possible pair of

		Imp-OAP$^{\#}$	Imp-OAP	TP
$M = 50$	$h_1 = 0$	$\mathbf{2.85\% \pm 0.73\%}$	$\mathbf{2.51\% \pm 0.64\%}$	$\mathbf{4.78\% \pm 0.49\%}$
	$h_1 = 0.2$	$\mathbf{2.73\% \pm 0.81\%}$	$\mathbf{2.53\% \pm 0.92\%}$	$\mathbf{5.00\% \pm 0.61\%}$
	$h_1 = 0.5$	$\mathbf{3.30\% \pm 1.59\%}$	$\mathbf{1.78\% \pm 1.45\%}$	$\mathbf{5.83\% \pm 1.45\%}$
$M = 45$	$h_1 = 0$	$\mathbf{0.64\% \pm 0.52\%}$	$-0.35\% \pm 0.51\%$	$\mathbf{-4.11\% \pm 0.74\%}$
	$h_1 = 0.2$	$\mathbf{1.04\% \pm 0.84\%}$	$-0.06\% \pm 0.73\%$	$\mathbf{-5.98\% \pm 0.95\%}$
	$h_1 = 0.5$	$0.71\% \pm 1.15\%$	$-0.26\% \pm 1.04\%$	$\mathbf{-13.34\% \pm 1.80\%}$
		BSP	RSP	
$M = 50$	$h_1 = 0$	$\mathbf{-31.30\% \pm 0.99\%}$	$\mathbf{-18.01\% \pm 1.14\%}$	
	$h_1 = 0.2$	$\mathbf{-35.95\% \pm 1.06\%}$	$\mathbf{-20.63\% \pm 1.32\%}$	
	$h_1 = 0.5$	$\mathbf{-53.82\% \pm 1.36\%}$	$\mathbf{-30.73\% \pm 2.14\%}$	
$M = 45$	$h_1 = 0$	$\mathbf{-25.06\% \pm 1.03\%}$	$\mathbf{-14.21\% \pm 1.08\%}$	
	$h_1 = 0.2$	$\mathbf{-29.84\% \pm 1.10\%}$	$\mathbf{-16.9\% \pm 1.32\%}$	
	$h_1 = 0.5$	$\mathbf{-48.61\% \pm 1.45\%}$	$\mathbf{-27.45\% \pm 2.34\%}$	

Table 4.4: Results of the Simulation Study (Patient Mix I) - (The first number indicates the mean percentage improvement of the corresponding policy over OAP, and the second number indicates the half width for the 95% confidence interval.)

policies and determined whether or not there is a statistical difference between their performances at a significance level of 0.05. For every scenario, we listed the policies whose performances are better than those of the others. Note that for some of the scenarios, there is more than one policy. That is because in those cases, the paired t-test is inconclusive meaning that the performances of the policies are not statistically different.

Table 4.7 suggests that, when the regular capacity is large, i.e. when $M = 50$, TP performs the best compared to other policies. Furthermore, as we see in Tables 4.4, 4.5 and 4.6, when the overtime capacity becomes more expensive, i.e. as h_1 increases, the improvement of TP over OAP is more significant under all patient mixes. This observation coincides with our intuition as well as our findings in the single-class patient scheduling in Chapter 3: TP performs well when the regular capacity is large.

When the regular capacity is small, i.e. $M = 45$, we see that Imp-OAP$^{\#}$ and Imp-OAP are the best two policies. However, it is not clear which one will outperform the other. One possible reason is that Imp-OAP$^{\#}$ is a greedy algorithm, which does not guarantee an exact policy improvement step. Thus the extend of the improvement of Imp-OAP$^{\#}$ over OAP is limited. On the other hand, though Imp-OAP does not use all available information of patients' no-shows and

		Imp-OAP$^{\#}$	Imp-OAP	TP
$M = 50$	$h_1 = 0$	$\mathbf{3.58\%} \pm \mathbf{0.54\%}$	$\mathbf{3.38\%} \pm \mathbf{0.59\%}$	$\mathbf{5.86\%} \pm \mathbf{0.68\%}$
	$h_1 = 0.2$	$\mathbf{3.57\%} \pm \mathbf{0.59\%}$	$\mathbf{2.37\%} \pm \mathbf{0.96\%}$	$\mathbf{6.33\%} \pm \mathbf{0.85\%}$
	$h_1 = 0.5$	$\mathbf{4.24\%} \pm \mathbf{0.91\%}$	$\mathbf{2.76\%} \pm \mathbf{1.77\%}$	$\mathbf{8.14\%} \pm \mathbf{1.58\%}$
$M = 45$	$h_1 = 0$	$\mathbf{1.15\%} \pm \mathbf{0.56\%}$	$\mathbf{1.46\%} \pm \mathbf{0.66\%}$	$-0.26\% \pm 0.59\%$
	$h_1 = 0.2$	$\mathbf{1.52\%} \pm \mathbf{0.87\%}$	$\mathbf{1.58\%} \pm \mathbf{1.04\%}$	$\mathbf{-3.96\%} \pm \mathbf{0.77\%}$
	$h_1 = 0.5$	$\mathbf{2.10\%} \pm \mathbf{1.73\%}$	$\mathbf{3.31\%} \pm \mathbf{1.48\%}$	$\mathbf{-9.25\%} \pm \mathbf{1.47\%}$
		BSP	RSP	
$M = 50$	$h_1 = 0$	$\mathbf{-23.04\%} \pm \mathbf{0.68\%}$	$\mathbf{-12.7\%} \pm \mathbf{0.60\%}$	
	$h_1 = 0.2$	$\mathbf{-26.4\%} \pm \mathbf{0.76\%}$	$\mathbf{-14.47\%} \pm \mathbf{0.66\%}$	
	$h_1 = 0.5$	$\mathbf{-39.19\%} \pm \mathbf{1.12\%}$	$\mathbf{-21.24\%} \pm \mathbf{0.93\%}$	
$M = 45$	$h_1 = 0$	$\mathbf{-16.14\%} \pm \mathbf{0.77\%}$	$\mathbf{-9.68\%} \pm \mathbf{0.52\%}$	
	$h_1 = 0.2$	$\mathbf{-19.49\%} \pm \mathbf{0.86\%}$	$\mathbf{-11.49\%} \pm \mathbf{0.62\%}$	
	$h_1 = 0.5$	$\mathbf{-32.51\%} \pm \mathbf{1.30\%}$	$\mathbf{-18.49\%} \pm \mathbf{1.02\%}$	

Table 4.5: Results of the Simulation Study (Patient Mix II) - (The first number indicates the mean percentage improvement of the corresponding policy over OAP, and the second number indicates the half width for the 95% confidence interval.)

cancellations (since it treats patients of two classes as one single class), it still takes into account patients' no-shows and cancellations in some aggregate way. Thus Imp-OAP may still make some improvements over OAP, though not guaranteed. TP is a simple choice which ignores patients' no-shows and cancellations altogether. When the regular capacity is large, this ignorance may not affect the system performance too much since the capacity can somehow absorb the randomness of demand. In other words, we can endure some waste of resources without hurting the system performance too much if the capacity is large. However, if this is not the case, it will be costly to neglect patients' no-shows and cancellations. On the contrary, one can make much smarter choices by taking into account patient behaviors, even if she only uses this information partially.

Here we arrive at a similar conclusion as we did in Chapter 3: One should seek simplicity if complexity does not bring any significant advantages. Therefore, it appears that if the capacity is large, simple policies like TP can perform quite well. On the other hand, if the regular capacity is relatively small compared to daily demand, then the heuristics we propose appear to be better than TP and other choices.

		Imp-OAP$^{\#}$	Imp-OAP	TP
$M = 50$	$h_1 = 0$	$\mathbf{4.24\%} \pm \mathbf{0.69\%}$	$\mathbf{3.72\%} \pm \mathbf{0.69\%}$	$\mathbf{6.19\%} \pm \mathbf{0.40\%}$
	$h_1 = 0.2$	$\mathbf{4.04\%} \pm \mathbf{0.53\%}$	$\mathbf{3.84\%} \pm \mathbf{0.48\%}$	$\mathbf{6.80\%} \pm \mathbf{0.52\%}$
	$h_1 = 0.5$	$\mathbf{5.60\%} \pm \mathbf{1.11\%}$	$\mathbf{4.71\%} \pm \mathbf{1.42\%}$	$\mathbf{9.14\%} \pm \mathbf{1.05\%}$
$M = 45$	$h_1 = 0$	$\mathbf{2.66\%} \pm \mathbf{0.58\%}$	$\mathbf{3.13\%} \pm \mathbf{0.63\%}$	$-\mathbf{1.74\%} \pm \mathbf{0.52\%}$
	$h_1 = 0.2$	$\mathbf{3.12\%} \pm \mathbf{0.78\%}$	$\mathbf{4.02\%} \pm \mathbf{0.68\%}$	$-\mathbf{2.81\%} \pm \mathbf{0.67\%}$
	$h_1 = 0.5$	$\mathbf{4.93\%} \pm \mathbf{1.30\%}$	$\mathbf{7.59\%} \pm \mathbf{0.96\%}$	$-\mathbf{6.94\%} \pm \mathbf{1.28\%}$
		BSP	RSP	
$M = 50$	$h_1 = 0$	$-\mathbf{23.04\%} \pm \mathbf{0.68\%}$	$-\mathbf{12.7\%} \pm \mathbf{0.60\%}$	
	$h_1 = 0.2$	$-\mathbf{17.95\%} \pm \mathbf{0.98\%}$	$-\mathbf{9.43\%} \pm \mathbf{0.79\%}$	
	$h_1 = 0.5$	$-\mathbf{25.92\%} \pm \mathbf{1.23\%}$	$-\mathbf{12.69\%} \pm \mathbf{1.24\%}$	
$M = 45$	$h_1 = 0$	$-\mathbf{8.57\%} \pm \mathbf{0.82\%}$	$-\mathbf{5.89\%} \pm \mathbf{0.52\%}$	
	$h_1 = 0.2$	$-\mathbf{10.56\%} \pm \mathbf{0.90\%}$	$-\mathbf{6.72\%} \pm \mathbf{0.67\%}$	
	$h_1 = 0.5$	$-\mathbf{18.20\%} \pm \mathbf{1.31\%}$	$-\mathbf{9.89\%} \pm \mathbf{1.31\%}$	

Table 4.6: Results of the Simulation Study (Patient Mix III) - (The first number indicates the mean percentage improvement of the corresponding policy over OAP, and the second number indicates the half width for the 95% confidence interval.)

Scenarios		Mix I	Mix II	Mix III
$M = 50$	$h_1 = 0$	TP	TP	TP
	$h_1 = 0.2$	TP	TP	TP
	$h_1 = 0.5$	TP	TP	TP
$M = 45$	$h_1 = 0$	Imp-OAP$^{\#}$	Imp-OAP$^{\#}$ Imp-OAP	Imp-OAP
	$h_1 = 0.2$	Imp-OAP$^{\#}$	Imp-OAP$^{\#}$ Imp-OAP	Imp-OAP
	$h_1 = 0.5$	Imp-OAP$^{\#}$ Imp-OAP	Imp-OAP$^{\#}$ Imp-OAP	Imp-OAP

Table 4.7: "Best" Policies for Each Scenario at the 5% Significance Level (All Patient Mixes)

CHAPTER 5

Static Design of Appointment Scheduling Systems: The Model

5.1. Introduction

For many firms in the service sector, appointment scheduling systems help regulate the customer demand by giving customers the convenience to know “exactly” when they will be served. However, while customers enjoy this convenience to a large extent, most firms suffer from the fact that a high percentage of their customers do not show up for their scheduled appointments. Empirical studies for outpatient health care clinics find that the longer is the appointment delay for a patient, i.e. the time between the patient’s call for an appointment and her appointment date, the higher is the chance that the patient will be a no-show (see, e.g. Gallucci et al. (2005) and our findings in Chapter 3). Therefore, there is a strong incentive to reduce customers’ waiting times before they receive service, or equivalently, to prevent the customer backlog queue from becoming too long. Given a fixed service capacity, there are several ways that a service provider can control the backlog. For example, she can limit the size of the population receiving service (just as physicians do by controlling their panel sizes) or she only schedules new appointments when the backlog is sufficiently low (just as physicians do by not allowing appointments beyond a certain date into the future). The objective of this research is to provide insights on how to optimize system performance, particularly system throughput, for a service provider that employs these control methods.

Recent books and articles highlight the detrimental effects of no-shows in various service industries including restaurants (Strianese and Strianese (2003)), veterinarians (Trotta (2006)), and most importantly health care clinics (see, e.g. Pesata et al. (1999) and Gallucci et al. (2005)). Pesata et al. (1999) perform a descriptive study of missed appointments using data from a children's hospital. They note a pattern of missed appointments by the week of the month. They find that, the no-show rates of the four weeks of the month range from 10% to 42%. Gallucci et al. (2005) study a sample of around 6000 patients who were appointed to a psychiatry outpatient program, and estimate that the total rate of no-shows and cancellations is 12%, 23%, 42%, and 44% corresponding to 0, 1, 7, and 13 days of appointment delay respectively. In Chapter 3, we found the same rates to be 17.99%, 18.48%, 21.37%, and 24.15%, respectively, using the data from the Family Medicine Center at UNC (refer to Section 3.1.1 for details). Not surprisingly, the no-show rates reported among literature are different. However, these numbers all point to the same fact: patient no-shows are significant phenomena observed and experienced by the health care sector.

Patient no-shows not only cause inconvenience in running a clinic/hospital, but more importantly, they can also lead to substantial loss of revenues and considerable waste of resources. Moore et al. (2001) study the time and financial effects of no-shows at a family practice residency clinic. They report that the revenue shortfalls of a family practice center under study due to patient no-shows can range from 3% to 14%. However, as reported in Langabeer (2007), the average hospital profit margins usually range from 4% to 7%. Thus the 3% to 14% revenue loss is in fact huge, and can be catastrophic to the financial standing of a practice. All these evidences suggest that patient no-shows are crucial problems faced by health care providers, and they can lead to serious consequences if not dealt with appropriately. Now, one corresponding question to ask is that how we can take into account patients' no-show behaviors while scheduling their appointments. This is the subject of this section: we study how to design an appointment scheduling system while considering patients' no-show behaviors.

More specifically, we model the appointment backlog as a single-server queue where new appointments join the backlog from the back of the queue. Motivated by empirical findings, we assume that customers do not show up for their appointments with probabilities that increase with their appointment delays, i.e. the times they spend in waiting for their appointments. The objective of the provider is to maximize the system throughput rate, the rate at which patients indeed show up and get served.

We make two different assumptions for service times. We assume that they are either deterministic or exponentially distributed. If service times are deterministic, several issues related to the way appointments are scheduled are automatically resolved since given the queue length (i.e., backlog) customers can be informed of their exact appointment times at the time of their requests for an appointment. However, this assumption leads to an analytically intractable model. If service times are exponentially distributed, then the queueing model is easier to analyze. But since the service times are random, customers cannot be told exactly when their appointments are. This makes modeling the appointment system a significant challenge and forces one to address the question of how one can inform the customers of their scheduled appointment times at the time of their requests. However, since the objective of this research is to address high-level system design questions, not how appointments should be scheduled, we ignore such modeling complications and simply assume that all customers show-up on time for their appointments if they do show-up at all.

In this chapter, we assume that service times are exponentially distributed to study the model analytically and prove the theoretical results which provide insights for designing appointment systems. Then we carry out extensive numerical studies in the next chapter to investigate the situation with deterministic service times. As our numerical study in Chapter 6 shows, the insights generated in this chapter under the exponential service time assumption continue to hold under the deterministic service time assumption.

We consider two controls to maximize the system throughput. First, we assume that we can put a limit on the system capacity, i.e. the size of the appointment backlog queue. On the one hand, if the system capacity is set too large, then it is more likely that the backlog queue will be long and customers will need to wait for a long time before receiving service. This inevitably increases customers' no-show rates. On the other hand, if one chooses a small capacity, then the probability that an arriving customer sees no room in the system becomes large and thus the chance of turning a new customer away is higher. This is the basic tradeoff we are dealing with if we choose to put a limit on the system capacity. In Section 5.3 and 5.4, we characterize the optimal system capacity and study how the optimal capacity changes as other model parameters change.

The second control we consider is managing the demand rate. We will discuss this in Section 5.5. The basic tradeoff under this control is that, for a fixed system capacity and a service rate, larger demand leads to a longer backlog appointment queue and thus larger no-show rates while lower demand results in higher frequency of the server being idle and thus inefficient utilization of resources. We show that the system throughput is either a strictly increasing or a unimodal function of the demand rate. Thus there exists an optimal demand rate that maximizes the system throughput. In Chapter 6, we study a special case where patients' show-up probabilities have a parametric form. Under this assumption, we obtain a closed-form expression for the optimal demand rate if we do not limit the size of the appointment backlog queue.

There is a rich operations research literature on appointment scheduling. Please refer to Section 2.2 for a comprehensive review. Although few articles considered customer no-shows, there has been an increasing interest in the last few years. Mercer (1960, 1973) is the first to consider a queue with customers scheduled to arrive at certain times but may not show up. Kaandorp and Koole (2007) propose a scheduling algorithm that optimizes an objective function which penalizes waiting times, idle time of the server, and tardiness and can also accommodate customer no-shows. Hassin and Mendel (2008) are interested in the optimal scheduling of a finite number of arrivals to a single server queue. Customers do not show up with a fixed probability and their objective is to minimize a weighted sum of expected customer waiting and server availability costs.

To the best of our knowledge, only two articles modeled appointment delay dependent no-show probabilities. Liu et al. (2009) develop heuristic dynamic appointment scheduling policies that explicitly take into account the fact that customers may cancel or not show up with probabilities that depend on their appointment delays. (Chapter 2 and 3 of this dissertation collectively present the results in Liu et al. (2009)). On the other hand, Green and Savin (2008) use a single-server queueing model in which each customer may not show up with a probability that depends on the queue length at the time of its service in order to estimate the ideal panel size for a clinic that employs the Open Access policy. Note that Green and Savin assume that no-show probabilities depend on the queue length at service times rather than the queue length at the time of arrivals, which would have been a more realistic choice. However, under the latter assumption, their model becomes analytically untractable and therefore they use queue length at service times as a proxy for the queue length at arrival points. In this paper, we consider a model that is similar to that of Green and Savin (2008). However, unlike Green and Savin, we assume that customers who do not show up do not join back at the end of the queue, which allows us to make customers' no-show probabilities dependent on the queue length at the time of their arrivals. The relatively simpler structure of our model also allows us to characterize the optimal arrival rate to the system, which provides insights on the "optimal" panel size for an outpatient clinic.

The rest of this chapter is organized as follows. We first describe our queueing model setup in Section 5.2. The optimal system capacity is studied in Section 5.3. We present the sensitivity analysis of the optimal system capacity in Section 5.4. Finally we discuss how to optimize the system throughput rate by controlling the demand rate in Section 5.5.

5.2. The Model

We model the appointment system as a single-server queue. Requests for appointments arrive according to a Poisson process with rate $\lambda > 0$ and each request is accepted if the number of outstanding appointments at the time of the request's arrival is less than $K \leq \infty$. Service times are independent and identically distributed with mean $1/\mu$ where $0 < \mu < \infty$. In this chapter, they are assumed to be exponentially distributed. In the next chapter, we will numerically investigate the situation where service times are deterministic.

Motivated by the empirical evidence, we assume that customers do not show-up for their appointments with probabilities increasing with their appointment delays, i.e. the times they spend waiting for their appointments. (We use the terms "increasing" and "decreasing" in the weak sense, i.e. increasing means nondecreasing and decreasing is equivalent to nonincreasing.) More specifically, we model this customer behavior as follows: For each customer n, there is a random time R_n that determines whether or not the customer will be a no-show. Customer n will show-up for her appointment if and only if her appointment delay is less than R_n. We assume that $\{R_n, n = 1, 2, \dots\}$ is a sequence of independent random variables with an identical probability distribution. For generality, we do not impose any assumptions on the distribution of R_n. Notice that the appointment delay of a customer solely depends on the backlog at the time she joins the backlog queue and how fast the server can work. Thus, without loss of generality, we assume that an incoming customer will show up for her appointment with probability $p_j(\mu)$ if there are j customers ahead of her upon her arrival and the service rate is μ, where $p_j(\mu) \geq p_{j+1}(\mu)$ and $p_j(\mu') \geq p_j(\mu)$, for $0 < \mu \leq \mu' < \infty$ and $j = 0, 1, 2 \dots$. To simplify the notation, we write $p_j(\mu)$ as p_j whenever this does not cause ambiguities. We assume that customers who do show up, show up on time. To avoid triviality, suppose that $p_0(\mu) > 0$, $0 < \mu \leq \infty$.

For $j \in \{0, 1, 2, \dots, K\}$, let π_j denote the steady-state probability that an arriving customer sees a backlog of size j. Define $T_K = T_K(\lambda, \mu)$ to be the throughput, i.e. the long-run average rate of customers who show up and get served. (Note that we will suppress the dependence of T_K on some or both of the parameters λ and μ when not necessary.) Then, we have

$$T_K = T_K(\lambda, \mu) = \begin{cases} 0 & K = 0 \\ \lambda \sum_{j=0}^{K-1} \pi_j p_j & 1 \leq K \leq \infty \end{cases} \tag{5.1}$$

where π_j is a function of λ, μ and K, and p_j is a function of μ.

In particular, if service times are exponentially distributed, then we know from the standard analysis of the $M/M/1/K$ queue that

$$\pi_j = \frac{\rho^j}{\sum_{i=0}^{K} \rho^i}, \quad j = 0, 1, \dots, K,$$

where $\rho = \lambda/\mu$ is the traffic intensity. We assume that $0 < \rho < 1$ if $K = \infty$ (see, e.g., Kulkarni (1995)). In this case (5.1) can be written as

$$T_K = T_K(\lambda, \mu) = \begin{cases} 0 & K = 0 \\ \lambda \frac{\sum_{j=0}^{K-1} \rho^j p_j}{\sum_{i=0}^{K} \rho^i} & 1 \leq K \leq \infty \end{cases} \tag{5.2}$$

The objective of the service provider is to maximize the throughput T_K. Recall the two controls discussed in the last section: the first is to limit the system capacity, i.e. the size of the appointment backlog queue, and the second is to control the demand rate, e.g. through limiting the population receiving the service. In our model setup, these two controls correspond to maximizing the system throughput with respect to model parameters K and λ. In the following sections, we investigate how the system throughput T_K changes with K and λ, obtain expressions for optimal values of K and λ, and study how these optimal values change as the other parameters change.

5.3. The Optimal System Capacity

Suppose that the service provider puts a limit $K \leq \infty$ on the system capacity, i.e., the number of outstanding appointments in the backlog queue. Recall that for the dynamic scheduling model in Chapter 2, the service provider has a fixed service capacity M for each day and a fixed scheduling horizon T, i.e. the maximum number of days into the future for which new appointments can be scheduled. Choosing K can be seen as choosing the scheduling horizon T. With this interpretation we have $K = MT$. Thus the analytical results presented in this section shed light into how to choose T for the service provider.

If all the customers showed up for their appointments, there would be no reason to limit the number of outstanding appointments and thus the optimal value for K would be infinity. However, when there exists the possibility of a customer no-show and the probability of that happening increases with the delay that the customer experiences, the service provider may choose not to accept all appointment requests especially if the customer demand is high. In the following, we assume that service times are exponentially distributed and prove a number of structural properties on the optimal value for K. In Chapter 6 we numerically test these findings under deterministic service times.

Suppose that the service times are exponentially distributed. Let

$$f_K(\lambda, \mu) = \begin{cases} 0, & K = 0, \\ \frac{\sum_{j=0}^{K-1} \rho^j p_j}{\sum_{i=0}^{K} \rho^i}, & 1 \leq K \leq \infty. \end{cases} \tag{5.3}$$

From (5.2), we have

$$T_K(\lambda, \mu) = \lambda f_K(\lambda, \mu).$$

In the following, for ease of exposition, we will sometimes suppress the dependence of $f_K(\cdot)$ in some of the parameters λ and μ. Denote the set of all nonnegative integers by $\mathbb{Z}_+$ and the set of all positive integers by $\mathbb{Z}^+$. Then, we can prove the following results (see the appendix for their proofs).

Theorem 2. *For fixed λ and μ, throughput T_K is a unimodal function of K over $K \in \mathbb{Z}_+$. Let $\mathcal{K}^*$ represent the set of all maximizers to T_K and $K^* = \sup \mathcal{K}^*$. Then, K^* maximizes T_K. Furthermore, $K^* = \infty$ if and only if $f_{K-1}(\lambda, \mu) \leq \frac{p_{K-1}}{\rho}, \forall K \in \mathbb{Z}^+$. Otherwise,*

$$K^* = \sup\{K : K \in \mathcal{S}\} < \infty \tag{5.4}$$

where

$$\mathcal{S} = \{K : f_{K-1}(\lambda, \mu) \leq \frac{p_{K-1}}{\rho}, K \in \mathbb{Z}^+\}, \tag{5.5}$$

if and only if there exists a finite K such that $f_{K-1}(\lambda, \mu) > \frac{p_{K-1}}{\rho}$.

Theorem 2 provides an expression for the optimal capacity. Notice that there can be multiple optimal capacities and K^* is the supremum of these optimal capacities. As a corollary of Theorem 2, the next result explores the properties of K^*.

Corollary 1. If K^* is finite, then $p_{K^*} < p_{K^*-1}$.

Corollary 1 implies that, if there exists a finite K such that $f_{K-1}(\lambda, \mu) > \frac{p_{K-1}}{\rho}$, we need not consider all positive integers to find the optimal capacity K^*. Instead, we only need to consider the set of backlog queue lengths seen by an arriving customer where her show-up probabilities strictly drop. To be more specific, we denote this set by $\mathcal{K}$ and it is defined as

$$\mathcal{K} = \{j : p_j < p_{j-1}, j = 1, 2, 3 \dots\}.$$

For example, if $p_0 = 1$ and $p_j = 0.5, \forall j \geq 1$, then $\mathcal{K} = \{1\}$. Now, define

$$\mathcal{S}^* = \{K : f_{K-1}(\lambda, \mu) \leq \frac{p_{K-1}}{\rho}, K \in \mathcal{K}\}. \tag{5.6}$$

The next corollary shows that in order to find K^*, one only needs to find the supremum of $\mathcal{S}^*$.

Corollary 2. If there exists a finite K such that $f_{K-1}(\lambda, \mu) > \frac{p_{K-1}}{\rho}$, then $K^* = \sup\{K : K \in \mathcal{S}^*\}$, where K^* is defined in (5.4).

Corollary 2 simplifies the process of identifying the optimal capacity K^*. Instead of checking if all positive integers belong to set $\mathcal{S}$, one now only needs to consider the integers in set $\mathcal{K}$ and verify whether they belong to set $\mathcal{S}$. Notice that the optimal backlog size can possibly be infinity. To avoid using a "brute-force" way to find K^*, the following two corollaries provide simple tests to check whether K^* will be infinite.

These tests are built on the limits of the sequences $\{p_K/\rho\}_{K=0}^{\infty}$ and $\{f_K(\lambda, \mu)\}_{K=0}^{\infty}$. We first examine their existences. First notice that $\lim_{K\to\infty} p_K/\rho$ exists since $\{p_j\}_{j=0}^{\infty}$ is a bounded decreasing sequence. Now, if $\rho < 1$, one can also prove the existence of $\lim_{K\to\infty} f_K(\lambda, \mu)$ by showing that $\{f_K(\lambda, \mu)\}_{K=0}^{\infty}$ is a Cauchy sequence. Else if $\rho \geq 1$, it is easy to see that $\lim_{K\to\infty} f_K(\lambda, \mu) = \lim_{K\to\infty} p_K/\rho$. For convenience, we write $p_\infty = \lim_{K\to\infty} p_K$ and $f_\infty(\lambda, \mu) = \lim_{K\to\infty} f_K(\lambda, \mu)$. The following corollary provides a way to verify, under the situation when $\rho < 1$, whether the optimal capacity is infinity or not by comparing p_∞/ρ and $f_\infty(\lambda, \mu)$.

Corollary 3. If $\rho < 1$, $K^* = \infty$ if and only if

$$\frac{p_\infty}{\rho} \geq f_\infty(\lambda, \mu).$$

However, the above statement does not hold if $\rho \geq 1$. When $\rho \geq 1$, then $f_\infty(\lambda, \mu) = p_\infty/\rho$, but K^* still can be finite. Consider the following case: if $\rho = \lambda/\mu \geq 1$, $p_0 > 0$ and $p_j = 0, \forall j = 1, 2, 3 \ldots$, then $f_2(\lambda, \mu) > p_1/\rho = 0$. Applying Corollary 2 yields that $K^* = 1 < \infty$.

As a direct result form Corollary 3 and the analysis above, the subsequent corollary provides a condition on the sequence $\{p_j\}_{j=0}^{\infty}$ which implies that the optimal capacity is finite, regardless of the value of ρ.

Corollary 4. If $p_\infty = 0$, then $K^* < \infty$.

Corollary 3 and 4 are particularly useful when p_j has a parametric form and p_∞ and $f_\infty(\lambda, \mu)$ are easy to compute. We will illustrate this when we study the special case in Section 6.1.

5.4. Sensitivity Analysis of the Optimal Capacity

In this section, we investigate how the optimal capacity K^* changes with other model parameters, λ, μ, and p_j's. Recall that p_j's depend on μ. For convenience, we assume that $p_j(\mu)$ is a differentiable function of μ for all j. In order to study how K^* changes with μ, we need some further conditions on the limiting behaviors of p_j as μ approaches zero and infinity.

(C1) $\lim_{\mu\to 0} p_0(\mu) > 0$,

(C2) $\lim_{\mu\to 0} p_j(\mu)/\mu^j < \infty$, $j = 1, 2, \ldots,$

(C3) $\lim_{\mu\to\infty} p_j(\mu) > 0$, $j = 0, 1, 2, \ldots$.

Condition (C1) implies that any patient who sees no one ahead of her will show up with some positive probabilities regardless of the service rate. Condition (C2) imposes some Lipschitz conditions on p_j. Condition (C3) tells that any customer will show up with some positive probabilities if the service rate is fast enough regardless of the number of patients ahead of her upon her arrival. Next, we define the following function for a fixed λ.

$$g_K(\mu) = p_K \sum_{j=0}^{K} \rho^j - \rho \sum_{j=0}^{K-1} \rho^j p_j. \tag{5.7}$$

In order to prove the monotonicity results presented in Theorem 3 subsequently, we need one more technical condition on $g_K(\mu)$.

(C4) For a fixed λ, $g_K(\mu) = 0$ has exactly one root on $(0, \infty)$.

We mark that, $g_K(\mu)$ has the same sign of $f_{K+1}(\lambda, \mu) - f_K(\lambda, \mu)$. Thus condition (C4) ensures that for a fixed λ, $f_{K+1}(\lambda, \mu)$ and $f_K(\lambda, \mu)$ crosses exactly once over $(0, \infty)$. Furthermore, notice that $\lim_{\mu\to 0} g_K(\mu) = -\infty < 0$ by condition (C1) and (C2) and $\lim_{\mu\to\infty} g_K(\mu) = \lim_{\mu\to\infty} p_K(\mu) > 0$ by condition (C3). Hence condition (C4) warrants that there exists a threshold value, denoted as μ_K, such that $g_K(\mu) < 0$ if $\mu < \mu_K$ and $g_K(\mu) > 0$ if $\mu > \mu_K$. The proof of Theorem 3 is based on this fact. Though condition (C4) looks technical, it is not difficult to be verified especially when p_j has a parametric form. We will illustrate how one can do this by one example in Chapter 6.

Next we present the results on the sensitivity of K^* with respect to λ and μ in the following theorem. The proof is long and technical, and can be found in the appendix.

Theorem 3. *If the service rate μ and patient show-up probabilities $\{p_j\}_{j=0}^{\infty}$ are fixed, the optimal capacity for the outstanding appointment queue, K^*,*
(i) decreases with the arrival rate λ.
If λ is fixed and conditions (C1), (C2), (C3) and (C4) hold, then K^*
(ii) increases with the service rate μ.

Theorem 3 provides insights regarding how the service provider should adjust the limit put on the appointment backlog as conditions change. Part (i) states that as the demand for appointment requests increases, the service provider should be stricter and allow fewer number of outstanding appointments. This might sound counterintuitive at first. In response to an increase in demand, one might be tempted to allow more appointments to benefit from the increase. However, in fact increase in demand is all the more reason to limit the size of the backlog queue. Higher demand means less need to accumulate customers in the queue since the service provider has less trouble to fill in the empty spots in the appointment queue. This suggests that for service providers that are in high demand, there is less incentive to offer appointments for too far into the future. Part (ii) of Theorem 3 states that as the service rate increases (i.e., as more appointments are offered on each day), the service provider should be less strict and allow more outstanding appointments since when more customers are served per unit time, there is less waiting and thus the service provider can "afford" to admit more customers.

Next we discuss how the optimal capacity changes when patient show-up probabilities alter. Consider two sequences of show-up probabilities: $\{p_j\}_{j=0}^{\infty}$ and $\{p'_j\}_{j=0}^{\infty}$. For fixed λ and μ, let K^* and K' be the optimal capacity computed by (5.4) based on $\{p_j\}_{j=0}^{\infty}$ and $\{p'_j\}_{j=0}^{\infty}$, respectively. The following theorem states the conditions on show-up probabilities under which $K^* \geq K'$ (see the appendix for its proof).

Theorem 4. *For fixed λ and μ, if*

$$\frac{p_j}{p_{j+1}} \leq \frac{p'_j}{p'_{j+1}}, \quad \forall j = 0, 1, 2 \ldots, \tag{5.8}$$

then $K^* \geq K'$.

For ease of interpretation, we call patients having show-up probabilities $\{p_j\}_{j=0}^{\infty}$ and $\{p'_j\}_{j=0}^{\infty}$ as patients of Type 1 and Type 2, respectively. Condition (5.8) implies that, if we let patients wait for one more slot, the show-up probability of Type 2 patients has a sharper drop in ratio than that of Type 1 patients. In other words, Type 2 patients are more sensitive to the change in their waiting times than Type 1 patients. Theorem 4 implies that, if patients become more impatient, the service provider should choose a smaller system capacity for the outstanding appointment queue (or use a shorter scheduling horizon if we follow the terminologies in Chapter 2). The reason behind this choice is that since customers have higher no-show probabilities, we should prevent them from joining the queue when it is too long.

5.5. Optimal Control of the Appointment Demand

Suppose that the service provider is required to operate with a fixed limit $K \leq \infty$ on the number of outstanding appointments but can choose the arrival rate λ of new appointment requests. In practice, the queue capacity K can easily be controlled but control of the arrival rate is in general more difficult. Nevertheless, the service provider can control the appointment demand by limiting the customer pool that the service will be provided to and not accepting any requests coming from customers outside the pool. For example, firms that provide subscription-based services (such as those that offer software or hardware maintenance services) can control the size of their customer pool by not selling subscriptions once the number of subscribers reaches a certain level. Another example is the outpatient clinics. As in the model of Green and Savin (2008), one can view the queueing system as an outpatient clinic attended to by a single physician, who controls the arrival rate by choosing her panel size. Since the physician accepts appointment requests that only come from her panel, choosing the panel size helps control the appointment arrival rate to the backlog queue. More specifically, Green and Savin assume that each customer submits jobs (i.e., each patient requests an appointment) with the same fixed rate so that there is a linear relationship between the panel size and the arrival rate. Assuming that each customer's job submission rate is λ_0, choosing λ is equivalent to choosing the panel size to be λ/λ_0.

In our model, suppose that there is a finite limit $K < \infty$ on the appointment backlog queue. If there are K appointments already scheduled, then incoming appointment requests are rejected. We can show that throughput is either an increasing or a unimodal function of the arrival rate. Thus there exists an optimal arrival rate (possibly infinity) which maximizes the system throughput rate.

Theorem 5. *For $K < \infty$ and for any fixed value of $0 < \mu < \infty$ and show-up probabilities, throughput $T_K(\lambda, \mu)$ as defined in (5.2) is either a strictly increasing or a unimodal function of the arrival rate λ.*

The proof of Theorem 5 appears in the appendix. From the proof, we know that when $K = 1$, then throughput $T_K(\lambda, \mu)$ is strictly increasing in the arrival rate λ and thus the optimal arrival rate is infinity. On the other hand, there exists a unique finite arrival rate, λ^*, which maximizes $T_K(\lambda, \mu)$ for a fixed $\mu > 0$, if and only if there exists a finite λ^* which satisfies the first order condition, i.e.

$$\frac{\partial T_K(\lambda, \mu)}{\partial \lambda}\Big|_{\lambda=\lambda^*} = 0.$$

CHAPTER 6

Static Design of Appointment Scheduling Systems: Special Cases

In Chapter 5 we consider a queueing system where customers may not show up for their appointments with probabilities increasing in their appointment delays. Recall that customer n will show up for her appointment if and only if her appointment delay, i.e. her waiting time in the queue before receiving service, is less than her tolerance time R_n, where $\{R_n, n = 1, 2, \dots\}$ is a sequence of i.i.d. random variables.

In this chapter, we consider a special case where R_n is assumed to follow an exponential distribution, and conduct numerical studies to investigate systems with deterministic service times. There are two purposes of studying this special case. (i) We illustrate how one can use the analytical results proved under the general setup in Chapter 5, and in particular, we demonstrate how one can verify condition **(C4)**. Recall that **(C4)** is a technical condition required to prove that the optimal capacity for the outstanding appointment queue increases in the service rate if other model parameters are fixed (see Theorem 3). (ii) We derive a simple closed-form expression for the optimal demand rate for the situation where the system administrator does not limit the size of the outstanding appointment queue. This expression can be used to estimate the "optimal" panel size for clinics. After studying this special case, we conduct an extensive numerical study on systems with deterministic service times. Our numerical results suggest that the theoretical results proved in Chapter 5 by assuming service times are exponentially distributed also hold for situations where service times are deterministic.

6.1. A Special Case: Exponential Tolerance Times

The model is the same as the one introduced in Section 5.2 except that customers' tolerance times $\{R_n, n = 1, 2, \dots\}$ is assumed to be a sequence of independent random variables that are exponentially distributed with mean $1/\theta$ such that $0 < \theta < \infty$. We call θ the "no-show rate." We follow the same notation in Chapter 5: an incoming customer will show up for her appointment with probability p_j if there are j customers ahead of her upon her arrival. Notice that in this setup, p_j also depends on μ and θ. Using the memoryless property of the exponential distribution, one can show that

$$p_j = (\frac{\mu}{\mu+\theta})^j \tag{6.1}$$

for $j \in \{0, 1, 2 \dots\}$. This immediately implies that p_j is strictly decreasing in j and strictly increasing in μ, satisfying the conditions on p_j as stated in Section 5.2. Furthermore, we note that p_j is strictly decreasing in θ, as expected.

Under the assumption of exponentially distributed service times, (5.2) can be written as

$$T_K = T_K(\lambda, \mu, \theta) = \begin{cases} 0 & K = 0 \\ \lambda \frac{\sum_{j=0}^{K-1} \gamma^j}{\sum_{i=0}^{K} \rho^i} = \lambda \frac{(1-\rho)(1-\gamma^K)}{(1-\gamma)(1-\rho^{K+1})} & 1 \leq K < \infty \\ \lambda \frac{\sum_{j=0}^{\infty} \gamma^j}{\sum_{i=0}^{\infty} \rho^i} = \lambda \frac{1-\rho}{1-\gamma} & K = \infty \end{cases} \tag{6.2}$$

where

$$\gamma = \frac{\lambda}{\mu+\theta}.$$

Notice that the throughput T_K is now a function of λ, μ, and θ. We assume that $0 < \rho < 1$ if $K = \infty$.

6.1.1 The Optimal System Capacity and Sensitivity Analysis

In this section we consider the optimal system capacity for this special case, denoted as K_e^* (where the subscript "e" stands for "exponential"). Theorem 2 is directly applicable since for fixed

λ, μ and θ, p_j is strictly decreasing in j. First, Corollary 4 implies that

$$K_e^* < \infty,$$

since

$$\lim_{K\to\infty} p_K = \lim_{K\to\infty} (\frac{\mu}{\mu+\theta})^K = 0.$$

Thus the optimal system capacity K_e^* can be characterized as follows:

$$K_e^* = \sup\{K : K \in \mathcal{S}_e\}$$

where

$$\mathcal{S}_e = \{K : f_{K-1}(\lambda, \mu, \theta) \le \frac{1}{\rho}(\frac{\gamma}{\rho})^{K-1}, K \in \mathbb{Z}^+\}.$$

Next, we investigate how this optimal capacity K_e^* changes with other model parameters, λ, μ, and θ. We summarize the results in the following theorem. In its proof, we will illustrate how one can verify condition (C4), a technical condition required to prove Part (ii) in Theorem 3 (refer to Section 5.4 for more details).

Theorem 6. *The optimal capacity for the outstanding appointment queue, K_e^*,*
(i) decreases with the arrival rate λ,
(ii) increases with the service rate μ,
(iii) decreases with the "no-show rate" θ.

Proof: Part (i) and (iii) directly follow from Theorem 3 part (i) and Theorem 4, respectively. In order to prove part (ii), we only need to verify conditions (C1), (C2), (C3) and (C4) as stated in Section 5.4. The verification of the first three conditions is trivial.
(C1) $\lim_{\mu\to 0} p_0(\mu) = 1 > 0$,
(C2) $\lim_{\mu\to 0} p_j(\mu)/\mu^j = \theta^{-j} < \infty$, $j = 1, 2, \ldots$,
(C3) $\lim_{\mu\to\infty} p_j(\mu) = 1 > 0$, $j = 0, 1, 2, \ldots$.
Next we show how to verify condition **(C4)**. Recall that **(C4)** is specified as follows.

(C4) For a fixed λ (and θ), $g_K(\mu) = 0$ has exactly one root on $(0, \infty)$, where

$$g_K(\mu) = p_K \sum_{j=0}^{K} \rho^j - \rho \sum_{j=0}^{K-1} \rho^j p_j$$

as defined in equation (5.7), where $\rho = \lambda/\mu$.

We first note that $\lim_{\mu \to 0} g_K(\mu) = -\infty < 0$ by condition (C1) and (C2), and $\lim_{\mu\to\infty} g_K(\mu) = \lim_{\mu\to\infty} p_K(\mu) > 0$ by condition (C3). Thus $g_K(\mu) = 0$ has at least one root on $(0, \infty)$. Denote the smallest root by μ_K and let $g'_K(\mu)$ represent the first order derivative of $g_K(\mu)$ with respect to μ. In order to verify the uniqueness of μ_K, it suffices to show that $g_K(\mu)$ is strictly increasing on $(0, \infty)$, i.e. $g'_K(\mu) > 0$ for all μ such that $0 < \mu < \infty$. Following this idea, we first provide an explicit expression for $g'_K(\mu)$.

$$g'_K(\mu) = K\frac{\mu^{K-1}\theta}{(\mu+\theta)^{K+1}}\sum_{j=0}^{K}(\frac{\lambda}{\mu})^j - (\frac{\mu}{\mu+\theta})^K\sum_{j=0}^{K} j\frac{\lambda^j}{\mu^{j+1}} + \frac{\lambda}{\mu^2}\sum_{j=0}^{K-1}(\frac{\lambda}{\mu+\theta})^j + \frac{\lambda}{\mu}\sum_{j=0}^{K-1} j\frac{\lambda^j}{(\mu+\theta)^{j+1}}.$$

Then, it follows that

$$\begin{aligned}
\mu g'_K(\mu) &= K\frac{\mu^{K}\theta}{(\mu+\theta)^{K+1}}\sum_{j=0}^{K}(\frac{\lambda}{\mu})^j - (\frac{\mu}{\mu+\theta})^K\sum_{j=0}^{K} j(\frac{\lambda}{\mu})^j + \frac{\lambda}{\mu}\sum_{j=0}^{K-1}(\frac{\lambda}{\mu+\theta})^j \\
&\quad + \sum_{j=0}^{K-1} j\frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}} \\
&> -(\frac{\mu}{\mu+\theta})^K\sum_{j=0}^{K} j(\frac{\lambda}{\mu})^j + \frac{\lambda}{\mu}\sum_{j=0}^{K-1}(\frac{\lambda}{\mu+\theta})^j + \sum_{j=0}^{K-1} j\frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}} \\
&= -(\frac{\mu}{\mu+\theta})^K\sum_{j=0}^{K} j(\frac{\lambda}{\mu})^j + (\frac{\mu}{\mu+\theta})^K\sum_{j=0}^{K}(\frac{\lambda}{\mu})^j + \sum_{j=0}^{K-1} j\frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}} \\
&> -(\frac{\mu}{\mu+\theta})^K\sum_{j=1}^{K} j(\frac{\lambda}{\mu})^j + (\frac{\mu}{\mu+\theta})^K\sum_{j=1}^{K}(\frac{\lambda}{\mu})^j + \sum_{j=0}^{K-1} j\frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}} \\
&= -(\frac{\mu}{\mu+\theta})^K\sum_{j=0}^{K-1} j(\frac{\lambda}{\mu})^{j+1} + \sum_{j=0}^{K-1} j\frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}} \\
&= -\sum_{j=0}^{K-1} j\frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}}(\frac{\mu}{\mu+\theta})^{K-(j+1)} + \sum_{j=0}^{K-1} j\frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}}
\end{aligned}$$

$$= \sum_{j=0}^{K-1} j \frac{\lambda^{j+1}}{(\mu+\theta)^{j+1}}[1-(\frac{\mu}{\mu+\theta})^{K-(j+1)}] \geq 0,$$

which immediately implies that $g_K(\mu) = 0$ has exactly one root on $(0, \infty)$ and thus condition (C4) holds. The proof is complete. ■

6.1.2 Optimal Control of the Appointment Demand

In this section, we consider the optimal demand rate for the special case. First, suppose that there is a finite limit $K < \infty$ on the appointment backlog queue, i.e. appointment requests will be rejected unless there are less than K appointments scheduled. The service provider's objective is to choose the demand rate so that throughput is maximized. Then, using Theorem 5 we immediately deduce that the throughput for this special case is either an increasing or a unimodal function of the arrival rate. We summarize this result in the following proposition.

Proposition 3. For $K < \infty$ and for any fixed value of $0 < \mu < \infty$ and $0 < \theta < \infty$, throughput $T_K(\lambda, \mu, \theta)$ as defined in (6.2) is either a strictly increasing or a unimodal function of the arrival rate λ.

Now, suppose that the service provider does not put any restrictions on the number of outstanding appointments (i.e., $K = \infty$) but determines the arrival rate. Assuming that there is no capacity restriction on the appointment backlog queue makes it possible to come up with an expression for the optimal arrival rate, which can then be used to determine the optimal customer pool size or the optimal panel size in the case of outpatient clinics. Using (6.2), we can show that there is a unique optimal arrival rate that is given by the unique solution to the first order condition. This result appears in Theorem 7 and its proof can be found in the appendix.

Theorem 7. *For $K = \infty$, and for any fixed μ and θ such that $0 < \mu < \infty$ and $0 < \theta < \infty$, the throughput given by* (6.2) *is a unimodal function of λ and there is a unique optimal arrival rate λ^*_∞ that maximizes the throughput. Furthermore,*

$$\lambda^*_\infty = (\mu + \theta) - \sqrt{(\mu + \theta)\theta}. \tag{6.3}$$

Using Theorem 7, we can also characterize the behavior of the optimal arrival rate with respect to the no-show parameter θ and the service rate μ. This result is summarized in Corollary 5 and its proof is trivial.

Corollary 5. When $K = \infty$, the unique optimal arrival rate λ_∞^* decreases with the no-show parameter and increases with the service rate, i.e.,

$$\frac{\partial \lambda_\infty^*}{\partial \theta} < 0, \frac{\partial \lambda_\infty^*}{\partial \mu} > \frac{1}{2} > 0.$$

According to Corollary 5, if customers are more prone to not showing up, the service provider chooses a smaller arrival rate (or a smaller customer population). She does not try to compensate a higher no-show rate by admitting more customers knowing that more customers will only lead to even more no-shows. On the other hand, if the service capacity increases, the service provider chooses a higher rate since more customers can be served while keeping the no-show rate relatively small.

6.1.3 An Approximation Scheme to Consider Rescheduling Patients

Theorem 7, in particular (6.3) can be used to come up with an estimate on the optimal customer pool size or patient panel size complementing a similar approach that Green and Savin (2008) proposed. Although Green and Savin also use single server queues with Poisson arrivals, there are some basic differences. Their objective is not to carry out an optimization but to come up with an estimate on the maximum panel size under which the Open Access policy is sustainable. They do this estimation by investigating how some of the key performance measures (expected appointment backlog and probability that an arriving patient is given a same-day appointment) change with λ and determine the arrival rates above which these measures indicate problems (high expected backlog, low same-day appointment probability) if Open Access were to be used. However, even though our objective and that of Green and Savin are not exactly the same, they are aligned to a certain extent. They increase the arrival rate as much as possible while still ensuring that Open Access policy is feasible. In our case, we do not worry about the applicability of Open Access. We are concerned with determining the arrival rate that will lead to the highest throughput be it under Open Access or another policy.

The important difference is in the way how the patient no-show behavior is formulated. Green and Savin assume a functional form for the no-show probability as a function of the backlog size at the time of the patient's entry to the queue and propose that the parameters of this function be estimated from the data. In fact, our no-show formulation can also be seen as assuming another functional form for the no-show probabilities and its parameter θ can also be easily estimated from the data. The most significant difference is that Green and Savin assume that patients who do not show up for their appointments join the queue from the very end with a certain probability arguing that some of the patients who fail to show up do so for personal reasons not because they do not need appointments anymore and thus make another appointment after the no-show. In our model, we do not allow rescheduling after no-shows. Allowing no-shows to join back at the queue with a certain probability is an important generality. However, it comes with a significant "cost" since it makes the model extremely complicated. In order to come up with an expression for the throughput, as Green and Savin also argue, one needs a complex formulation in which system state description contains not only the current backlog but also backlogs at the arrival time of each of the patients in the backlog queue. Therefore, exact analysis is practically not feasible and resorting to approximations appears to be a reasonable option. Green and Savin do precisely that by assuming that patients do not show-up with probabilities that depend on the backlog at the time of their service, not at the time of their arrivals. With this assumption, the system state can be represented by a single-dimensional formulation where the state simply is the current size of the backlog queue. Obviously, the analysis under this assumption only gives approximate values for the performance measures of interest but it appears to be a very reasonable approximation especially since the authors carry out steady-state analysis. Furthermore, the authors numerically demonstrate that this approximation is very reliable.

Although our formulation did not explicitly consider the possibility of no-show customers joining back in the queue, we can still use our results, particularly the optimal arrival rate expression (6.3) given in Theorem 7 to come up with an approximation for the optimal arrival rate for the case where customers do rejoin with a certain probability.

Using the notation of Green and Savin, suppose that r denotes the probability that a no-show customer rejoins the queue right away. Suppose that for a fixed arrival rate λ (including new arrivals as well as rejoining customers), $p(\lambda)$ denotes the probability that in the steady-state a random customer is a no-show. Then, the expected number of times that a new arrival will join the queue at a single visit is $\sum_{k=0}^{\infty}(rp(\lambda))^k = \frac{1}{1-rp(\lambda)}$. Now, even though that is not the case, in order to obtain an approximation, suppose that the arrival process to the queue (including new and rejoining customers) is Poisson. Then, we know that λ^*_∞ as defined in (6.3) is the optimal arrival rate. Our objective is to determine the arrival rate of new appointments $\hat{\lambda}$ so that the effective arrival rate to the queue is λ^*_∞. Thus, we must have

$$\lambda^*_\infty = \hat{\lambda}\frac{1}{1-rp(\lambda^*_\infty)}. \tag{6.4}$$

Since the effective arrival to the queue is approximated by a Poisson process with rate λ^*_∞, using PASTA and the standard analysis of the $M/M/1$ queue, the no-show probability can be shown to be (see, e.g., Kulkarni (1995))

$$p(\lambda^*_\infty) = \frac{\lambda^*_\infty \theta}{\mu(\mu+\theta-\lambda^*_\infty)}.$$

Using this together with (6.4), we obtain

$$\hat{\lambda} = \lambda^*_\infty - \frac{(\lambda^*_\infty)^2\theta r}{\mu(\mu+\theta-\lambda^*_\infty)} \tag{6.5}$$

where λ^*_∞ is as given by (6.3).

6.2. Numerical Studies

In this section we carry out extensive numerical experiments to study the cases where service times are deterministic. We assume that customers' tolerance times follow an exponential distribution with the "no-show" rate θ. For all the numerical studies we conducted, the results established in Theorem 2, 3, 4 and 5 based on the assumption of exponentially distributed service times also hold for the cases that assume deterministic service times. We select some of the numerical results and present them in the following.

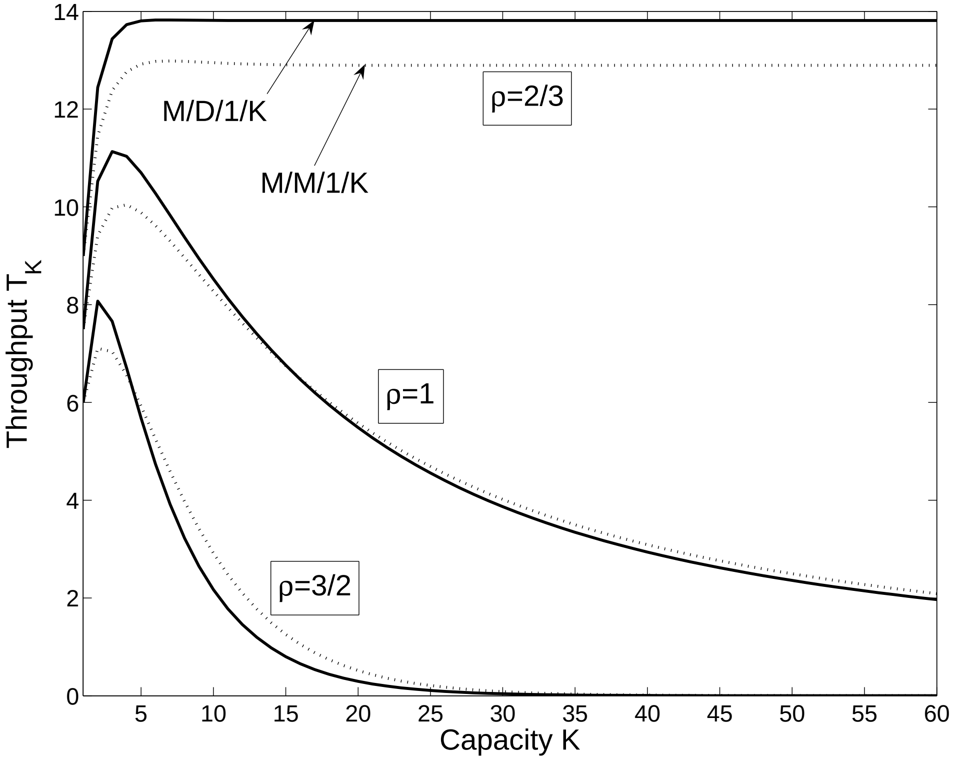

Figure 6.1: Throughput Rate T_K vs. System Capacity K

Figure 6.1 shows how the throughput rate changes with respect to system capacity K. We fix the arrival rate $\lambda = 15$ and the "no-show" rate $\theta = 2$ while varying the service rate μ, or equivalently the traffic intensity ρ. The solid lines represent the cases under the assumption of deterministic service times while the dotted lines exhibit the cases that assume exponentially distributed service times. We observe that the throughput rate T_K is a unimodal function of the capacity K for all cases.

Next, we study the sensitivity of K^* with respect to the arrival rate λ, the service rate μ and the "no-show" rate θ. We consider the following combination of model parameters: $\lambda = 15$, $\mu = 10$ and $\theta = 2$. When studying the sensitivity of K^* with respect to one particular model parameter, say λ, we fix the other two and vary λ. Figure 6.2 and Table 6.1 summarize the results. Similar to

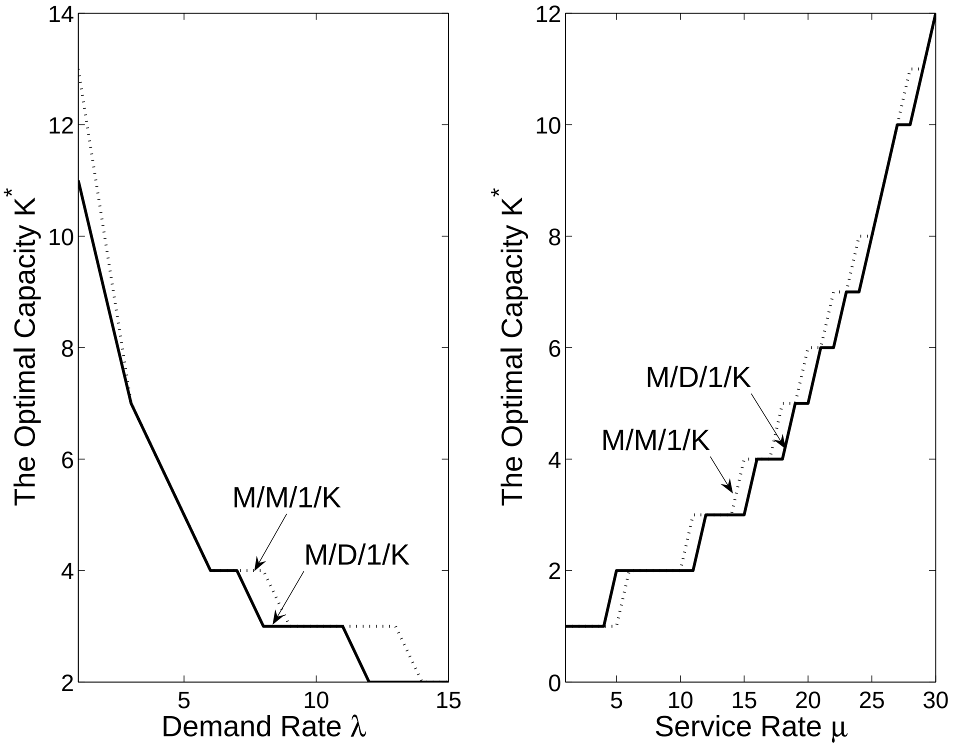

Figure 6.2: The Optimal Capacity K^* vs. Demand Rate λ and Service Rate μ

Figure 6.1, Figure 6.2 uses solid lines to represent the deterministic service time cases while uses dotted lines to show the cases with exponentially distributed service times. From Figure 6.2 and Table 6.1, we observe that K^* decreases in λ, increases in μ and decreases in θ regardless of the assumptions on service times.

K^*	M/M/1/K	M/D/1/K
3	θ=1	θ=1
2	θ=2, 3, 4, 5, 6	θ=2, 3, . . . , 11
1	θ=7, 8, . . . , 20	θ=12, 13, . . . , 20

Table 6.1: The Optimal Capacity K^* vs. "No-show Rate" θ

Finally, we study how system throughput T_K changes in demand rate λ if other model parameters are fixed. We first set the service rate $\mu = 15$ and the "no-show rate" $\theta = 2$, and then vary demand rate λ from 1 to 30 for $K = 3$, 5 and 10, respectively. We compute the system throughput for different combinations of parameters and Figure 6.3 summarizes the results. Similar to the figures above, we use solid lines and dotted lines to represent the situations with deterministic service times and with exponentially distributed service times, respectively. We observe that under both service time assumptions, the throughput is a unimodal function of the demand rate for all values of K tested, and thus there exists a unique optimal demand rate that maximizes the system throughput.

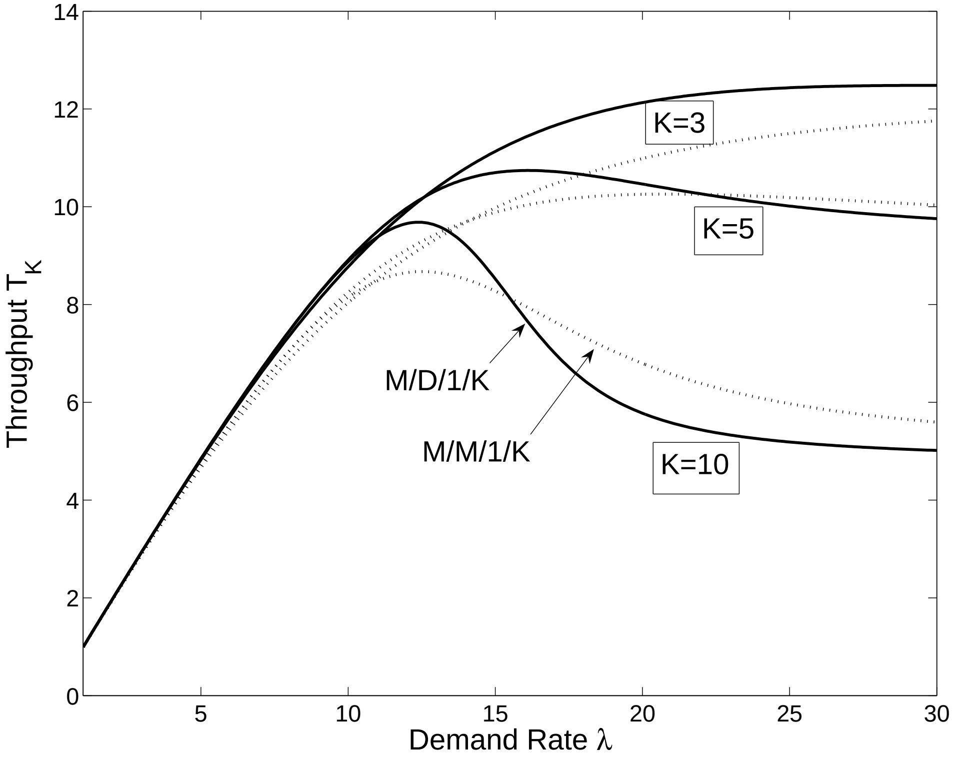

Figure 6.3: Throughput Rate T_K vs. Demand Rate λ

CHAPTER 7

Conclusion

7.1. Summary

As many practitioners would agree and as supported by research including our own, the Open Access policy can be an ideal solution to the appointment scheduling problem if patient demand and appointment supply can be kept "in balance." The problem, however, is that as we discuss in detail in Section 2.1, keeping that balance may not be possible for many clinics. There are already physician shortages in certain parts of the United States, which are expected to go worse in the next few years. In addition, as the Massachusetts experience suggests, if "universal health care" is widely adopted, increasingly more clinics will have to deal with high patient loads. Therefore, it is necessary to design efficient appointment scheduling systems for clinics that handle high patient loads, and as a result suffer from high no-show and cancellation rates. In this dissertation, we propose two complementary approaches to solve patient scheduling problems in health care, especially for those which can not implement OA. The first approach is to dynamically schedule patients based on the current appointment schedule; the other one is to statically design the system by managing the appointment backlog or regulating the demand rate. We briefly summarize our work in the following.

In Chapter 2 we develop an MDP model for the dynamic appointment scheduling decisions for a clinic that explicitly takes into account the possibility that patients may cancel or not show up for their appointments with probabilities that increase with their appointment delays. The model parameters can easily be estimated from data typically available for most clinics. However, any realistically sized problem can not be solved to optimality using the standard numerical procedures like the policy improvement algorithm or the value iteration algorithm, due to its large state space. Thus we develop dynamic heuristic policies using the idea of applying one step of the policy improvement algorithm on a "good" initial policy, and our proposed heuristics are very easy to implement.

In Chapter 3 we test the performances of the heuristic policies in a simulation study where we consider a "model clinic" generated by data provided to us by the Department of Family Medicine at the University of North Carolina. The results of the study clearly indicate that the heuristic policies we propose perform quite well particularly when the patient load is high. The policies outperform the Open Access policy even when the load is low, but the differences in the performances are not as significant and thus the Open Access policy might be a more reasonable choice if one wishes to keep the scheduling simple. However, for high patient loads, the policies we propose appear to be much better than the Open Access policy as well as the other benchmark heuristics.

In Chapter 4, we consider the heterogeneity of no-show behaviors among patients by grouping them into different classes according to their characteristics that correlate with their no-show behaviors, as in the model of Muthuraman and Lawley (2008). We extend the MDP model developed in Chapter 2 to take into account the patient heterogeneity. Due to its significantly enlarged state space, we derive a greedy algorithm for this problem using the idea of policy improvement heuristics. In order to evaluate the performances of different policies, we conduct an extensive simulation study using simulated data. The numerical results suggest that the proposed heuristics perform well when the regular capacity is small.

In Chapter 5, we model the appointment scheduling system as a single-server queue where new appointment requests join the backlog from the back of the queue. We consider two controls to maximize the system throughput rate: one is to put a limit on the appointment backlog and the other is to control the demand rate. We assume the service times to be exponentially distributed and derive a series of theoretical results. First, we provide an expression for the optimal size of appointment backlog and study how it changes as other model parameters change. Second, we show that the system throughput is either a strictly increasing or a unimodal function of the demand rate. Thus there exists a unique optimal demand rate that maximizes the system throughput.

In Chapter 6, we study a special case and conduct extensive numerical studies for the single-queueing model considered in Chapter 5. In the special case, we assume a specific parametric form for customers' no-show probabilities. We illustrate how the theoretical results obtained in Chapter 5 can be applied. Furthermore, we obtain a simple closed-form expression for the optimal arrival rate, assuming that we do not put a limit on the appointment backlog. In the numerical study, we assume the service times to be deterministic, and observe that all insights generated under the assumption of exponential service times remain valid. Thus these insights can still be used to direct the design of appointment scheduling systems when service times are deterministic.

It is important to note that in this dissertation we use the term "Open Access" (OA) strictly to refer to the policy of scheduling all the appointments for the day when the requests are made. In practice, OA is sometimes interpreted or implemented in a more flexible way. For example, some clinics that use OA allow a certain portion of their patients to make appointments for the future. Some clinics use open access mostly as a guiding principle rather than a specific prescription of how appointments should be scheduled. On the other hand, some others implement OA but maintain more flexibility by scheduling appointments over a period of 2-5 days rather than a single day. Such more flexible implementations of OA, where appointments are scheduled over a number of days, fit into our dynamic scheduling model. Thus the policy improvement heuristics we propose can be seen as not necessarily alternatives to open access in the broader meaning of the term, but policies that can also be used by clinics who are implementing some of the more flexible versions of OA.

7.2. Future Research Directions

There are many ways to extend our work. We briefly discuss some future research directions here.

Sequential Decision Model: First, our policy improvement heuristics specify on which day to schedule an incoming appointment request, but not the time of the appointment on that day. It will be interesting to develop a sequential decision model to jointly decide on which day and at what time to schedule appointments. To the best of our knowledge, no prior work has investigated that problem. One possibility is to merge the existing slot scheduling algorithms already developed in the literature with our heuristic methods. More precisely, one can use the methods we propose to determine the day of the appointment and use a slot scheduling algorithm (such as the one proposed by Muthuraman and Lawley (2008)) to determine the time of the appointment. Obviously, performance of such policies should be investigated with extensive numerical and simulation studies.

Patient Preferences: Another important yet challenging extension would be to incorporate the patients' appointment time preferences into account. In the dynamic scheduling model, we assumed that patients accept the first appointment date offered to them, which can be a reasonable assumption in many cases. However, in practice not all patients will be satisfied with the day offered to them and some will ask for another day. In fact, the heuristics we propose can be used even when the patients have preferences since they do not simply tell us on which day the next patient should be scheduled for. They also provide a ranking of alternative appointment dates since the index that the heuristics calculate for each day can be seen as a heuristic score of how much the clinic should rather assign that day instead of the others. Thus, if a patient is not happy with the day offered, the scheduler can offer the day with the next highest score. However, the performance of such a policy should be investigated carefully since the heuristics are in fact derived under the assumption that patients accept the first appointment time offered to them.

Patient Heterogeneity: We have modeled the heterogeneity of no-show behaviors among patients by grouping them into different classes according to their attributes, as we did in Chapter 4. Though it is straightforward to develop an MDP model to incorporate such an extension, solving the problem becomes much more difficult as we have already seen. The reasons include its huge state space and its much more complicated system dynamics compared to the single-class patient scheduling problem. We have developed a greedy algorithm for this problem, but by no means it is a universal solution. It will be of interest to develop heuristics using ideas other than policy improvement heuristics. Furthermore, it will be interesting to carry out a simulation study to compare the performances of different policies based on data from practice rather than the simulated data as we used in this dissertation.

Rescheduling Patients: We have proposed an approximation formula, equation (6.5), to estimate the optimal demand rate for situations where each no-show patient has a certain probability to reschedule her appointment. It will be interesting to test this formula using data from practice. More specifically, one can plug the demand rate computed by (6.5) into a simulation, and evaluate the system performance, e.g. the system throughput. Then one can further compare this performance with that under the current practice, and draw managerial insights.

Unified Design of Appointment Scheduling Systems: An ideal process to design appointment scheduling systems will be first to set the "right" panel size and pick the "right" scheduling horizon, then to determine the daily scheduling policy. The two approaches proposed in this dissertation, dynamic scheduling and static design, provide methodologies for each step in this process, respectively. However, a strict implementation of this two-step process may not work well since (i) when setting the panel size and picking the scheduling horizon in the fist step, one has not yet determined the daily scheduling policy and needs to make assumptions on it, and (ii) when determining the daily scheduling policy in the second step, one uses the panel size and the scheduling horizon chosen in the first step, which, however, may not be the "right" choice. We conjecture that in order to achieve a good design, one would need to fine-tune the system in an iterative way so as to reduce the bias and be coherent in assumptions made in each step. However, how exactly to carry out such a unified design procedure remains an interesting problem to study in the future. We believe that work on this topic will lead to fruitful research that can be very useful for practice.

Finally, it is important to note that although this dissertation has been motivated by the scheduling of outpatient appointments for clinics, our framework and the methodologies we propose are in fact relevant to any system that schedules jobs in a similar fashion and experiences customer no-shows and cancellations (e.g., hair salons, mechanics, computer support services) that depend on appointment delays.

Appendix

Proofs of the Results in Chapter 2

Proof of Proposition 1: To show that $f_j(y_j, \mathbf{x}, \mathbf{p})$ is concave in y_j, it suffices to show that $\mathbf{E}\left[r\left(w + \bar{W}_j(\mathbf{y}), v + \bar{V}_j(\mathbf{y})\right)\right]$ is concave in y_j for any real numbers w and v. Now, notice that $\bar{W}_j(\mathbf{y})$ and $\bar{V}_j(\mathbf{y}) = i$ are dependent. Furthermore, conditioning on $\bar{V}_j(\mathbf{y}) = i$, $\bar{W}_j(\mathbf{y})$ is a binomial random variable with parameters i and α_{0j}/β_{0j}. A direct application of Lemma 1 shown and proved below yields the result. $\blacksquare$.

Lemma 1. Define $B(i,p)$ to be a Binomial random variable with parameters i and p. Let $\{W_j, j = 0, 1, 2. \dots\}$ and $\{V_j, j = 0, 1, 2. \dots\}$ be two sequences of random variables such that (1) $W_j \overset{d}{=} B(j,\alpha)$, (2) $V_j \overset{d}{=} B(j,\beta)$ and (3) conditioning on $V_j = i$, $W_j \overset{d}{=} B(i, \alpha/\beta)$ for $0 \leq i \leq j$, where $0 \leq \alpha \leq \beta \leq 1$. Suppose that $r(x,z)$ is submodular and jointly concave in x and z. For any arbitrary real numbers w and v, let $g(j) = \mathbf{E}\{r(W_j + w, V_j + v)\}$. Then $g(j)$ is a concave function over the set of non-negative integers, i.e., $g(j+2) + g(j) \leq 2g(j+1)$, $\forall j = 0, 1, 2, \dots$.

Proof: The proof uses a coupling argument. Let $\{Z_{i,k}, k = 0, 1, 2, \dots\}$ be a sequence of i.i.d. Bernoulli random variables with parameter β for each $k \in \{1, 2, 3, 4\}$. First, we write

$$\begin{aligned}
&g(j+2) + g(j) - 2g(j+1) \\
=&\mathbf{E}\{r(B(\sum_{k=1}^{j+2} Z_{1,k}, \alpha/\beta) + w, \sum_{k=1}^{j+2} Z_{1,k} + v) + r(B(\sum_{k=1}^{j} Z_{2,k}, \alpha/\beta) + w, \sum_{k=1}^{j} Z_{2,k} + v) \\
&- r(B(\sum_{k=1}^{j+1} Z_{3,k}, \alpha/\beta) + w, \sum_{k=1}^{j+1} Z_{3,k} + v) - r(B(\sum_{k=1}^{j+1} Z_{4,k}, \alpha/\beta) + w, \sum_{k=1}^{j+1} Z_{4,k} + v)\}. \qquad \text{(A.1)}
\end{aligned}$$

Now we couple the random variables so that $Z_{1,k} = Z_{2,k} = Z_{3,k} = Z_{4,k} = z_k$ for $k = 0, 1, 2, \dots, j$, $Z_{1,j+1} = Z_{3,j+1} = \hat{z}$ and $Z_{1,j+2} = Z_{4,j+1} = \tilde{z}$. There are four possible cases for the pair $(\hat{z}, \tilde{z})$: (i) $\hat{z} = \tilde{z} = 0$, (ii) $\hat{z} = 1$ and $\tilde{z} = 0$, (iii) $\hat{z} = 0$ and $\tilde{z} = 1$ and (iv) $\hat{z} = \tilde{z} = 1$. For case (i), (ii) and (iii), (A.1) reduces to 0. In the rest of the proof, we show that (A.1) is also non-positive under case (iv).

Let $z = \sum_{k=1}^{j} z_k$. Then, under case (iv), we have

$$\begin{aligned} & g(j+2) + g(j) - 2g(j+1) \\ =& \mathbf{E}\{r(B(z+2, \alpha/\beta) + w, z+2+v) + r(B(z, \alpha/\beta) + w, z+v) \\ & - r(B(z+1, \alpha/\beta) + w, z+1+v) - r(B(z+1, \alpha/\beta) + w, z+1+v)\}. \end{aligned} \tag{A.2}$$

Following as above, we define $\{U_{i,k}, k = 0, 1, 2, \ldots\}$ be a sequence of i.i.d. Bernoulli random variables with parameter α/β for each $k \in \{1, 2, 3, 4\}$. Then we write equation (A.2) as

$$\begin{aligned} & g(j+2) + g(j) - 2g(j+1) \\ =& \mathbf{E}\{r(\sum_{k=1}^{z+2} U_{1,k} + w, z+2+v) + r(\sum_{k=1}^{z} U_{2,k} + w, z+v) \\ & - r(\sum_{k=1}^{z+1} U_{3,k} + w, z+1+v) - r(\sum_{k=1}^{z+1} U_{4,k} + w, z+1+v)\}. \end{aligned} \tag{A.3}$$

Now we couple the random variables so that $U_{1,k} = U_{2,k} = U_{3,k} = U_{4,k} = u_k$ for $k = 0, 1, 2, \ldots, z$, $U_{1,z+1} = U_{3,z+1} = \hat{u}$ and $U_{1,z+2} = U_{4,z+1} = \tilde{u}$. There are four possible cases for the pair $(\hat{u}, \tilde{u})$:

Case 1: $\hat{u} = \tilde{u} = 0$. Then the term inside the expectation in (A.3) becomes

$$r(\sum_{k=1}^{z} u_k + w, z+2+v) + r(\sum_{k=1}^{z} u_k + w, z+v) - r(\sum_{k=1}^{z} u_k + w, z+1+v) - r(\sum_{k=1}^{z} u_k + w, z+1+v) \leq 0$$

due to the concavity of $r(\cdot, \cdot)$.

Case 2: $\hat{u} = 1$ and $\tilde{u} = 0$. Then the term inside the expectation in (A.3) becomes

$$r(\sum_{k=1}^{z} u_k + 1 + w, z+2+v) + r(\sum_{k=1}^{z} u_k + w, z+v) - r(\sum_{k=1}^{z} u_k + 1 + w, z+1+v) - r(\sum_{k=1}^{z} u_k + w, z+1+v) \leq 0$$

since

$$r(\sum_{k=1}^{z} u_k + 1 + w, z+2+v) - r(\sum_{k=1}^{z} u_k + 1 + w, z+1+v)$$

$$\leq r(\sum_{k=1}^{z} u_k + 1 + w, z + 1 + v) - r(\sum_{k=1}^{z} u_k + 1 + w, z + v) \quad \text{(due to the concavity of } r(\cdot))$$

$$\leq r(\sum_{k=1}^{z} u_k + w, z + 1 + v) - r(\sum_{k=1}^{z} u_k + w, z + v). \quad \text{(due to the submodularity of } r(\cdot, \cdot))$$

Case 3: $\hat{u} = 0$ and $\tilde{u} = 1$. Then the term inside the expectation in (A.3) becomes

$$r(\sum_{k=1}^{z} u_k + 1 + w, z + 2 + v) + r(\sum_{k=1}^{z} u_k + w, z + v) - r(\sum_{k=1}^{z} u_k + w, z + 1 + v) - r(\sum_{k=1}^{z} u_k + 1 + w, z + 1 + v) \leq 0,$$

following the same proof in Case (2) above.

Case 4: $\hat{u} = \tilde{u} = 1$. Then the term inside the expectation in (A.3) becomes

$$r(\sum_{k=1}^{z} u_k + 2 + w, z + 2 + v) + r(\sum_{k=1}^{z} u_k + w, z + v) - r(\sum_{k=1}^{z} u_k + 1 + w, z + 1 + v) - r(\sum_{k=1}^{z} u_k + 1 + w, z + 1 + v) \leq 0$$

due to the concavity of $r(\cdot, \cdot)$.

Hence, the result follows.■

Proof of Proposition 2: Let λ denote the mean number of daily arrivals and $PO(\alpha)$ denote a Poisson random variable with mean α. Then,

$$\tilde{V}_0(p_0) \overset{d}{=} B(\tilde{A}_0, p_0) \overset{d}{=} PO(\lambda p_0), \tilde{V}_1(p_0) \overset{d}{=} B(\tilde{A}_1, (1 - p_0)\beta_{01}) \overset{d}{=} PO(\lambda(1 - p_0)\beta_{01}),$$

and

$$\tilde{W}_0(p_0) \overset{d}{=} B(\tilde{A}_0, p_0\alpha_{00}) \overset{d}{=} PO(\lambda p_0\alpha_{00}), \tilde{W}_1(p_0) \overset{d}{=} B(\tilde{A}_1, (1 - p_0)\alpha_{01}) \overset{d}{=} PO(\lambda(1 - p_0)\alpha_{01}).$$

Note that $\tilde{V}_0(p_0)$ is independent of $\tilde{V}_1(p_0)$ and $\tilde{W}_0(p_0)$ is independent of $\tilde{W}_1(p_0)$ since $\tilde{A}_0$ is independent of $\tilde{A}_1$. Let $C_1(p_0)$ be a random variable such that $C_1(p_0) \overset{d}{=} PO(\lambda p_0(\alpha_{00} - \alpha_{01}))$ (note that $\alpha_{00} - \alpha_{01} \geq 0$), and $D_1 \overset{d}{=} PO(\lambda\alpha_{01})$ be a random variable which is independent of $C_1(p_0)$. Also, let $C_2(p_0)$ be a random variable such that $C_2(p_0) \overset{d}{=} PO(\lambda p_0(1 - \beta_{01}))$, and $D_2 \overset{d}{=} PO(\lambda\beta_{01})$

be a random variable which is independent of $C_2(p_0)$. Then, using the fact that the sum of two independent Poisson random variables is another Poisson random variable, we have

$$\tilde{W}_0(p_0) + \tilde{W}_1(p_0) \overset{d}{=} PO(\lambda p_0(\alpha_{00} - \alpha_{01}) + \lambda\alpha_{01}) \overset{d}{=} C_1(p_0) + D_1$$

and

$$\tilde{V}_0(p_0) + \tilde{V}_1(p_0) \overset{d}{=} PO(\lambda p_0(1 - \beta_{01}) + \lambda\beta_{01}) \overset{d}{=} C_2(p_0) + D_2.$$

Assumption 1 implies that $w(\cdot)$ is an increasing convex function and $r(\cdot)$ is an increasing concave function. Let $z^+ = \max\{z, 0\}$. Since the class of functions $g_s(z) = (z - s)^+$ for all $s \in \mathbb{R}$ generates all the increasing convex functions and the class of functions $h_s(z) = -(s - z)^+$ for all $s \in \mathbb{R}$ generates all the increasing concave functions (see, e.g. page 173 in Shaked and Shanthikumar (1994)), in order to show

$$\mathbf{E}\left[r\left(\tilde{W}_0(p_0) + \tilde{W}_1(p_0)\right) - w\left(\tilde{V}_0(p_0) + \tilde{V}_1(p_0)\right)\right]$$

is concave in p_0 it suffices to show that

$$\mathbf{E}\{-[s - (C_1(p_0) + D_1)]^+\}$$

is concave in p_0 for all $s \in \mathbb{R}$ and

$$\mathbf{E}\{[(C_2(p_0) + D_2) - s]^+\}$$

is convex in p_0 for all $s \in \mathbb{R}$, both of which immediately follow from Lemma 2 shown and proved below. ■.

Lemma 2. Let $z^+ = \max\{z, 0\}$. Suppose that $\tilde{C}(p)$ is a Poisson random variable with mean $p\theta$ where $p, \theta > 0$ and A is a random variable that is independent of $\tilde{C}(p)$. Then $\mathbf{E}\{[(\tilde{C}(p) + A) - s]^+\}$ and $\mathbf{E}\{[s - (\tilde{C}(p) + A)]^+\}$ are both convex in p for any given $s \in \mathbb{R}$.

Proof: It is sufficient to show that for all $a \in \mathbb{R}$, $\mathbf{E}\{[(\tilde{C}(p) + A) - s]^+ | A = a\}$ and $\mathbf{E}\{[s - (\tilde{C}(p) + A)]^+ | A = a\}$ are both convex in p for an arbitrary $s \in \mathbb{R}$. We first show that $\mathbf{E}\{[(\tilde{C}(p)+A)-s]^+|A = a\}$ is convex in p for any given $a, s \in \mathbb{R}$. This is equivalent to showing that $\mathbf{E}\{[\tilde{C}(p) - s]^+\}$ is convex in p for any given $s \in \mathbb{R}$.

Now, to show that $\mathbf{E}\{[\tilde{C}(p) - s]^+\}$ is convex in p, first note that the proof is trivial if $s \leq 0$, and thus we only need to consider $s > 0$. Denote $\lceil s \rceil$ to be the smallest integer that is larger than or equal to s, and let $g(\cdot|\lambda)$ be the probability mass function of a Poisson random variable with mean λ, i.e.

$$g(i|\lambda) = \begin{cases} e^{-\lambda}\frac{\lambda^i}{i!}, & \text{if } i = 0, 1, 2, \ldots, \\ 0, & \text{otherwise.} \end{cases}$$

After some straightforward but tedious calculus, one can show that for any $s > 0$,

$$\frac{d^2}{dp^2}\mathbf{E}\{[\tilde{C}(p) - s]^+\} = \theta^2[(s + 1 - \lceil s \rceil)g(\lceil s \rceil - 1|\theta p) + (\lceil s \rceil - s)g(\lceil s \rceil - 2|\theta p)] \geq 0,$$

which establishes that $\mathbf{E}\{[\tilde{C}(p) - s]^+\}$ is convex in p for any given $s \in \mathbb{R}$.

Similarly, to show $\mathbf{E}\{[s - (\tilde{C}(p) + A)]^+\}$ is convex in p for any given $s \in \mathbb{R}$, it is sufficient to show that $\mathbf{E}\{[s - \tilde{C}(p)]^+\}$ is convex in p for any given $s \in \mathbb{R}$. But then since $\mathbf{E}\{[s - \tilde{C}(p)]^+\} = \mathbf{E}\{[\tilde{C}(p) - s]^+ + [s - \tilde{C}(p)]\} = \mathbf{E}\{[\tilde{C}(p] - s)^+\} + s - \theta p$, the result immediately follows. ■.

Proof of Theorem 1: Let f^* denote the policy that schedules all appointment requests received today for the j^*th day from today where $j^* = \arg\max_{j \in 0,1,\ldots,T} I_j$ where I_j is as described in (2.12).

First, the long-run average net reward under policy f^* is $\phi_{f^*} = (\tau\alpha_{0j^*} - \nu_1\beta_{0j^*})\mu$. For almost all sample paths w of $\{A^1, A^2, A^3, \ldots\}$, we show that, for any arbitrary policy f, the long-run average net reward along this path, denoted by $\phi_f(\omega)$, is no larger than ϕ_{f^*}. Without loss of generality, we assume that the initial backlog is empty, i.e. $\mathbf{X}^0 = \mathbf{0}$. Let $N_j(t, \omega)$ be the total number of patients scheduled with appointment delay j days up to day t along sample path ω under policy f. Now, since the cost and reward are both linear, we can view the total net reward as the sum of net rewards contributed by individual patients. Let $R_{ij}(t, \omega)$ be the net reward generated by the ith patient of those $N_j(t,\omega)$ patients whose appointment delay is j days along sample path ω up to day t, $i = 1, 2, \ldots, N_j(t,\omega)$. Notice that $\{R_{ij}(t,\omega), i = 1, 2, \ldots, N_j(t,\omega)\}$ is the realization of a sequence of i.i.d. random variables with mean $\tau\alpha_{0j} - \nu_1\beta_{0j}$ along sample path ω. Let $\mathcal{J} = \{j : \lim_{t\to\infty} N_j(t,\omega) = \infty\}$. Then,

$$\begin{aligned}
\phi_f(\omega) &= \lim_{t\to\infty} \frac{\sum_{j=0}^{T}\sum_{i=1}^{N_j(t,\omega)} R_{ij}(t,\omega)}{t} \\
&= \lim_{t\to\infty} \frac{1}{t}\sum_{j=0}^{T} N_j(t,\omega)\frac{\sum_{i=1}^{N_j(t,\omega)} R_{ij}(t,\omega)}{N_j(t,\omega)} \\
&= \lim_{t\to\infty} \frac{1}{t}\sum_{j\in\mathcal{J}} N_j(t,\omega)\frac{\sum_{i=1}^{N_j(t,\omega)} R_{ij}(t,\omega)}{N_j(t,\omega)} \\
&= \lim_{t\to\infty} \frac{1}{t}\sum_{j\in\mathcal{J}} N_j(t,\omega)\lim_{t\to\infty}\frac{\sum_{i=1}^{N_j(t,\omega)} R_{ij}(t,\omega)}{N_j(t,\omega)} \\
&= \lim_{t\to\infty} \frac{1}{t}\sum_{j\in\mathcal{J}} N_j(t,\omega)(\tau\alpha_{0j} - \nu_1\beta_{0j}) \quad \text{(by Strong Law of Large Numbers)} \\
&\le (\tau\alpha_{0j^*} - \nu_1\beta_{0j^*})\lim_{t\to\infty}\frac{1}{t}\sum_{j\in\mathcal{J}} N_j(t,\omega) \quad \text{(by the definition of } j^*\text{)} \\
&\le (\tau\alpha_{0j^*} - \nu_1\beta_{0j^*})\mu \quad \left(\text{since } \lim_{t\to\infty}\frac{1}{t}\sum_{j\in\mathcal{J}} N_j(t,\omega) \le \lim_{t\to\infty}\frac{1}{t}\sum_{j=0}^{T} N_j(t,\omega) = \mu\right) \\
&= \phi_{f^*}.
\end{aligned}$$

This completes the proof. ■.

Proofs of the Results in Chapter 5

For ease of exposition, we first present some results as the following lemmas that are frequently used in the proofs of the results in Chapter 5.

Lemma 3. Let $a, b, c, d \ge 0$. If $a/b \ge (>) c/d$, then

$$\frac{c}{d} \le (<) \frac{a+c}{b+d} \le (<) \frac{a}{b}.$$

Proof: We prove the first inequality, and the second inequality follows a similar proof.

$$\frac{a}{b} \geq (>) \frac{c}{d} \Leftrightarrow bc \leq (<) ad \Leftrightarrow bc + dc \leq (<) ad + dc \Leftrightarrow \frac{c}{d} \leq (<) \frac{a+c}{b+d}.$$

■

Lemma 4. For fix λ and μ, if $K \geq 2$ and $K \in \mathcal{S}$, then $K - 1 \in \mathcal{S}$.

Proof: In the following, we let $f_K = f_K(\lambda, \mu)$. First, we rewrite (5.3) in the following explicit form:

$$f_K = \begin{cases} 0 & K = 0 \\ \frac{p_0 + \rho p_1 + \cdots + \rho^{K-1} p_{K-1}}{1 + \rho + \cdots + \rho^K} & 1 \leq K \leq \infty. \end{cases} \tag{A.4}$$

Let $K \geq 2$ and $K \in \mathcal{S}$ but suppose for contradiction that $K - 1 \notin \mathcal{S}$. Then,

$$f_{K-2} > \frac{p_{K-2}}{\rho}.$$

Now, if $K = 2$, this is an immediate contradiction since $f_0 = 0$ and the right hand side equals p_0/ρ, which is strictly positive.

If $K \geq 3$, we have

$$\frac{p_0 + \rho p_1 + \cdots + \rho^{K-3} p_{K-3}}{1 + \rho + \cdots + \rho^{K-2}} = f_{K-2} > \frac{p_{K-2}}{\rho} = \frac{\rho^{K-2} p_{K-2}}{\rho^{K-1}}.$$

It then follows from Lemma 3 and the fact that p_j is decreasing in j,

$$f_{K-1} = \frac{p_0 + \rho p_1 + \cdots + \rho^{K-2} p_{K-2}}{1 + \rho + \cdots + \rho^{K-1}} > \frac{p_{K-2}}{\rho} \geq \frac{p_{K-1}}{\rho},$$

which contradicts the assumption that $K \in \mathcal{S}$. ■

Lemma 5. Let $\{a_i, i = 1, 2, \ldots, K\}$ be a sequence of real numbers. Suppose that there exists a $k \in \{2, \ldots, K\}$ such that if $i < k$, $a_i \geq 0$ and if $i \geq k$, $a_i < 0$. If $\sum_{i=1}^{K} i a_i = 0$, then $\sum_{i=1}^{K} i^2 a_i < 0$.

Proof:

$$\sum_{i=1}^{K} i^2 a_i = \sum_{i=1}^{K} i^2 a_i - (k-1)\sum_{i=1}^{K} i a_i = \sum_{i=1}^{k-1} i[i-(k-1)]a_i + \sum_{i=k}^{K} i[i-(k-1)]a_i < 0.$$

■

Next we present the proofs of the results in Chapter 5.

Proof of Theorem 2:

In this proof, we let $f_K = f_K(\lambda, \mu)$. Recall that λ is fixed and $T_K = \lambda f_K$. Thus, to show T_K is unimodal in K, it suffices to show f_K in unimodal in K. Similarly, to show that the maximizer of T_K can be characterized by equation (5.4), it suffices to show the the maximizer of f_K also satisfies (5.4) (since K^* maximizes T_K if and only if K^* maximizes f_K). Now, notice that the set $\mathcal{S}$ is nonempty since $1 \in \mathcal{S}$ for any value of λ and μ. We consider two cases.

Case 1: $K^* = \infty \Leftrightarrow f_{K-1} \leq \frac{p_{K-1}}{\rho}, \forall K \in \mathbb{Z}^+$.

$\Leftarrow$ direction: We show that f_K is an increasing function in K and thus $K^* = \infty$. First, it is clear that $f_1 \geq f_0 = 0$ since $p_0 > 0$. Now, since

$$f_{K-1} = \frac{\sum_{j=0}^{K-2} \rho^j p_j}{\sum_{i=0}^{K-1} \rho^i} \leq \frac{p_{K-1}}{\rho} = \frac{\rho^{K-1} p_{K-1}}{\rho^K},$$

we conclude that, for all $K \geq 2$,

$$\begin{aligned} f_K &= \frac{\sum_{j=0}^{K-2} \rho^j p_j + \rho^{K-1} p_{K-1}}{\sum_{i=0}^{K-1} \rho^i + \rho^K} \\ &\geq \frac{\sum_{j=0}^{K-2} \rho^j p_j}{\sum_{i=0}^{K-1} \rho^i} \quad \text{(by Lemma 3)} \\ &= f_{K-1}. \end{aligned}$$

$\Rightarrow$ direction: it follows from the analysis of case 2 below.

This completes the analysis of Case 1.

Case 2: $K^* = \sup\{K : K \in \mathcal{S}\} < \infty \Leftrightarrow$ if there exists a $\bar{K}$ such that $\bar{K} \notin \mathcal{S}$.

$\Leftarrow$ direction: by Lemma 4 we immediately know that $K \notin \mathcal{S}$ if $K \geq \bar{K}$ and thus $\mathcal{S}$ is a finite set which is nonempty. Let $K^* = \sup\{K : K \in \mathcal{S}\}$ as given in (5.4) and we know that $K^* < \infty$. To show f_K is a unimodal function of K and $K^* = \sup \mathcal{K}^*$, it suffices to show that f_K is increasing in K when $K < K^*$ and (strictly) decreasing when $K > K^*$. We consider two cases: $K^* = 1$ and $K^* \geq 2$. We prove the results for the case when $K^* \geq 2$. The case with $K^* = 1$ follows similarly.

Suppose that $K^* \geq 2$, i.e. $\sup\{K : K \in \mathcal{S}\} \geq 2$. By Lemma 4, we know that $2 \in \mathcal{S}$, and thus

$$f_1 = \frac{p_0}{1+\rho} \leq \frac{p_1}{\rho} = \frac{\rho p_1}{\rho^2}.$$

By Lemma 3, we have

$$f_1 = \frac{p_0}{1+\rho} \leq \frac{p_0 + \rho p_1}{1+\rho+\rho^2} = f_2.$$

Using a simple induction, we can establish that $f_{K-1} \leq f_K$ if $K \leq K^*$. Similarly, we can show that $f_{K-1} > f_K$ if $K > K^*$.

$\Rightarrow$ direction: it follows from the analysis of case 1.

Hence, the result follows. ■

Proof of Corollary 1:

We prove the results by contradiction. Suppose that $p_{K^*} = p_{K^*-1}$. Since K^* is finite, thus $K^* = \sup \mathcal{S}$ and $K^* \in \mathcal{S}$, then

$$f_{K^*-1}(\lambda, \mu) = \frac{\sum_{j=0}^{K^*-2} \rho^j p_j}{\sum_{i=0}^{K^*-1} \rho^i} \leq \frac{p_{K^*-1}}{\rho} = \frac{p_{K^*}}{\rho}.$$

By Lemma 3, we know that

$$f_{K^*}(\lambda, \mu) = \frac{\sum_{j=0}^{K^*-2} \rho^j p_j + \rho^{K^*-1} p_{k^*-1}}{\sum_{i=0}^{K^*-1} \rho^i + \rho^{K^*}} \leq \frac{p_{K^*}}{\rho},$$

which implies that $K^* + 1 \in \mathcal{S}$. Now we arrive at a contradiction that $K^* = \sup \mathcal{S} \geq K^* + 1$, and hence we must have $p_{K^*} < p_{K^*-1}$. This completes the proof. ■

Proof of Corollary 2:

We know that $p_{K^*} < p_{K^*-1}$ by Corollary 1, thus $K^* \in \mathcal{K}$ by the definition of $\mathcal{K}$. Since $\mathcal{S}$ contains only integer numbers and $K^* = \sup\{K : K \in \mathcal{S}\} < \infty$, thus we know $K^* \in \mathcal{S}$. Hence

$$f_{K^*-1}(\lambda, \mu) \leq \frac{p_{K^*-1}}{\rho},$$

which implies that $K^* \in \mathcal{S}^*$. Now, since $K \notin \mathcal{S}$ if $K > K^*$, we conclude that $K \notin \mathcal{S}^*$ if $K > K^*$ since $\mathcal{S}^* \subseteq \mathcal{S}$. This immediately implies that $K^* = \sup\{K : K \in \mathcal{S}^*\}$. The proof is complete. ■

Proof of Corollary 3:

⇒ direction:

If $K^* = \infty$, it follows from Theorem 2 that $K \in \mathcal{S}$, $\forall K \in \mathbb{Z}_+$. Thus

$$\frac{p_K}{\rho} \geq f_K, \forall K \in \mathbb{Z}_+,$$

which implies that

$$\frac{p_\infty}{\rho} = \lim_{K\to\infty} \frac{p_K}{\rho} \geq \lim_{K\to\infty} f_K = f_\infty.$$

⇐ direction:

It suffices to show that if $K^* < \infty$, then

$$\frac{p_\infty}{\rho} < \lim_{K\to\infty} f_K.$$

First, it follows from Theorem 2 that

$$f_{K^*} = \frac{\sum_{j=0}^{K^*-1} \rho^j p_j}{\sum_{i=0}^{K^*} \rho^i} > \frac{\rho^{K^*} p_{K^*}}{\rho^{K^*+1}} \geq \frac{p_\infty}{\rho}. \tag{A.5}$$

Now we can establish the following inequalities

$$\lim_{K\to\infty} f_K \quad = \quad \frac{\sum_{j=0}^{\infty} \rho^j p_j}{\sum_{i=0}^{\infty} \rho^i}$$

$$
\begin{aligned}
&= \frac{\sum_{j=0}^{K^*-1}\rho^j p_j + \sum_{K^*}^{\infty}\rho^j p_j}{\sum_{i=0}^{K^*}\rho^i + \sum_{K^*+1}^{\infty}\rho^j} \\
&\geq \frac{\sum_{j=0}^{K^*-1}\rho^j p_j + p_\infty\sum_{K^*}^{\infty}\rho^j}{\sum_{i=0}^{K^*}\rho^i + \sum_{K^*+1}^{\infty}\rho^j} \quad \text{(since } p_j \geq p_\infty, \forall j \in \mathbb{Z}_+\text{)} \\
&= \frac{\sum_{j=0}^{K^*-1}\rho^j p_j + p_\infty \rho^{K^*}/(1-\rho)}{\sum_{i=0}^{K^*}\rho^i + \rho^{K^*+1}/(1-\rho)} \quad \text{(since } \rho < 1\text{)} \\
&> \frac{p_\infty}{\rho} \quad \text{(by inequality (A.5) and Lemma 3).}
\end{aligned}
$$

This completes the proof. ■

Proof of Corollary 4:

Let $f_K = f_K(\lambda, \mu)$. If $\rho < 1$, then

$$
f_\infty \geq \frac{p_0}{\sum_{i=0}^{\infty}\rho^i} = (1-\rho)p_0 > \frac{p_\infty}{\rho} = 0.
$$

The result immediately follows from Corollary 3. If $\rho \geq 1$, then we have $f_1 > f_\infty = p_\infty/\rho = 0$. The proof is complete. ■

Proof of Theorem 3:

Part (i): Fix μ and $\{p_j\}_{j=0}^{\infty}$. With a slight abuse of notation, we rewrite (5.3) as

$$
f_K(\rho) = \frac{p_0 + \rho p_1 + \cdots + \rho^{K-1}p_{K-1}}{1 + \rho + \cdots + \rho^K}.
$$

We observe that K^* decreases with λ if and only if K^* decreases with ρ. First notice that if there exists a finite $\bar{K}$ such that $p_j > 0$ if $j < \bar{K} - 1$ and $p_j = 0$ if $j \geq \bar{K} - 1$, then it is easy to verify that $K^* < \bar{K}$ since $f_{\bar{K}-1}(\rho) > f_K(\rho)$, for all $K \geq \bar{K}$. Now, if $p_j > 0$, $\forall j = 0, 1, 2 \ldots,$ we let $\bar{K} = \infty$. We establish the results by proving the following. For a fixed K such that $1 \leq K < \bar{K} - 1$, if $f_{K+1}(\rho) - f_K(\rho) = 0$ does not have a real root on $(0, \infty)$, then $f_{K+1}(\rho) > f_K(\rho)$ and hence $K^* \geq K + 1$; else if $f_{K+1}(\rho) - f_K(\rho) = 0$ have real roots, then

(a) it has exactly one real root $\rho_K \in (0, \infty)$;

(b) If $\rho < \rho_K$, then $f_{K+1}(\rho) > f_K(\rho)$; if $\rho > \rho_K$, then $f_{K+1}(\rho) < f_K(\rho)$; and

(c) $\rho_{K+1} \leq \rho_K$, $K = 0, 1, 2, \ldots$ where $\rho_0 = \infty$.

Note that if (a) and (b) hold as well as (c) holds as a strict inequality, then $K^* = K+1$ if and only if $\rho_{K+1} < \rho \leq \rho_K$. On the other hand, if (a) and (b) hold but (c) holds as an equality, then $K^* = \sup\{j+1 : \rho_j = \rho_K\}$ if and only if $\rho_{K^*} < \rho \leq \rho_{K^*-1}$. Thus, any increase in ρ can only make K^* smaller and the proof is complete once we prove these three conditions.

Now, we compute the follows:

$$f_{K+1}(\rho) - f_K(\rho) = \frac{\rho^K}{(1+\rho+\cdots+\rho^K)(1+\rho+\cdots+\rho^{K+1})} g_K(\rho)$$

where

$$g_K(\rho) = p_K \sum_{j=0}^{K} \rho^j - \rho \sum_{j=0}^{K-1} \rho^j p_j.$$

If $p_j = p_K$ for all $j = 0, 1, \ldots, K$, then $g_K(\rho) = \rho p_K > 0$, which implies that $f_{K+1}(\rho) - f_K(\rho) > 0$. Otherwise, if there exist some i and j such that $0 \leq i < j \leq K$ and $p_i > p_j$, we will show statements (a), (b) and (c) hold. First we compute the first order derivative of $g_K(\rho)$ with respect to ρ as follows:

$$g_K'(\rho) = \sum_{j=1}^{K} (p_K - p_{j-1}) j \rho^{j-1} < 0.$$

Notice that $g_K(\rho)$ is a continuous and strictly decreasing function. Since $g_K(0) = p_K > 0$ and $\lim_{\rho\to\infty} g_K(\rho) = -\infty$, we conclude that $g_K(\rho)$ has exactly one real root on $(0, \infty)$, which we denote by ρ_K. This completes the proof of (a). Statement (b) directly follows from (a) and the fact that $g_K(0) = p_K > 0$ and $\lim_{\rho\to\infty} g_K(\rho) = -\infty$.

To establish (c), first using the fact that $g_K(\rho_K) = 0$ and $g_{K+1}(\rho_{K+1}) = 0$, we have

$$\frac{\rho_K p_0 + \rho_K^2 p_1 + \ldots \rho_K^K p_{K-1}}{p_K + \rho_K p_K + \ldots \rho_K^K p_K} = 1,$$

and

$$\frac{\rho_{K+1} p_0 + \rho_{K+1}^2 p_1 + \ldots \rho_{K+1}^K p_{K-1} + \rho_{K+1}^{K+1} p_K}{p_{K+1} + \rho_{K+1} p_{K+1} + \ldots \rho_{K+1}^K p_{K+1} + \rho_{K+1}^{K+1} p_{K+1}} = 1. \tag{A.6}$$

Suppose for contradiction that $\rho_{K+1} > \rho_K$. Then, from (b), we know that $g_K(\rho_{K+1}) < 0$, i.e.

$$\frac{\rho_{K+1}p_0 + \rho_{K+1}^2 p_1 + \dots \rho_{K+1}^K p_{K-1}}{p_K + \rho_{K+1}p_K + \dots \rho_{K+1}^K p_K} > 1.$$

Multiplying both sides by p_K/p_{K+1}, we get

$$\frac{\rho_{K+1}p_0 + \rho_{K+1}^2 p_1 + \dots \rho_{K+1}^K p_{K-1}}{p_{K+1} + \rho_{K+1}p_{K+1} + \dots \rho_{K+1}^K p_{K+1}} > \frac{p_K}{p_{K+1}} \geq 1.$$

Then, using

$$\frac{\rho_{K+1}^{K+1} p_K}{\rho_{K+1}^{K+1} p_{K+1}} = \frac{p_K}{p_{K+1}} \geq 1$$

together with Lemma 3, we find

$$\frac{\rho_{K+1}p_0 + \rho_{K+1}^2 p_1 + \dots \rho_{K+1}^K p_{K-1} + \rho_{K+1}^{K+1} p_K}{p_{K+1} + \rho_{K+1}p_{K+1} + \dots \rho_{K+1}^K p_{K+1} + \rho_{K+1}^{K+1} p_{K+1}} > 1,$$

which is a contradiction to (A.6). This completes the proof of part (c).

Part (ii): The proof is similar to the proof of part (i). We first fix λ and note that ρ and p_j are functions of μ now. Let

$$f_K(\mu) = \frac{p_0 + \rho p_1 + \dots + \rho^{K-1} p_{K-1}}{1 + \rho + \dots + \rho^K}.$$

We establish the results by proving the following. For a fixed $K \geq 1$, we show that

(a) $f_{K+1}(\mu) - f_K(\mu) = 0$ has exactly one real root $\mu_K \in (0, \infty)$;

(b) If $\mu > \mu_K$, then $f_{K+1}(\mu) > f_K(\mu)$; if $\mu < \mu_K$, then $f_{K+1}(\mu) < f_K(\mu)$; and

(c) $\mu_{K+1} \geq \mu_K$, $K = 0, 1, 2, \dots$ where $\mu_0 = 0$.

Note that if (a) and (b) hold as well as (c) holds as a strictly inequality, then $K^* = K + 1$ if and only if $\mu_K \leq \mu < \mu_{K+1}$. On the other hand, if (a) and (b) hold but (c) holds as an equality, then $K^* = \sup\{j + 1 : \mu_j = \mu_K\}$ if and only if $\mu_{K^*-1} \leq \mu < \mu_{K^*}$. Thus, any increase in μ can only make K^* larger and the proof is complete once we prove these three conditions.

We first compute the follows:

$$f_{K+1}(\mu) - f_K(\mu) = \frac{\rho^K}{(1+\rho+\cdots+\rho^K)(1+\rho+\cdots+\rho^{K+1})} g_K(\mu),$$

where $g_K(\mu)$ has the form of (7.2). Recall conditions (C1), (C2), (C3) and (C4), which imply that $\lim_{\mu\to 0} g_K(\mu) = -\infty < 0$, $\lim_{\mu\to\infty} g_K(\mu) > 0$ and $g_K(\mu) = 0$ has exactly one root on $(0,\infty)$. Let μ_K denote this root. We conclude that if $\mu > \mu_K$, then $f_{K+1}(\mu) > f_K(\mu)$; and else if $\mu < \mu_K$, then $f_{K+1}(\mu) < f_K(\mu)$. Now the proofs of (a) and (b) are complete.

To prove (c), we follow as in the proof of part (i). Let $\rho_K = \lambda/\mu_K$, $p_j = p_j(\mu_K)$ and $\bar{p}_j = p_j(\mu_{K+1})$. Using the fact that $g_K(\mu_K) = 0$ and $g_{K+1}(\mu_{K+1}) = 0$, we have

$$\frac{p_K + \rho_K p_K + \cdots + \rho_K^K p_K}{\rho_K p_0 + \rho_K^2 p_1 + \cdots + \rho_K^K p_{K-1}} = 1$$

and

$$\frac{\bar{p}_{K+1} + \rho_{K+1}\bar{p}_{K+1} + \cdots + \rho_{K+1}^K \bar{p}_{K+1} + \rho_{K+1}^{K+1}\bar{p}_{K+1}}{\rho_{K+1}\bar{p}_0 + \rho_{K+1}^2\bar{p}_1 + \cdots + \rho_{K+1}^K\bar{p}_{K-1} + \rho_{K+1}^{K+1}\bar{p}_K} = 1. \tag{A.7}$$

Suppose for contradiction that $\mu_{K+1} < \mu_K$. Then, from (b), we know that $g_K(\mu_{K+1}) < 0$, i.e.

$$\frac{\bar{p}_K + \rho_{K+1}\bar{p}_K + \cdots + \rho_{K+1}^K\bar{p}_K}{\rho_{K+1}\bar{p}_0 + \rho_{K+1}^2\bar{p}_1 + \cdots + \rho_{K+1}^K\bar{p}_{K-1}} < 1.$$

Multiplying both sides by $\bar{p}_{K+1}/\bar{p}_K \le 1$, we get

$$\frac{\bar{p}_{K+1} + \rho_{K+1}\bar{p}_{K+1} + \cdots + \rho_{K+1}^K\bar{p}_{K+1}}{\rho_{K+1}\bar{p}_0 + \rho_{K+1}^2\bar{p}_1 + \cdots + \rho_{K+1}^K\bar{p}_{K-1}} < 1.$$

Then, using

$$\frac{\rho_{K+1}^{K+1}\bar{p}_{K+1}}{\rho_{K+1}^{K+1}\bar{p}_K} = \frac{\bar{p}_{K+1}}{\bar{p}_K} \le 1$$

together with Lemma 1, we find

$$\frac{\bar{p}_{K+1} + \rho_{K+1}\bar{p}_{K+1} + \cdots + \rho_{K+1}^K\bar{p}_{K+1} + \rho_{K+1}^{K+1}\bar{p}_{K+1}}{\rho_{K+1}\bar{p}_0 + \rho_{K+1}^2\bar{p}_1 + \cdots + \rho_{K+1}^K\bar{p}_{K-1} + \rho_{K+1}^{K+1}\bar{p}_K} < 1,$$

which is a contradiction to (A.7). Thus, (c) holds and the proof is complete. ■

Proof of Theorem 4:

Define $\mathcal{S}'$ as follows.

$$\mathcal{S}' = \{K : \frac{\sum_{j=0}^{K-1} \rho^j p'_j}{\sum_{i=0}^{K} \rho^i} \leq \frac{p'_{K-1}}{\rho}, K \in \mathbb{Z}^+\}.$$

It suffices to show that $\mathcal{S}' \subseteq \mathcal{S}$. Equivalently, we show that $K \in \mathcal{S}, \forall K \in \mathcal{S}'$. Now, fix a $K \in \mathcal{S}'$. Notice that condition (5.8) implies that

$$\frac{p_j}{p_{K-1}} \leq \frac{p'_j}{p'_{K-1}}, \quad ,\forall j = 0, 1, \ldots, K-1.$$

Thus

$$\frac{\sum_{j=0}^{K-1} \rho^j \frac{p_j}{p_{K-1}}}{\sum_{i=0}^{K} \rho^i} \leq \frac{\sum_{j=0}^{K-1} \rho^j \frac{p'_j}{p'_{K-1}}}{\sum_{i=0}^{K} \rho^i} \leq \frac{1}{\rho},$$

which immediately implies that $K \in \mathcal{S}$ and completes the proof. ■

Proof of Theorem 5:

First, for convenience we rewrite throughput (5.1) as

$$T_K(\rho) = \mu \frac{\rho p_0 + \rho^2 p_1 + \cdots + \rho^K p_{K-1}}{1 + \rho + \cdots + \rho^K}.$$

Note that in the following, we prove that $T_K(\rho)$ is a strictly increasing or a unimodal function of ρ for fixed μ and p_j's. Since μ is fixed, our proof will directly imply that the throughput $T_K(\lambda)$ is also a strictly increasing or a unimodal function of λ.

For $K = 1$, we have

$$T_1(\rho) = \mu \frac{\rho p_0}{1 + \rho},$$

which is strictly increasing in ρ.

For $K \geq 2$. First consider the case when $p_j = p_0$ for $j = 0, 1, \ldots, K-1$. In this case we have

$$T_K(\rho) = \mu p_0 (1 - \frac{1}{\sum_{j=0}^{K} \rho^j}),$$

which is strictly increasing in ρ.

Now consider the case when there exist some j and k such that $0 \leq j < k \leq K-1$ and $p_j > p_k$. Let $T'_K(\rho)$ denote the first order derivative of $T_K(\rho)$ with respect to ρ. Then,

$$T'_K(\rho) = \mu \frac{h_K(\rho)}{(1 + \rho + \cdots + \rho^K)^2} \tag{A.8}$$

where

$$h_K(\rho) = \sum_{i=1}^{K} i\rho^{i-1} p_{i-1} \sum_{j=0}^{K} \rho^j - \sum_{j=1}^{K} \rho^j p_{j-1} \sum_{i=1}^{K} i\rho^{i-1}$$

and the first order condition is given by

$$h_K(\rho) = 0.$$

We will show that there can exist at most one $\bar{\rho} \in (0, \infty)$ so that $h_K(\bar{\rho}) = 0$. If there is a solution, since $T_K(0) = 0$ and $T_K(\rho) > 0$ for any $\rho > 0$, we can conclude that $T_K(\rho)$ is a unimodal function of ρ. However, if there exists no $\bar{\rho}$ for which $h_K(\bar{\rho}) = 0$, then $T'_K(\rho) > 0$ for all $\rho \in (0, \infty)$ (and thus $T_K(\rho)$ is a strictly increasing function of λ) since $T'_K(0) = \mu p_0 > 0$ and $T'_K(\rho)$ is a continuous function.

Now suppose that there exists a $\bar{\rho} \in (0, \infty)$ so that $h_K(\bar{\rho}) = 0$. Since $h_K(0) = p_0 > 0$, it suffices to show that $h'_K(\bar{\rho}) < 0$ where $h'_K(\rho)$ is the first order derivative of $h_K(\rho)$ with respect to ρ. First, rewrite $h_K(\rho)$ as

$$h_K(\rho) = \sum_{i=1}^{K} \sum_{j=1}^{K} i\rho^{i+j-1} p_{i-1} + \sum_{i=1}^{K} i\rho^{i-1} p_{i-1} - \sum_{i=1}^{K} \sum_{j=1}^{K} i\rho^{i+j-1} p_{j-1}.$$

Then,

$$\begin{aligned}
h'_K(\rho) &= \sum_{i=1}^{K}\sum_{j=1}^{K} i(i+j-1)\rho^{i+j-2}p_{i-1} + \sum_{i=1}^{K} i(i-1)\rho^{i-2}p_{i-1} - \sum_{i=1}^{K}\sum_{j=1}^{K} i(i+j-1)\rho^{i+j-2}p_{j-1} \\
&= \frac{1}{\rho}[\sum_{i=1}^{K}\sum_{j=1}^{K} i^2\rho^{i+j-1}p_{i-1} + \sum_{i=1}^{K} i^2\rho^{i-1}p_{i-1} - \sum_{i=1}^{K}\sum_{j=1}^{K} i^2\rho^{i+j-1}p_{j-1} \\
&\quad +(\sum_{i=1}^{K}\sum_{j=1}^{K} ij\rho^{i+j-1}p_{i-1} - \sum_{i=1}^{K}\sum_{j=1}^{K} ij\rho^{i+j-1}p_{j-1} - h_K(\rho)] \\
&= \frac{1}{\rho}[\sum_{i=1}^{K}\sum_{j=1}^{K} i^2\rho^{i+j-1}p_{i-1} + \sum_{i=1}^{K} i^2\rho^{i-1}p_{i-1} - \sum_{i=1}^{K}\sum_{j=1}^{K} i^2\rho^{i+j-1}p_{j-1} - h_K(\rho)].
\end{aligned}$$

Then, using the fact that $h_K(\bar{\rho}) = 0$, we can write

$$\begin{aligned}
\bar{\rho}h'_K(\bar{\rho}) &= \sum_{i=1}^{K}\sum_{j=1}^{K} i^2\bar{\rho}^{i+j-1}p_{i-1} + \sum_{i=1}^{K} i^2\bar{\rho}^{i-1}p_{i-1} - \sum_{i=1}^{K}\sum_{j=1}^{K} i^2\bar{\rho}^{i+j-1}p_{j-1} \\
&= \sum_{i=1}^{K} i^2\bar{\rho}^{i-1}b_i
\end{aligned}$$

where

$$b_i = \sum_{j=1}^{K}\bar{\rho}^j p_{i-1} + p_{i-1} - \sum_{j=1}^{K}\bar{\rho}^j p_{j-1}, i = 1, 2, \ldots, K.$$

Recall that there exist some j and k such that $0 \le j < k \le K-1$ and $p_j > p_k$. This implies that there exists at least one $b_i \neq 0$. Since $\{b_i\}$ is a decreasing sequence and

$$h_K(\bar{\rho}) = \sum_{i=1}^{K} i\bar{\rho}^{i-1}(\sum_{j=1}^{K}\bar{\rho}^j p_{i-1} + p_{i-1} - \sum_{j=1}^{K}\bar{\rho}^j p_{j-1}) = \sum_{i=1}^{K} i\bar{\rho}^{i-1}b_i = 0,$$

we deduce that there exists a $k \in \{2, \ldots, K\}$ such that if $i < k$, $b_i \ge 0$ and if $i \ge k$, $b_i < 0$. Let $a_i = \bar{\rho}^{i-1}b_i$. Then, it immediately follows from Lemma 5 that $h'_K(\bar{\rho}) < 0$, which completes the proof. ■

Proof of Theorem 7:

From (6.2), we know

$$T_K = \lambda \frac{1 - \lambda/\mu}{1 - \lambda/(\mu+\theta)} = \lambda \frac{\mu+\theta}{\mu} \frac{\mu - \lambda}{\mu+\theta-\lambda},$$

if $K = \infty$. Let T_K' represent the first order derivative of T_K with respect to λ. The first order condition yields that

$$T_K' = \frac{\mu+\theta}{\mu} \frac{(\mu - 2\lambda)(\mu+\theta-\lambda) + \lambda(\mu-\lambda)}{(\mu+\theta-\lambda)^2} = 0.$$

Recall that we assume $\lambda < \mu$ for $k = \infty$. Thus the first order condition is equivalent to

$$(\mu - 2\lambda)(\mu+\theta-\lambda) + \lambda(\mu-\lambda) = \lambda^2 - 2(\mu+\theta)\lambda + \mu(\mu+\theta) = 0. \qquad \text{(A.9)}$$

Two roots satisfying condition (A.9) are $\lambda_1 = (\mu+\theta) - \sqrt{(\mu+\theta)\theta} < \mu$ and $\lambda_2 = (\mu+\theta) + \sqrt{(\mu+\theta)\theta} > \mu$. It is easy to verify that $T_K' > 0$ if $0 < \lambda < \lambda_1$ and $T_K' < 0$ if $\lambda_1 < \lambda < \mu$, which implies that T_K increases in λ if $0 < \lambda < \lambda_1$ and decreases in λ if $\lambda_1 < \lambda < \mu$. Thus λ_1 is a local maximizer. Since we require $0 < \lambda < \mu$, hence λ_1 is a global maximizer and $\lambda_\infty^* = \lambda_1 = (\mu+\theta) - \sqrt{(\mu+\theta)\theta}$. ■

Bibliography

Argon, N. T., L. Ding, K. D. Glazebrook, S. Ziya. 2009. Dynamic routing of customers with general delay costs in a multiserver queuing system. *Probability in the Engineering and Informational Sciences* **23**(2) 175–203.

Arvantes, J. 2007. Survey confirms growing demand for primary care physicians. *American Academy of Family Physicians News Now*. Oct 16, 2007.

Barron, W. M. 1980. Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care* **7**(4) 563–574.

Belardi, F. G., S. Weir, F. W. Craig. 2004. A controlled trial of an advanced access appointment system in a residency family medicine center. *Fam. Med.* **36**(5) 341–345.

Blumenthal, D. 2004. New steam from an old cauldron — The physician-supply debate. *N. Engl. J. Med.* **350**(17) 1780–1787.

Cauchon, D. 2005. Medical miscalculation creates doctor shortage. *USA Today*. Mar 2, 2005.

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

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* **35**(11) 1003–1016.

Dixon, S., F. C. Sampson, A. O'Cathain, M. Pickin. 2006. Advanced access: More than just GP waiting times? *Fam. Pract.* **23**(2) 233–239.

Ferlie, E. B., S. M. Shortell. 2001. Improving the quality of health care in the united kingdom and the united states: A framework for change. *Milbank Quarterly* **79**(2) 281.

Gallucci, G., W. Swartz, F. Hackerman. 2005. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatr. Serv.* **56**(3) 344–346.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56**(6) 1526–1538.

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

Gupta, D., L. Wang. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research* **56**(3) 576–592.

Hassin, R., S. Mendel. 2008. Scheduling arrivals to queues: A single-server model with no-shows. *Management Science* **54**(3) 565.

Ibaraki, T., N. Katoh. 1988. *Resource Allocation Problems : Algorithmic Approaches*. M.I.T. press, Cambridge, MA.

Kaandorp, G. C., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Science* **10**(3) 217–229.

Klassen, K. J., T. R. Rohleder. 2004. Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. *International Journal of Service Industry Management* **15**(2) 167–186.

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

Krishnan, K. R. 1990. Joining the right queue: A state-dependent decision rule. *IEEE Transactions on Automatic Control* **35**(1) 104–108.

Kulkarni, V. G. 1995. *Modeling and Analysis of Stochastic Systems*. Chapman & Hall/CRC.

LaGanga, L. R., S. R. Lawrence. 2007. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences* **38**(2) 251–276.

Lamb, A. 2002. Why advanced access is a retrograde step. *Br. J. Gen. Pract.* **52**(485) 1035.

Langabeer, J. R. 2007. *Health Care Operations Management: A Quantitative Approach to Business and Logistics*. Jones & Bartlett Publishers.

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling and Analysis*. McGraw-Hill, New York.

Liu, N, S. Ziya, V. G. Kulkarni. 2009. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. To appear in *Manufacturing and Service Operations Management.*

Massachusetts Medical Society. 2007. MMS physician workforce study. Retrieved on May 14, 2009, http://www.massmed.org.

Mercer, A. 1960. A queueing problem in which the arrival times of the customers are scheduled. *Journal of the Royal Statistical Society, Series B* **22** 108–113.

Mercer, A. 1973. Queues with scheduled arrivals: a correction, simplification and extension. *Journal of the Royal Statistical Society, Series B* **35**(1) 104–116.

Moore, C. G., P. Wilson-Witherspoon, J. C. Probst. 2001. Time and money: Effects of no-showsat a family practice residency clinic. *Fam. Med.* **33**(7) 522–527.

Murray, M., T. Bodenheimer, D. Rittenhouse, K. Grumbach. 2003. Improving timely access to primary care: Case studies of the advanced access model. *J. of the American Medical Association* **289**(8) 1042–1046.

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

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

NAE and IOM. 2005. Building a better delivery system: a new engineering/health care parternship Report, National Adademy of Engineering and Institute of Medicine.

Opp, M., K. Glazebrook, V. G. Kulkarni. 2005. Outsourcing warranty repairs: Dynamic allocation. *Naval Research Logistics* **52**(5) 381–398.

Oppenheim, G. L., J. J. Bergman, E. C. English. 1979. Failed appointments: A review. *J. Fam. Pract.* **8**(4) 789–96.

Patrick, J., M. L. Puterman, M. Queyranne. 2008. Dynamic multi-priority patient scheduling for a diagnostic resource. *Operations Research* **56**(6) 1507–1525.

Pesata, V., G. Pallija, A. A. Webb. 1999. A descriptive study of missed appointments: Families' perceptions of barriers to care. *J. Pediatr. Health Care* **13**(4) 178–182.

Qu, X., R. L. Rardin, J. A. S. Williams, D. R. Willis. 2007. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research* **183**(2) 812–826.

Robinson, L. W., R. R. Chen. 2003. Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Transactions* **35**(3) 295–307.

Robinson, L. W., R. R. Chen. 2008. The effects of patient no-shows on appointment scheduling policies. Working Paper, Graduate School of Management, University of California, Davis, CA, USA.

Sack, K. 2008. In Massachusetts, universal coverage strains care. *The New York Times*. April 5, 2008.

Shaked, M., J. G. Shanthikumar. 1994. *Stochastic Orders and Their Applications*. Academic Press, San Diego, CA.

Solberg, L. I., M. C. Hroscikoski, J. M. Sperl-Hillen, P. J. O'Connor, B. F. Crabtree. 2004. Key issues in transforming healthcare organizations for quality: The case of advanced access. *Jt. Comm. J. Qual. Saf.* **30**(1) 15–24.

Strianese, A. J., P. P. Strianese. 2003. *Dining Room and Banquet Management (3rd edition)*. Thomson, New York.

Tijms, H. C. 1994. *Stochastic Models : An Algorithmic Approach*. John Wiley & Sons, Chichester, England.

Trotta, E. 2006. Quit waiting around for no-shows. *Veterinary Economics* 68–79.

U.S. Census Bureau. 2008. The 2008 statistical abstract. Retrieved on May 14, 2009, http://www.census.gov/prod/2007pubs/08abstract/health.pdf.

Vikander, T., K. Parnicky, R. Demers, K. Frisof, P. Demers, N. Chase. 1986. New-patient no-shows in an urban family practice center: analysis and intervention. *Journal of Family Practice* **22**(3) 263–268.

Weingarten, N., D. L. Meyer, J. A. Schneid. 1997. Failed appointments in residency practices: who misses them and what providers are most affected? *Journal of the American Board of Family Practice* **10**(6) 407–411.

York, M. 2007. Few young doctors step in as upstate population ages. *The New York Times*. July 23, 2007.

Zeitlin, Z. 1981. Integer resource allocations with the objective function separable into pairs of variables. *European Journal of Operational Research* **8**(2) 152–158.

Zenios, S. A., G. M. Chertow, L. M. Wein. 2000. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research* **48**(4) 549–569.