ABSTRACT

AKIN ARAS, GÖKÇE. Capacity and Yield Management in Outpatient Clinics. (Under the direction of Julie S. Ivy).

Outpatient clinics receive appointment requests from a variety of patient types who exhibit different cancellation, reschedule and no-show behaviors. It is important to provide timely appointments to these patients. In fact, our analysis reveals that the later the appointment date, the less likely the patient will be seen. Therefore it is imperative that we minimize delay in order to reduce patient loss and provide patient access. This is challenging and requires significant effort due to the number of uncertainties in the demand structure and appointment calendars in outpatient clinics. Given the increasing demand and limited capacity of outpatient clinics, it is particularly important to utilize capacity effectively and efficiently. Clinics strive to avoid appointment schedules that result in physician idle time or delay patient appointment dates due to improper capacity management.

In this study, we consider outpatient clinics, in which there are multiple patient classes with different demand rates; revenues; and behavioral functions associated with reschedules, cancels, and no-shows, each of which depends on the appointment delay (i.e., the number of days between the appointment request and the actual appointment). To address this problem we first develop discrete-event simulation models to evaluate different appointment policies. The simulation models show that for some patient classes it is especially important to provide timely appointments in order to increase the net profit as well as the seen-patient percentages. We use these insights to develop mathematical programming models to identify the optimal appointment assignment policy for each patient class, so that the net revenue is maximized

subject to the clinic's daily capacity constraints. The initial optimization model assumes stationary demand and focuses on the stochasticity in the delay-based patient behaviors, i.e., cancellation, rescheduling and no-shows. We show that under certain conditions, this model reduces to a Multiple Choice Knapsack Problem. Using this structure of the problem, we derive optimal policy properties and show that the optimal policy is either to assign next-day appointments or the latest-allowed-day (based on maximum allowed delay) appointments to patients considering the available capacity.

In the last part of this dissertation, we develop a time-varying optimization model, to incorporate time-varying demand and the timing of the reschedule requests. We also modify this model and obtain a restricted which allows only next day or latest-allowed day appointments. We compare the policies obtained from the stationary optimization model, the time-varying optimization model, and the restricted time-varying optimization model with the simulation. We observe that all the optimal policies perform significantly better than the current system in the outpatient clinic (that motivates this research), in terms of net profit and the seen-patient percentages. The time-varying optimization model solutions perform much better than the stationary optimization model solution, due to additional components, such as time-varying demand, that it takes into account.

Capacity and Yield Management in Outpatient Clinics

by
Gökçe Akın Aras

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Operations Research

Raleigh, North Carolina

2014

APPROVED BY:

_____	_____
Julie S. Ivy Committee Chair	James R. Wilson
_____	_____
Yunan Liu	Serhan Ziya

DEDICATION

This thesis is dedicated to

My mother, Semiha

My father, Levent

My brother, Gökhan

and

My husband, Korhan

BIOGRAPHY

Gökçe Akın Aras was born in Ankara, Turkey. She received her Bachelor of Science degree in Industrial Engineering in 2008 and Master of Science degree in Industrial Engineering in 2010 at Bilkent University in Ankara, Turkey. In 2010, she joined the Ph.D program in Operations Research at North Carolina State University. Since January 2014, she has been working at SAS Institute. Her research interests include capacity and revenue management, inventory control, computer simulation, and optimization in healthcare.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Julie S. Ivy, for her invaluable guidance and support during my graduate study. She has supervised me with everlasting interest and motivation. I am very lucky to have her as my advisor and she is definitely more than an advisor to me.

I am also grateful to my committee members Dr. Yunan Liu, Dr. James R. Wilson, and Dr. Serhan Ziya for accepting to read and review this thesis and for their invaluable suggestions.

I would like to thank our collaborators Todd R. Huschka, Dr. Thomas R. Rohleder, and Dr. Yariv N. Marmor from Mayo Clinic for their support and invaluable input to this research.

I am indebted to my husband Korhan Aras for his incredible love and support for nine years, especially for his patience, encouragment and understanding for the last four years.

I would also like to thank my friends, which I was lucky to have met with in North Carolina; Erinç Albey, Elif Çar Albey, Müge Çapan, İrem Şengül Örgüt, Resulali Emre Örgüt, Carl Pankok, and Cenk Türkmen. I am thankful to Uğur Cakova, Hatice Çalık, Ece Demirci, Caner Göçmen, Gülşah Hançerlioğulları, Çağatay Karan, Pelin Damcı Kurt, Mehmet Can Kurt, Aslı Sırman, Emre Uzun, and all my great friends from Bilkent University that I failed to mention here, for their invaluable support and friendship during my graduate study, I am so lucky to have you all around me.

I would like to express my sincere thanks to my professors from Bilkent University, for their continued support. I also would like to thank my co-workers at SAS Institute for their

support during the last seven months, they have made me feel as a part of their family in a very short period of time.

Last but not least, I would like to express my deepest gratitude to all my family; especially Semiha Akın, Levent Akın, Gökhan Akın, and Ömür Aras for their endless love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Outpatient clinics are faced with frequent appointment requests and visits from different types of patients (e.g., new patients, subsequent visit patients, etc.) each day. Although the patients arriving to the outpatient clinic may not be in a critical condition, it is important for the clinic to have adequate appointments to serve patients without significant delay in order to retain patients and provide patient access. Considering the various uncertainties in the demand structure and appointment calendars in outpatient clinics, this task is very challenging and requires significant time and effort. Furthermore, due to high demand and limited capacity of these clinics, it is crucial to manage capacity effectively and efficiently.

These clinics strive to avoid schedules that result in physician idle time or delay patient appointment dates due to improper capacity management.

Management of capacity in outpatient clinics requires the analysis of demand and the scheduling (or the available capacity) of the resources. Demand is generated by multiple patient groups. In the large outpatient clinic that motivates our research, there are four patient categories: new external patients, internal patients, established patients, and subsequent visit patients. New external patients are new to the clinic and usually have longer appointments. Internal patients are referred to this department from another department within the clinic but they are new to this department. These patients can have shorter appointments compared with the external patients. Established patients are patients who have been seen previously and have appointments of various lengths: short, medium or long. Subsequent visit patients are patients who return for follow-up control, and usually require shorter appointments. Another important factor influencing the demand structure is the patient behavior. Patients have different tendencies to cancel or reschedule their appointments, or fail to show up for their appointments. Gallucci, Swartz and Hackerman (2005) found that cancellation and no-show rates increase with appointment delay, i.e., the time between the appointment request date and the actual appointment date. In addition to these behaviors, we also observe that the reschedule rates are affected by the appointment delay. Thus in this study the attendance behaviors are predicated as a function of the appointment delay.

Neither the delay-based patient behaviors nor the reschedule behaviors have been addressed in the literature. However, reschedules are particularly important for scheduling because an optimal allocation policy that ignores reschedules will result in a suboptimal and

possibly an infeasible solution. This is due to the fact that rescheduled appointments are still in the system and in fact they are consuming capacity on other days. For any given day there are both new appointment requests and appointment reschedule requests, both of which require capacity. Moreover, assuming fixed behavior rates, i.e., fixed probabilities for changing an appointment, may also inaccurately represent the system, because the later the appointment, the more likely the patient will cancel, reschedule or fail to show up for his/her appointment.

Another complicating factor in this problem is that different classes of patients have different revenue potential. Therefore, it is important to evaluate the trade-off between scheduling a patient of a certain type to an earlier versus a later date considering the behavior and the revenue of different patient classes.

A traditional scheduling system, like the one used in the clinic that motivates this research, has unique calendars devoted to each physician. In general physicians manage their own calendar deciding when to schedule internal patient visits or external patient visits or the amount of time that they reserve for research, etc. Thus each calendar is uniquely designed and filled for a physician. These personal calendars have predefined slots for each patient category. Appointments are assigned to these slots based on the appointment type – referred to as slot designation matching (see Figure 1.1). If a certain type of slot is full and that type of patient requests an appointment, then that patient will be scheduled for an appointment further into the future when the next available slot for that type is available. Moreover, according to our data analysis we observe that the further into the future the appointment is,

the more likely that it will be cancelled or rescheduled. This results in unfilled or overbooked slots.

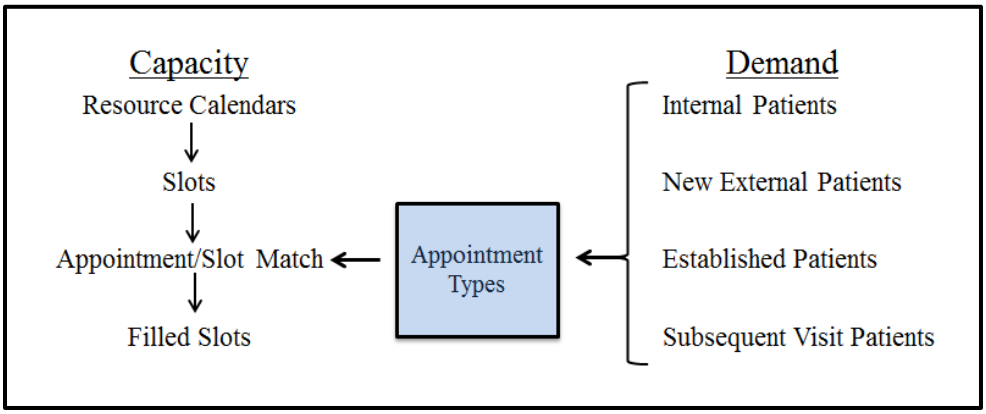

Figure 1.1: Traditional scheduling system

In our models, instead of predefined slots we consider slots of pre-specified length. The number of slots is based on the current usable total capacity for each subspecialty in the clinic, where we have seven subspecialties: ES (Esophageal), GF (Gut Failure), HB (Hepatobiliary), IBD (IBD), MO (Motility), NE (Neoplasia) and PA (Pancreas) plus 1 GE (General or Comprehensive) type of appointment. We initially assume that appointments are independent and each has its own allocated capacity. Moreover, in the proposed system, instead of having separate calendars for each physician, calendars for physician pools that are formed based on the physician subspecialties will be used. Our proposed solutions will be explained in detail in the following chapters.

This patient scheduling problem can also be analyzed from a yield management perspective. As stated previously, in the current system physicians decide on when to accept which class of patients. However, this may result in inefficient capacity management. In

reality this clinic serves various patient types (internal, external, etc.) and receives requests from each type of patient so the physician is making choices regarding not only the appointment slot to assign but how to prioritize the various patient types which may result in poor capacity management. For example, if the overall capacity allocation for external patients is insufficient because the system is filled with established or subsequent visit patients, as the physicians may choose to prioritize those due to the fact that they know those patients, this may result in the clinic losing some external patients. Given that the different patient types have different potential for bringing revenue, daily capacity needs to be allocated to the patient types, considering their expected service time, behaviors (cancellation, reschedule, no-show functions) as well as their potential revenues. We develop simulation models and mathematical programming models to address this problem.

This dissertation is organized as follows: In Chapter 2 we review the relevant literature in the patient scheduling and capacity management area. In Chapter 3, we review the literature on simulation-based approaches to patient scheduling, and introduce two simulation models. In Chapter 4 we present two mathematical programming models developed based on the insights gained from the simulation models. In Chapter 5, we introduce time-varying optimization models to improve the solution that we get in Chapter 4. Finally in Chapter 6, we conclude and discuss possible extensions of this work.

Chapter 2

Literature Review

Patient scheduling and capacity management have been explored extensively in the literature. We refer the interested readers to Cayirli and Veral (2003) and Gupta and Denton (2008) which have reviewed the recent papers within their surveys. In this chapter we consider the papers published after those reviews focusing on those papers that are closely related to our problem. In the next three sections we classify these papers into three major groups based on their problem settings: single patient class with a single resource, multi patient classes with a single resource, and multi patient classes with multiple resources. Then, we analyze each paper based on the behaviors that are taken into account, objectives of concern and the

methodologies that are used. We conclude this chapter by summarizing our contribution to the patient scheduling and capacity management literature.

2.1 Single-Class Patients with a Single Resource

Within a single class of patients and a single resource group, authors have modeled systems in which all patients have the same demand rates, rewards and behaviors (if any), and the system has only one resource.

Kim and Giachetti (2006) developed a model to find the optimal number of appointments while maximizing the expected profits. They considered overbooking to compensate for the no-shows, where the no-show rates were conditional probabilities that only depend on the number of appointments made for a given day. They assumed that the service times were deterministic but the arrivals were stochastic including probabilities for no-shows and walk-ins. In this work, the authors did not determine the appointment times but instead they obtained the daily number of appointments. They used enumeration to find the optimal solution.

LaGanga and Lawrence (2007a, 2007b) also considered overbooking to reduce the impact of no-shows. They considered sequencing of appointments on a given day. Their aim was to show when the overbooking was advantageous if the objective was to serve additional patients or to minimize waiting times or to minimize overtime. They used fixed deterministic service times and fixed appointment requests. In their problem, patients were assumed to have a probability of being a no-show, although the no-show rate that they considered was a fixed probability. They used simulation modeling to experiment with different cases.

Kaandorp and Koole (2007) had the objective of minimizing physician's idle time and overtime, and the patient's waiting time. Their focus was on sequencing the appointments on a given day. They assumed that the patients had exponentially distributed service times and also the patients were punctual (if they showed up). They also considered a fixed no-show probability for the patients. They used local search to find the schedule with the lowest objective function which was a combination of the idle time, tardiness and patient's waiting time.

Hassin and Mendel (2008) handled the problem of determining a schedule for a fixed number of customers while minimizing the total expected costs of customer waiting and server availability. They assumed that the patients had a fixed no-show probability and an exponentially distributed service time. They modeled this system as a $S(n,p)/M/1$ queue, where n was the number of scheduled customers and p was the show-up probability. They then used an optimization model to find the best policy to set intervals between scheduled arrivals.

Green and Savin (2008) found the largest panel size that could be handled by the clinic. They modeled the system as $M/D/1/K$ and $M/M/1/K$ queues with state-dependent no-shows. Here the state was defined as the length of the queue at the time when a patient requested an appointment. In addition to no-show (or last-minute cancellation) rates that depended on the queue length, they also considered the reschedule possibility that arises from the no-show patients. In this case, they simply assumed that some of the no-show patients were then rescheduled, but actually they assumed that these patients joined at the end of the queue.

Patrick (2012) used a Markov decision process model to determine the optimal outpatient scheduling considering the revenue gained by serving a patient, cost of overtime, idle time and appointment delay. The author analyzed several cases to show the trade-off between appointment delays and revenue, overtime, and idle time. Moreover, in this study the no-shows were assumed to be appointment delay-based. Instead of tracking each patient, the author assumed that the patients who scheduled in advance are given the first available slot so that the length of the queue could be used as the waiting time. Based on this model, one could dynamically decide how many patients to book in advance on a given day.

Liu, Ziya, and Kulkarni (2010) provided a dynamic programming approach for determining a dynamic scheduling policy for the outpatient appointments by maximizing the long-run average net reward of the clinic. Their aim was to select the appointment date of a patient based on the availability in the schedule at the time of the appointment request. They considered delay-based no-shows and cancellations. They modeled this problem as an MDP; and because the state space is too large and intractable, they developed heuristics to obtain a policy.

Feldman et al. (2012) sought to maximize the net revenue per day, using the revenue obtained from each seen patient and the cost related to the service of the scheduled patients. They assumed that the patients' cancellation and no-show rates were both delay dependent. They also considered the patient preferences regarding when to be seen. In order to find the optimal set of appointment days to offer the patients before assigning appointments, they first built a static model (mathematical programming model). They then extended it to a dynamic model (an MDP formulation) that considered the most recent information about the

scheduled appointments. Similar to Liu, Ziya and Kulkarni (2010), this model also had a very large state space. For this reason they proposed a similar approximation method to solve the problem.

In summary, Kim and Giachetti (2006), LaGanga and Lawrence (2007), Kaandorp and Koole (2007), and Hassin and Mendel (2008) considered no-shows by assuming that they were deterministic and fixed, while delay-dependent no-show rates were used by Patrick (2012). Green and Savin (2008) considered both delay-dependent no-show rates and probability of reschedule only after a patient does not show up. In addition to delay-based no-show rates, Liu Ziya, and Kulkarni (2010) and Feldman et al. (2012) also incorporated delay-based cancellation rates into account.

2.2 Multiple-Class Patients with a Single Resource

Papers with multiple classes of patients with a single resource modeled systems with different patient classes that had different demand rates and/or different behavior function rates and/or different priorities (or rewards) associated with them.

Gerchak, Gupta and Henig (1996) modeled the decision problem of surgical capacity reservations for the emergency patients while the same rooms were used for the elective patients, with stochastic service times. They modeled this problem as a stochastic dynamic programming model. They included a penalty for delaying elective surgeries, so their objective, the expected net profit per day, depended on the number of elective patients that were admitted or postponed on a given day, and the overtime needed to complete unfinished surgeries.

Vanden Bosch and Dietz (2000, 2001), and Vanden Bosch, Dietz and Simeoni (1999) sought to minimize the weighted sum of physician's expected overtime and patients' expected direct waiting time under the following assumptions: (i) a fixed number of appointment requests each day; (ii) punctual arrivals (if the patient showed up); (iii) known service time distributions; and (iv) fixed probabilities of no-shows. They classified patients based on the service time distributions, and obtained no-show rates for each class. They modeled this system as a transient queue with deterministic arrivals and used an optimization model to obtain an optimal sequence of arrivals.

Green, Savin, and Wang (2006) considered a hospital diagnostic facility which was used by scheduled outpatients, walk-in inpatients, and emergency patients. They assumed that the scheduled outpatients may not show up with a known fixed probability, and the arrivals of the emergencies and inpatients were based on known probabilities. Their purpose was to find the optimal dynamic policy on admission of the patients, while maximizing the net profit. When calculating the net profit, they considered different revenues gained by seeing an outpatient and inpatient, waiting costs for an outpatient and inpatient and penalty costs for unseen outpatients and inpatients. They solved this problem using a finite-horizon dynamic program.

Patrick, Puterman and Queyranne (2008) also considered a diagnostic facility which was used by inpatients and outpatients with different priorities. Their objective was a function of costs generated by not meeting the waiting time target, using overtime, and demand that was not booked or diverted. In their model they assumed that the decisions regarding the outpatients were determined after the inpatient demand was known at the beginning of a

given day, since the outpatients request appointments with certain probabilities. They used a rolling horizon of length N and in their setting at the beginning of each day the scheduler assigned appointments to the following N days or decided to hold (divert) a patient until the next day to assign an appointment. They used an infinite horizon MDP to model this problem and then solved with a linear program. However, due to the size of the problem both of the models were intractable; thus they used approximate dynamic programming (ADP) to solve the problem.

Stanciu (2009) sought to find the optimal capacity allocation for different patient classes in a system with random demand and service times while maximizing the expected revenue over the planning horizon using a mathematical model. In this problem, different classes had different revenues for the clinic but the author did not consider any type of attendance behaviors. A revenue management approach was used to solve this problem. Since the model constructed was intractable, Stanciu et al. (2010) modeled the same problem using simulation-optimization to get near-optimal solutions.

Ayvaz and Huh (2010) also considered the problem of capacity allocation with multiple types of patients (electives and emergencies), each of which had the same service times but different profit potential. The profit potential was calculated using revenue from the visit, penalties for losing a patient, or penalties for causing a patient to wait. They did not include patient behaviors in their model. Dynamic programming was used to find the optimal policy that maximized the total net revenue while dynamically allocating capacity after backlogged elective patients from the previous day were observed.

Cayirli, Veral and Rosen (2008) focused on a clinic that had demand from new, return and walk-in patients, each of which required different amounts of physician time. In this system, patients were assumed to have a fixed no-show probability, which was the same for each group. Their purposes were to determine the sequencing of the appointments and to obtain the appointment intervals by taking different service times into account. They used simulation to experimentally investigate different decisions with the objective of reducing the patients' waiting times, the physician's idle plus overtime.

Muthuraman and Lawley (2008) focused on an outpatient clinic in which there were multiple patient types, each of which had the same exponentially distributed service time but different fixed no-show probabilities. Their aim was to maximize the expected profit by minimizing waiting times and the number of waiting patients at the end of the day (causing overtime), and maximizing the resource utility. The decisions available in their stochastic model were whether to accept an appointment request and assign it to any slot available that particular day or to deny it. Thus their model was a myopic decision making model. Later, Chakraborty, Muthuraman and Lawley (2010) extended Muthuraman and Lawley (2008) by using general service times.

Zeng et al. (2010) focused on a scheduling problem in which different patient types had different no-show probabilities that were fixed. Their objective was to maximize the net profit, which included the physician's overtime cost, the revenue obtained from seen patients and the patients' waiting time costs. The patient-related costs and revenues were assumed to be the same for all patient groups. They only modeled the single-day scheduling problem with a purpose of sequencing the daily appointments. They used an optimization model to

solve this problem, but because their objective function was multimodular, they obtained local optimal schedules using a local search algorithm.

Ratcliffe, Gilland and Marucheck (2011) studied the problem of joint capacity control in an outpatient clinic, which had two patient classes with different delay-based no-show probabilities but the same revenue and service time. They assumed that the number of patients that showed up was a binomially distributed as a function of the number of booked appointments for that patient class. In other words, the appointment delay was measured by the number of booked appointments on a given day. They also consider overbooking and overtime when solving for the best single day scheduling policy. Stochastic dynamic programming was used to accomplish this.

Samorani and LaGanga (2011) also assumed that each patient class had a different delay based upon no-show probabilities. Unlike the previous paper, the authors modeled a multi-day scheduling problem. Their objective consisted of the revenue obtained by seeing a patient, the waiting time costs of a patient, and the overtime costs of the physician. In this setting, whenever an appointment request was observed, a slot was assigned to that patient based on the solution of the optimization problem taking the no-show probability of that patient into account. Based on their experiments via simulation, they also defined a scheduling rule that yielded near optimal solutions.

Schutz and Kolisch (2013) focused on a system with different classes of patients each of which had different revenue potential and fixed no-show and cancellation probabilities. With their capacity allocation model maximized the net profit which was calculated using revenues for each patient seen, refunds for cancellations and no-shows, and overtime costs. They also

allowed for overbooking. They used an MDP model in which the decision was to accept or deny the appointment requests as they arrive; and they solved this MDP optimally using stochastic dynamic programming.

In review, within this class Gerchak, Gupta and Henig (1996), Patrick, Puterman and Queyranne (2008), Stanciu (2009), and Ayvaz and Huh (2010) did not consider any type of behaviors. Green, Savin and Wang (2006) assumed a fixed no-show rate for a single patient class, whereas Vanden Bosch and Dietz (2000, 2001), and Vanden Bosch, Dietz and Simeoni (1999) used fixed no-show rates that differ for each class. Moreover, Cayirli, Veral and Rosen (2008), Muthuraman and Lawley (2008), Chakraborty, Muthuraman, and Lawley (2010), and Zeng et al. (2010) assumed patient class-based fixed no-show rates. In addition to the class-based fixed no-show rates, Schutz and Kolisch (2013) also considered class-based cancellation rates. Finally, Ratcliffe, Gilland, and Marucheck (2011) and Samorani and LaGanga (2011) used delay-dependent class-based no-show rates. It is important to note that none of these papers consider reschedules (either delay based or fixed) in addition to delay-based no-show and cancellation rates.

2.3 Multiple-Class Patients with Multiple Resources

In the last group of papers, which model multiple classes of patients and multiple resources, there were different characteristics associated with each patient class. For this category of papers the models had multiple possible resources. This group can also be divided into two subgroups: Papers in the first subgroup consider systems in which multiple resources were

used during a single patient visit, and the papers in the second group model systems in which a single resource was chosen from a set of resources based on the appointment type.

Gupta and Wang (2008) developed policies for determining whether to accept or deny an appointment request so as to maximize the net revenue considering the patients' selection behavior in terms of the time of their appointment and provider preference. They considered a primary care clinic with *n* physicians. The patient groups that they considered were same-day patients, who tend to accept any slot available that day, and regular patients who had preferences for providers and appointment timing. The net revenue had the following components: revenue gained by seeing a patient, cost to deny a regular patient, and cost of insufficient same-day capacity for same-day patients. They modeled this problem as an MDP with a rolling horizon that was divided into periods where there could be at most one appointment request each.

White, Froehle and Klassen (2011) used a simulation model to analyze the effect of different appointment scheduling and capacity allocation strategies on resource utilization and patient waiting time using a full factorial analysis. This paper used multiple resources for a patient; i.e., an appointment occupies slots both in the physician's and the exam room's calendars. They also considered patient pathways in the clinic while determining the appointment rules and allocation capacity.

Huh, Liu, and Truong (2013) also worked on a problem in which multiple resources were assigned to a single patient. Based on the waiting time sensitivity, they used two patient groups in their setting; elective and emergency. By using an MDP model, they minimized the total cost which included the cost of adding elective patients to the wait list, the loss of

revenue for lost elective patients and surge capacity usage for the emergency patients. This model allowed elective patients to leave the waiting list with a time-dependent rather than waiting-time dependent probability.

With this set of papers, neither Gupta and Wang (2008) nor White Froehle and Klassen (2011) considered patient behaviors, whereas Huh, Liu, and Truong (2013) used time dependent no-show rates.

2.4 Our Contribution to the Patient Scheduling and Capacity Management Literature

The models presented in this dissertation include multiple patient classes, each of which has different revenue potentials and delay-based seen, no-show, cancellation and reschedule rates. We constructed a literature review table based on the problem setting and behaviors taken into account in the relevant papers and highlight our literary contributions (Table 2.1). Due to the complicated nature of the system considered in this research, we first use simulation to analyze the effects of changing appointment windows for differing patient classes, then construct mathematical programming models. An important consideration in our work is the inclusion of reschedules, because prior literature typically ignored this behavior due to its complexity, despite the fact that reschedule rates may be particularly significant in outpatient clinics. Ignoring the effect of reschedules can yield solutions that are infeasible in terms of capacity, because with rescheduling-based demand is simply moved to other days, thus it still creates load on the system.

Table 2.1: Categorized Summary of the Patient Scheduling and Capacity Management Lit.

Setting / *Behaviors*	**Single Class Patients Single Resource**	**Multiple Class Patients Single Resource**	**Multiple Class Patients Multiple Resources**
None	Not included	➢ Gerchak, Gupta and Henig (1996) ➢ Patrick, Puterman and Queyranne (2008) ➢ Stanciu (2009) ➢ Stanciu et al. (2010) ➢ Ayvaz, Huh (2010)	➢ Gupta, Wang (2008) ➢ White, Froehle and Klassen (2011)
No-Show	➢ Kim, Giachetti (2006) ➢ Kaandorp, Koole (2007) ➢ LaGanga, Lawrence (2007a, 2007b) ➢ Hassin, Mendel (2008) ➢ Patrick (2012)	➢ Vanden Bosch et al. (2000, 2001) ➢ Vanden Bosch et al. (1999) ➢ Green, Savin and Wang (2006) ➢ Muthuraman, Lawley (2008) ➢ Cayirli, Veral and Rosen (2008) ➢ Zeng et al. (2010) ➢ Chakraborty, Muthuraman, Lawley (2010) ➢ Ratcliffe, Gilland, Marucheck (2011) ➢ Samorani LaGanga (2011)	➢ Huh, Liu and Truong (2012)
No-Show, Cancellation	➢ Liu, Ziya and Kulkarni (2010) ➢ Feldman et al. (2012)	➢ Schutz, Kolisch (2010)	
No-Show, Reschedule	➢ Green, Savin(2008)		
No-Show, Cancellation, Reschedule	_	***Our work***	

Our first simulation model contributes to the patient scheduling literature by incorporating delay-based cancellation, reschedules and no-show probabilities. Then, building on the first simulation model, we develop our second simulation model in which we link the patient classes so that the subsequent visits are generated by other types of patients. To the best of our knowledge the previous multi-patient studies do not consider dependency between the patient classes.

Next, using mathematical programming, we model the patient behaviors analytically while determining how many days out we should schedule a patient so that capacity is effectively utilized. Note that the analytical models we present in Chapter 4, assume stationary demand on all days. In Chapter 5 we introduce a time-varying model in which we allow for different daily demands and decision variables on each day. Moreover we also incorporate the timing of reschedule requests to our model.

Finally, note that we only discussed papers related to scheduling, capacity management and revenue management (RM) in the healthcare field. There are no closely related capacity allocation papers in the revenue management literature. Generally, RM has been applied to hotel and airline management but in those studies the reschedule requests arising from the customers have also not been modeled to the best of our knowledge. Other differentiating features of hotel and airline RM, are that, airlines charge large penalties for changes/reschedules and hotels allow changes within a given window. This is one of the reasons that we focus on yield management problems in healthcare. We refer the interested readers to Talluri and Van Ryzin (2004) for more information about revenue management.

Chapter 3

Simulation Models

3.1 Introduction

We develop two simulation models to explore the patient scheduling and capacity allocation system. In the first model, assuming independent patient classes, we evaluate the effects of reducing appointment windows for all patient types at the same time by the same amount, under two performance indicators, the seen-patient percentage and utilization. Using this model we also compared the effect of using calendars with traditional slots versus calendars with standardized length of slots. We use the capacity utilization and the “seen” patient proportions as the performance indicators in this model. In the second simulation

model we incorporate the fact that the subsequent visits are generated by other types of appointments. In addition, we consider (i) different appointment windows for different patient classes, and (ii) different revenues obtained by seeing patients of different classes. In this model capacity utilization, patient access, and financial rewards are used as performance indicators. These two simulation models are explained in our related publications, Akin et al. (2013a) and Akin et al. (2013b).

The remainder of this chapter is organized as follows: in Section 3.2 we introduce relevant simulation papers in the outpatient clinic area, and in Section 3.3 the first simulation model is explained in detail with the data analysis, model components and numerical results. In Section 3.4, the second simulation model is presented with additional data analysis and numerical results. Finally Section 3.4 discusses the findings of this work and the conclusion.

3.2 Literature Review

There are several studies in which discrete event simulation has been used to model outpatient clinics. Jun, Jacobson, and Swisher (1999), Cayirli and Veral (2003) and Gunal and Pidd (2010) include extensive reviews of these applications. Here we introduce the closely related papers.

One of the earliest papers on the simulation modeling of the outpatient clinics is by Fetter and Thompson (1965). Fetter and Thompson (1965) compared patients' direct waiting time, defined as the difference between the patient's appointment time and the time that the patient was seen by the provider, to the physician utilization rates under different input variables, like no-show rates, walk-in versus appointment scheduling rates, and service times. They

considered two appointment types (walk-ins and by appointments), with no cancellations, reschedules or no-shows.

Hashimoto and Bell (1996) modeled an appointment-based outpatient clinic where there were sequential providers and a single patient class. They used simulation to get insights about the bottleneck processes and patient waiting times in the clinic. In the system that they considered, once the patient was admitted to the hospital, he/she was seen by a sequence of providers, and left the clinic. They did not consider cancellations, no-shows or reschedules in this process.

Klassen and Rohleder (1996) used simulation to model an outpatient scheduling system and analyzed different rules to identify the solution which minimizes physician idle times and patient waiting times. Their scheduling rules identified where to schedule patients with longer service times and where to reserve slots for unscheduled urgent patients. They incorporated a fixed no-show rate for patients; cancellations and reschedules are not taken into account.

Harper and Gamlin (2003) considered multiple patient classes each of which required different service times, however they ignored the patient's no-show, cancellation and reschedule behaviors. This paper used the patient's direct waiting time as the performance indicator. They used a simulation model to test different scheduling policies which assigned appointment times to patients and assigned the times that the providers should begin accepting patients.

Similarly, Guo, Wagner, and West (2004) considered direct waiting time and fixed no-show probabilities that were different for different patient classes. Using a discrete event

simulation model, they predicted the effect of different scheduling rules on patient waiting times and overall utilization rates.

Wijewickrama and Takakuwa (2005) modeled an outpatient clinic with two independent patient classes with different consultation times where patients may not show up (no-shows are not delay-based). They modeled the entire process in the outpatient clinic, starting from check-in with the receptionist to billing. By doing so, they had the ability to observe the direct waiting times and bottlenecks in the system. They assumed that there were incoming patients with appointments and walk-in patients. They did not consider when to schedule (how far out) an appointment, instead they focused on the daily organization of the schedules. They did not incorporate cancellations and reschedules by the appointment-based patients.

Cayirli, Veral, and Rosen (2006) considered two patient classes (new-N and return-R) each of which required a different service time distribution but each class had the same fixed no-show probability. They modeled walk-in patients and the punctuality of the patients (including both earliness and lateness with respect to their assigned appointment times). They focused on daily sequencing of the appointments (i.e., RNRNR, NNN, …, RRR, etc.); and they use simulation to evaluate different sequencing rules. They did not consider cancellations and reschedules.

Wijewickrama and Takakuwa (2008) modeled outpatient appointment scheduling in a multi-facility system via simulation. They had two patient classes (new and appointment patients) and assumed that these patient classes had the same fixed no-show and lateness (or earliness) probabilities. They evaluated different appointment sequencing rules and chose the

best one in terms of patients' direct waiting times and physicians' idle times. Similar to the previous papers, cancellations and reschedules were not considered in this paper.

White, Froehle, and Klassen (2011) used simulation modeling to analyze the effects of capacity allocation and appointment policy decisions on an outpatient clinic. They grouped appointments based on the appointment length mean and variance, and determined the order the appointments should be assigned on a given day. They did not consider a planning horizon, instead they focused on planning a single day appointment sequences. The performance indicators included patients' direct waiting times, physicians' utilizations and overtime. They assumed patients always show up and did not cancel or reschedule their appointments.

In our research, through our analysis of data from a large outpatient clinic we observe that patients frequently cancel or reschedule their appointments. Also they may not show up for their appointments. We also observe that the indirect waiting time, i.e., the difference between the appointment request date and the actual appointment date, has a significant effect on the system's performance, as the patients' behaviors are highly dependent on this appointment delay. Using indirect waiting times we can more accurately represent an outpatient clinic, because the patients' behaviors all depend on the indirect waiting time. As discussed in Chapters 1 and 2, patients that are assigned appointments later are more likely to cancel, reschedule or not show up for their appointments. Instead of direct waiting times we use indirect waiting times in our simulation models to capture these effects. Our two simulation models are the first to handle a system with multiple patient classes that have *delay-based no-show*, *reschedule* and *cancellation* probabilities. Additionally, in the second

model we incorporate the relationship between of the subsequent visit and the other patient classes. Patient-based revenues and penalties associated with cancellation and reschedules are also taken into account as we allow for different patient classes to have different appointment windows.

3.3 Simulation Model – I

Two years of data has been used to characterize the current state of the clinic, and to estimate the patient-related attributes (demand and behavioral functions) and capacity related parameters (daily available capacities for each subspecialty). After estimating the necessary parameters and functions we build a simulation model in Arena, validate the current system model with the actual data, and we experiment with different scenarios to see the effect of changes.

3.3.1 Current State and Data Analysis

In the current scheduling system, physician calendars are available for appointment scheduling for up to 12 weeks into the future. Different classes of patients have different appointment request frequencies as well as different cancellation, no-show, reschedule, and seen rates which depend on the appointment delay. In addition to distinct daily demand volumes for each patient class, each of the patient behaviors should be characterized by different functions for each patient classes. Each patient class can schedule appointments of different lengths (either 20, 40 or 60 minutes) depending on the appointment type that they need. Moreover, the clinic has varying capacity based on the day of the week, and the

number of physicians in each of the eight subspecialties also differs, so it is important to represent the daily capacities by taking the day of the week and the subspecialties into account. In the next three subsections the data analyses that correspond to the above components are explained briefly.

Demand Data:

According to the demand data, 9.7% of the appointment requests arise from new external patients, while 15.9% are for internal patients, 33.6% are for established patients and 40.8% are for subsequent visit patients, respectively. In this setting, it is assumed that daily appointments are deterministic. We use the mean values, obtained from the data, for the number of appointment requests of each patient class.

Demand for a subspecialty can be observed for each of the four patient groups. The required appointment length can be 20, 40 or 60 minutes depending on the appointment type within each subspecialty. Specifically, new external patient appointments usually take 60 minutes, internal patient appointments take 40 minutes and subsequent visit patients take 20 minutes regardless of the subspecialty physician that they need to visit. On the other hand, established patients can take 20, 40 or 60 minutes depending on the appointment type that they need, and based on the data analysis, it has been observed that 7.7% of the time appointments take 20 minutes, 78.1% take 40 minutes and 14.2% take 60 minutes.

Patient Behaviors:

Patients can cancel or reschedule their appointments before their appointment day, or they may wait until their appointment day, but then they may or may not show up for their appointment. Moreover, different types of patients have different delay-dependent probabilities of cancellation (C), reschedule (R), no-show (NS) or seen (S) (see Figure 3.1). Note that the analysis for the probability curves in Figure 3.1 was based on a nine week horizon due to the small number of appointments scheduled more than nine weeks out. As can be seen from the graphs in Figure 3.1, the longer the delay between the date of the

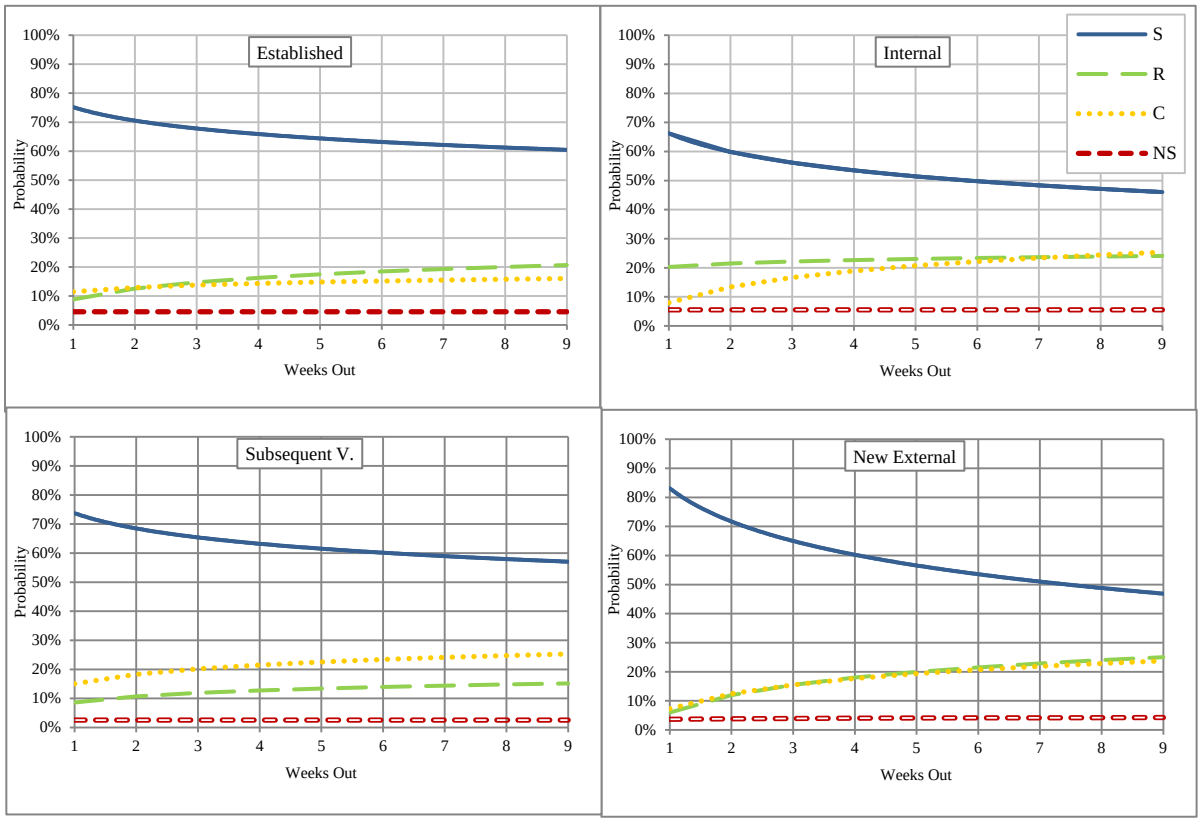

Figure 3.1: Probability curves representing patient behaviors by patient type

appointment request and the appointment date, the more likely the patient will not be seen. Moreover, the probability of no-show is very low for all patient categories (much less than 10%), while the reschedule and cancellation probabilities are considerably higher and tend to increase with the appointment delay. This further suggests that longer delays result in more changes in the appointment. Furthermore, the rate at which the probability of being seen decreases is higher for new external and internal patients than it is for established and subsequent visit patients. This is believed to be due in part to the fact that new external patients have no commitment to the hospital yet, so if their initial appointment is too far out it is much more likely that they will not be seen, compared with established and subsequent visit patients. This behavior is more significant in the internal patients, because they are already in the hospital and referred to this department during their visit. Thus if they cannot be accommodated at the earliest convenience, their probability of being seen decreases drastically compared with the established and subsequent visit patients.

As stated earlier, on average 17% of all appointments are rescheduled (either to an earlier day *or* a later day). Thus it is important to consider this movement of appointments between the days. It should be noted that the patient behaviors are slightly different after a reschedule. In other words, as can be seen in Figure 3.2, once a patient reschedules her appointment, the probabilities of cancellation, reschedule, no-show and seen are different than the initial behaviors that were shown in Figure 3.1.

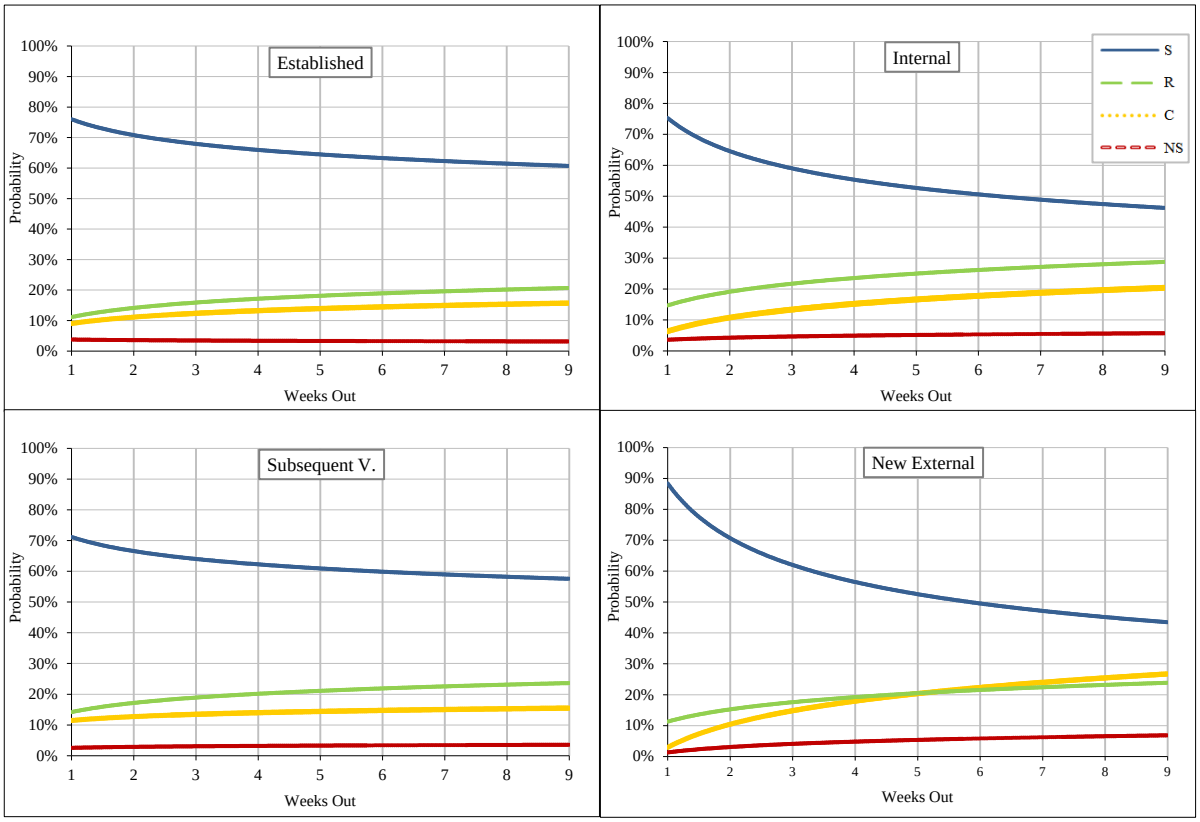

Figure 3.2: Probability curves representing patient behaviors after rescheduling by patient type

In Figure 3.3, the probability curves associated with not changing the appointment until the appointment day (i.e., “make the day”) for all patient groups are plotted. From this graph it can be observed that new external patients have the greatest reduction in their probability of “making the day.” The established patients and subsequent visit patients have less steep curves suggesting that the rate of change in the probability of “making the day” declines more gradually with the appointment delay (see Figure 3.3).

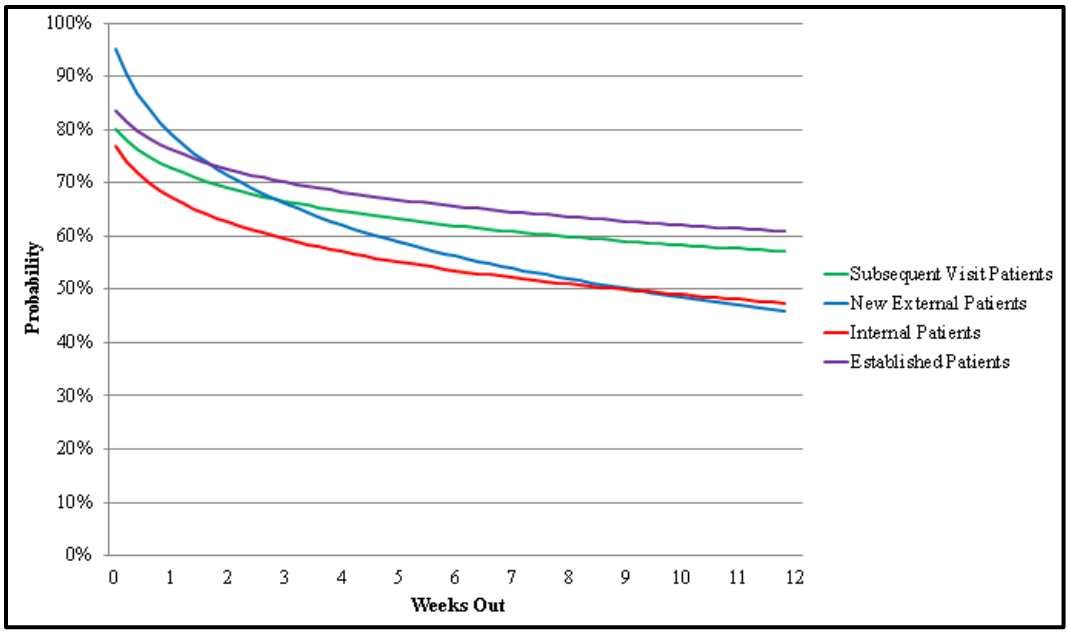

Figure 3.3: Probability of not changing the appointment until the appointment day (making the day), given that the patient has x weeks until her appointment

Capacity Data:

Capturing the available capacity is one of the most challenging aspects of this type of study. There are lots of physician calendars that are being used in the current scheduling system and in each calendar there are lots of different slot types defined. Thus this capacity capturing process had to be completed manually with expert review over several data files. The daily capacities are calculated based on the available hours for slots in the current calendars. Since there are several calendars, they are grouped into eight subspecialties with certain probabilities and each of these is considered separately. Their proportional capacities within the total capacity are used in the simulation model. It is important to note that in this problem setting there are multiple resources which behave like multiple single resources because we

have grouped them by subspecialties. For this reason, we use overall utilizations to evaluate various scenarios.

3.3.2 Model I

Due to the complexity of this system, a simulation model developed in Arena is used to analyze it under various scenarios, see Figure 3.4 for the process flow used in Arena. In all scenarios demand for different appointment types, patient behaviors (that are characterized by functions that depend on appointment delay), and daily capacities are assumed to be the same as in the current system as explained in Section 3.2.1.

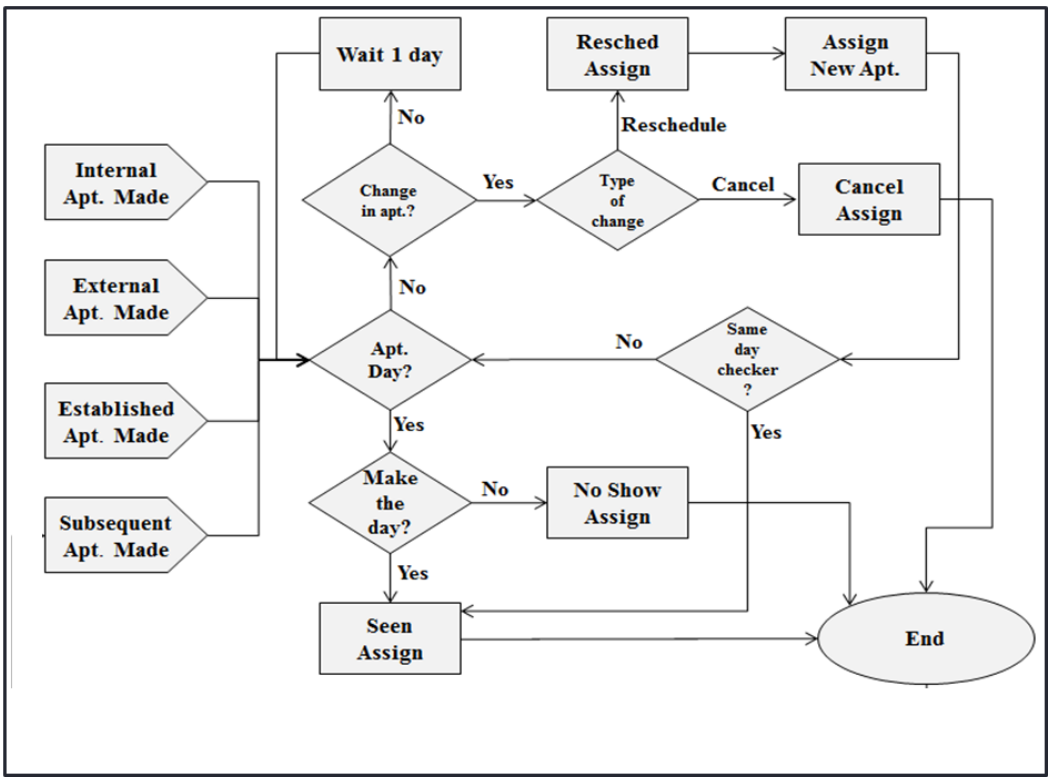

Figure 3.4: Process flow model used in simulation model I

First the current scheduling system is modeled. In this model the current days out functions are used for determining the appointment date. This represents current delays in appointments due to inefficient capacity management. If an appointment slot is not available, then the patient is moved to the following day until an appointment slot is available. A 12-week planning horizon is used as is the case in the current system. We allow the simulation model to run for 1300 days but we set the warm up period to be 1000 days, in order to make sure that the results are steady state results. The model is validated based on the days spent in the system by the patients that are seen without any change in their appointment dates (Table 3.1), as well as the proportions of seen, cancelled and no-show patient appointments (Table 3.2).

Table 3.1: Model validation based on number of days that seen patients spent in the system

Patient Type	**Data Analysis**	**Simulation Results**
New External	19.40	18.18
Internal Consult	10.53	11.63
Established	24.43	23.07
Subsequent Visit	10.29	11.01

Next, the simulation model is modified to evaluate the effect of a slotless design. Under this design, there are no predefined slots of fixed lengths. In other words, when an appointment is requested, it is directly scheduled to the requested day (if available) and blocks the required amount of time, depending on the appointment type, on that day of the

Table 3.2: Model validation based on seen/cancelled/no-show patient percentages

Patient Type	Disposition	Data Analysis	Simulation Results
New External	C - Cancelled	20.0%	20.0%
New External	N - No Show	4.7%	4.8%
New External	S - Seen	75.3%	75.2%
Internal	C - Cancelled	18.3%	16.9%
Internal	N - No Show	5.7%	6.9%
Internal	S - Seen	76.1%	76.1%
Established	C - Cancelled	18.4%	16.6%
Established	N - No Show	4.3%	4.2%
Established	S - Seen	77.4%	79.3%
Subsequent V.	C - Cancelled	20.2%	19.6%
Subsequent V.	N - No Show	2.4%	3.3%
Subsequent V.	S - Seen	77.4%	77.2%

calendar. If the appointment is cancelled or rescheduled prior to the appointment day, then that amount of time becomes available for the new appointment requests. This design is analyzed with various planning horizon lengths: 12 weeks, 2 weeks, and 1 day. For the 12 week planning horizon, the current days out functions are used. For the 2 week case, the current days out functions are skewed by using a multiplier so that the maximum appointment delay would be 2 weeks. For the 1 day planning horizon, we set the days out function equal to 1 so that each patient is given an appointment on the next day, if available (if it is not available the patient is moved to the next day until an available day is found).

After evaluating the slotless design with various planning horizon lengths, the simulation model is modified to observe the effect of standardized slots. Slot lengths of

20 and 30 minutes are evaluated using the simulation model. Under the 12 week planning horizon, the slotless design, standardized 30 min slots, and standardized 20 min slots are compared. In addition, using the standardized 20 min slots, we vary the planning horizon length (12 weeks, 2 weeks, 1 day) to see the planning horizon effect when we have standardized slots.

Under each of these scenarios the performance indicators, i.e., the utilization of physicians and the percentage of seen patients, are analyzed with 95% confidence intervals (shown by black bars in each graph). Finally, because the clinic is interested in the effects of different scheduling scenarios on the number of seen patients of different types, each of these scenarios is evaluated with respect to this rate. The corresponding results are discussed in the next section.

3.3.3 Results Based on Model I

As shown in Figure 3.5A, if we reduce the planning horizon under the slotless design, the utilization of physicians is higher. Reducing the planning horizon from 12 weeks to 2 weeks, increases the utilization from 82.8% to 87.9%; and reducing it to 1 day yields 91.5% utilization, while the current utilization of the physicians is 81.6% (see Figure 3.5A). The planning horizon reduction also has positive effects on the seen patient rates, which are the proportion amounts of seen patients out of all appointments. Figure 3.5B shows that the seen patient rate can be increased from 77.67% to 81.88% (or 84.93%) if the 2 week (or 1 day) planning horizon is used instead of 12 weeks. However, we should note that although the seen rate increases and the cancellation rate decreases significantly, the no-show rate also

increases slightly with these changes (4.20% to 4.46% (or 4.76%) compared to 4.26% in the current system). The adverse effect of this change is very small and has little impact. However, if the no-show rate is higher, then it would be important to conduct a revenue analysis evaluating the trade-off between the increase in the seen rate versus the increase in the no-show rate as a function of the planning horizon.

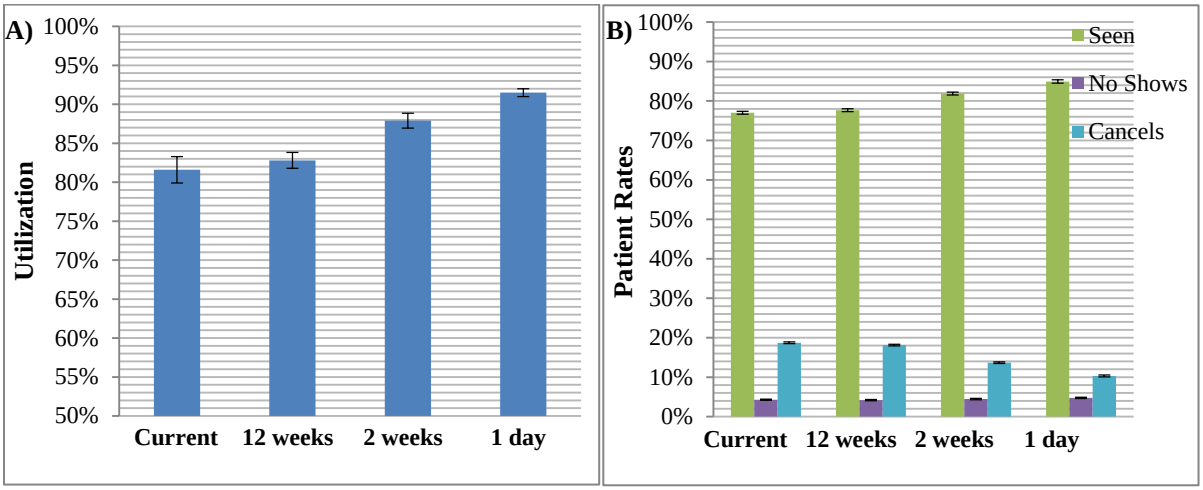

Figure 3.5: Utilization rates (A) and Patient rates (B) under different planning horizons with slotless design

In Figure 3.6B, the effect of using the slotless design, 20 minute standardized slots and 30 minute standardized slots on the utilization of physicians given that the planning horizon is 12 weeks is shown. If 30 minute slots are used rather than 20 minute slots, the utilization will be 67.1% compared to 81.6%, while the utilization for the slotless design is 82.8%. A similar effect is observed in the seen patient rates, i.e., with the slotless design the seen patient rate is 77.67%, while it is 77% for the 20 minute slots and 73.41% for the 30 minute slots (see

Figure 3.6A). Thus we can conclude that if the standardized slots are preferred by the clinic, then 20 minute slots should be used instead of 30 minute in order to maximize utilization as well as the seen patient rates.

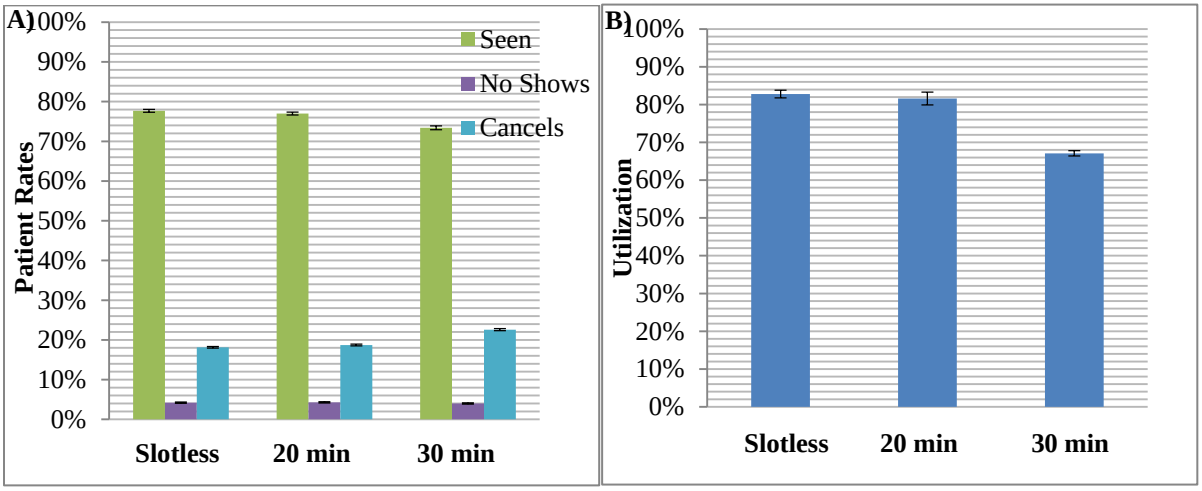

Figure 3.6: Patient rates (A) and Utilization rates (B) under different slot structures with 12 weeks horizon

After observing that 20 minute standardized slots achieve better results compared with 30 minute slots, we vary the planning horizon with 20 minute slots to see the effect of the planning horizon (see Figure 3.7). Figure 3.7 shows that while using standardized 20 min slots, reducing the planning horizon increases the seen patient rates and the utilization of physicians.

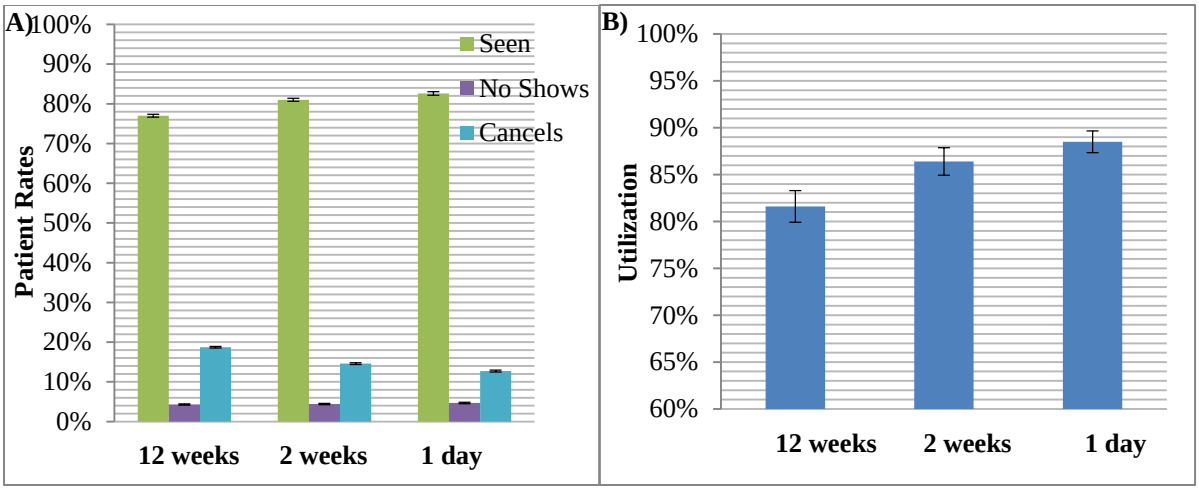

Figure 3.7: Patient rates (A) and Utilization rates (B) under different planning horizons with 20 min slots

Finally all of these scenarios are compared with respect to the average number of seen patients of each type (Established, Internal, Subsequent Visit and New External) per day. As discussed for each scenario, the seen patient rates can be improved by modifying the planning horizon and/or slot designs. Figure 3.8 shows that the average number of seen patients of all types can be increased significantly, if the slotless design with 1 day of planning horizon is used, resulting in a 9.89% increase in Established patients, 11.77% increase in Internal patients, 8.58% increase in Subsequent Visit patients, and 25.4% increase in New External patients. According to this criterion the worst setting is one with 30 minute slots and a 12 week planning horizon, which yields the fewest seen patients per day and is even worse than the current system. The reason for this is using slots of length 30 minutes results in more underutilized slots because as explained previously, there are three different appointment lengths, 20 minutes, 40 minutes and 60 minutes. If a patient requiring a 20

minute appointment is given a 30 minute slot, then 10 minutes of the physician time will be wasted. Similarly, if a patient requires a 40 minute appointment, then two 30 minute slots must be assigned to that patient, which yields 20 minutes of idle physician time that cannot be used for another appointment. Thus, inefficient usage of physician times yields lower rates of seen patients as well as lower utilization of the overall capacity.

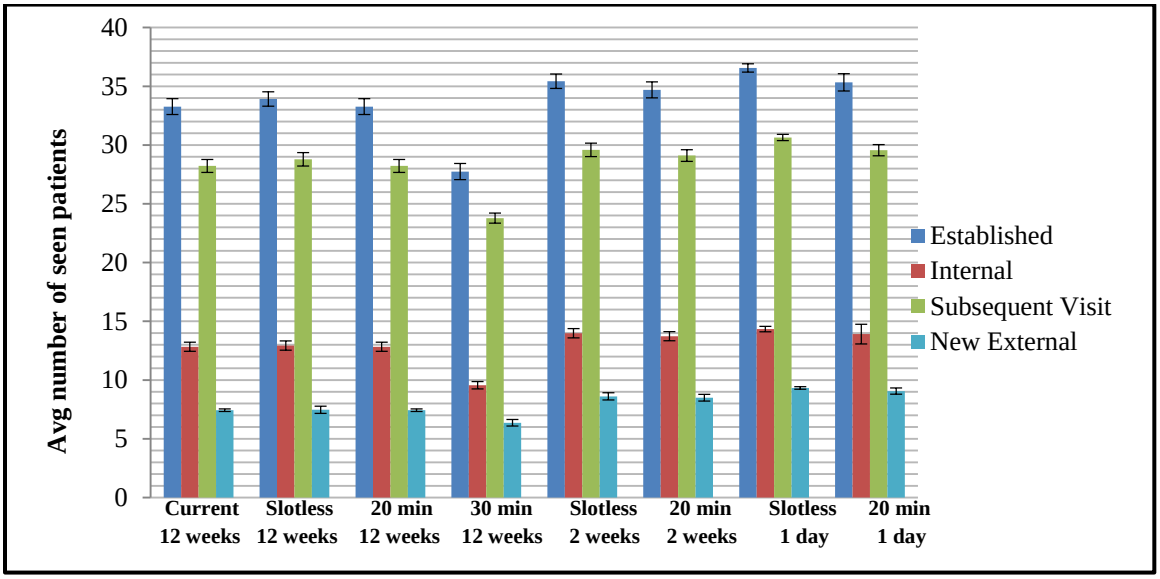

Figure 3.8: Average Number of Seen Patients per day

3.3.4 Discussion about the Model I

With this model, the effects of the slot designs and planning horizon lengths on the appointment scheduling system are evaluated using simulation. The clinic on which this study is based has multiple resources (physicians) and multiple classes of patients each of which have different delay-dependent cancellation, reschedule and no-show rates. Moreover,

each class of patients requires different service times depending on the appointment type, and has different rewards. To the best of our knowledge, this study is the first to consider this complicated problem setting and analyze various scheduling decisions. In this work, because the system that we are handling is complex, we use simulation to analyze scheduling decisions in the clinic, in order to understand the system and the benefits of using different horizon lengths and/or slot designs. We use seen patient rates and overall physician utilization as our performance indicators.

Our results indicate that we can significantly increase physician utilization and the percentage of seen patients by reducing the appointment planning horizon. As a result of our initial data analysis, we observe that the patients are less likely to make a change to their appointment when the appointment is close (recall Figure 3.3). Intuitively this can be described as follows: considering that each day is a decision point for the patient, although the chance to make a change is small each day, over many days it starts to add up. However, patients only have one decision point to not show. This may explain why no-show rates are relatively consistent regardless of how far in the future an appointment is made. Thus, the less time a patient has time to think about making a change, the less likely he/she is to make a change and the more likely the patient is to be seen, which yields higher rates of seen patients as well as higher utilization of the physician calendars.

We also observe improvements in the performance indicators, if the slotless design or standardized slots of shorter length (20 min) are used. The underlying reason for this is that, by using a slotless design we do not restrict calendars with predefined slots so different appointment length can be assigned consecutively without causing idle time. Similarly, using

20 minute slots, we obtain better results for the performance indicators compared to 30 minute slots. Depending on the type of the appointment, each patient may require a 20, 40 or 60 minute and using 30 min slots would cause physicians to have idle time, which cannot be used for another appointment, and cause the seen patient rates to be lower.

Based on the simulation analysis, we have observed that we can improve the system by reducing the appointment windows and using 20 minutes appointment slots. In the next section, we extended our simulation model to allow for different appointment windows for different patient groups in order to balance or prioritize different types of patient visits per day while removing the assumption of independent patient classes.

3.4 Simulation Model – II

The main distinctions of the Simulation Model II compared with Model I are, the dependency of the subsequent visits on the other patient classes, the consideration of patient-based revenues and penalties associated with cancellation and reschedules, and the allowance for different appointment windows for different patient classes. In the next four subsections we explain our analysis of the additional components, the second simulation model, scenarios and our findings.

3.4.1 Additional Data Analysis

In order to capture the dependency between patient classes, we use the 2-year demand data. From Section 3.2.1 we know that the average number of appointment requests per day is 17 for the internal patients, 10 for the new external patients, 36 for the established patients and 44 for the subsequent visit patients. However, the subsequent visits are generated by other

patients classes since this type of visit is due to follow up from a previous appointment. For this reason, we analyze the subsequent visit impact created by other types of patients. Based on our analysis, an internal patient visit creates 0.55 subsequent visits per day, new external patients create 1.12 subsequent visits and established patients create 0.65 subsequent visits. In our model we assume that subsequent visits are generated with Poisson distribution with the above means.

Additional data components that we use are revenues and penalties associated with seeing a patient of certain type, and cancelling or rescheduling an appointment. For these values, we use relative value units that are determined by expert opinion. We assume that each seen internal patient yields revenue of 0.8 units, a new external patient yields 3 units, a subsequent visit patient yields 2 units, and an established patient yields 1 unit. Moreover, we set the penalties for cancellations and rescheduling based on the effort that the appointment scheduler spends. We assume that cancelling or rescheduling an appointment has a penalty of 0.05 units based on an expert opinion from the clinic.

3.4.2 Model II

We develop a simulation model in Arena, whose flow chart is shown in Figure 3.9, in order to conduct an experimental study based on several scenarios. The scenarios will be explained in Section 3.4.3. In this simulation model, internal, new external and established patients arrive based on their daily appointment request rates, while subsequent patient visits are generated from these appointment types as they are follow up visits of the other types. In this model, similar to our previous model, if an appointment is cancelled or rescheduled, that slot

becomes available for other appointment requests, otherwise the patient will either be seen or will no-show based on behavioral functions of that class.

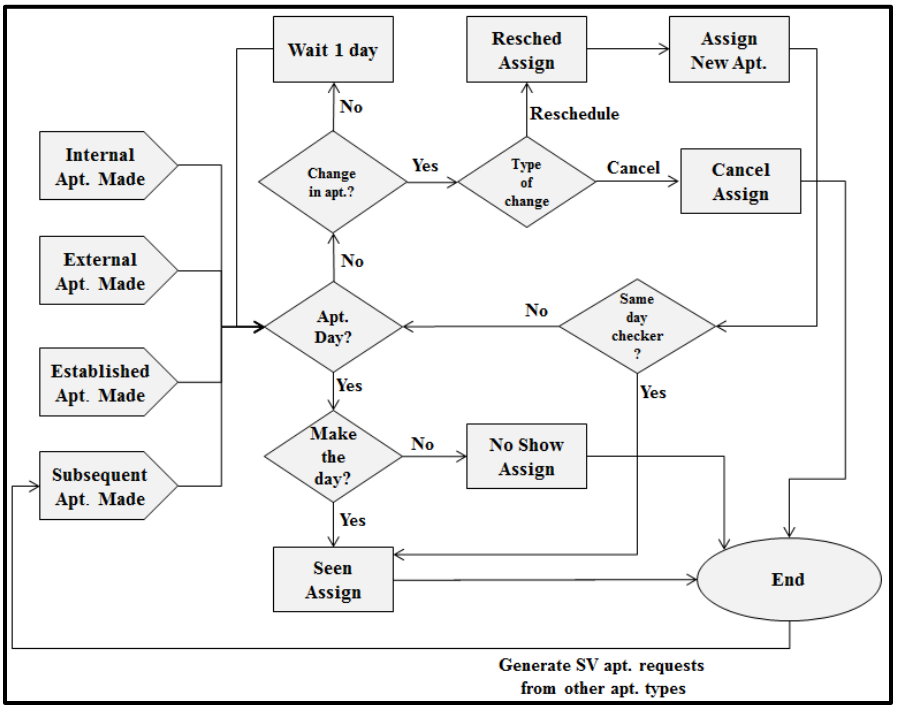

Figure 3.9: Process flow chart of the simulation model II

We first simulate the current scheduling system, in which the "current days out" functions are used to determine the assigned appointment days for incoming appointment requests, so that we can represent the current system with all its delays due to flaws in the capacity management. Appointment windows for all patient classes are 12-weeks in the current system, which we also refer to as the baseline model.

The simulation model is then modified to evaluate different scenarios for our experimental study. Under this system we do not have predefined slots of different types. Instead we have generic 20 minute slots to which any appointment can be assigned. Note that when assigning appointments, a sufficient number of slots should be assigned to the each patient based on his/her class, i.e., two slots to an internal patient, three slots to a new external patient, one slot to subsequent visit patient, one, two or three slots to an established patient depending on the appointment type he/she needs. We use this model to evaluate various scenarios with different appointment windows for different patient classes. While changing the appointment windows, we skew the current days out functions of each group by using multipliers in such a way that the maximum appointment delays will be the target appointment windows for that scenario. Note that in order to prevent infeasible cases, if there is no available slot on the assigned day, the patient is moved to the next day until an available day is found.

In each scenario explained in the next subsection, daily net profit, seen patient percentages and overall utilization of the physicians are used as the performance indicators with 95% confidence intervals which are shown by black bars in the graphs. For each scenario, we use 1000 days to warm up the system in order to reach steady state, and once the system reaches the steady state we collect results for 260 weekdays (52 weeks).

3.4.3 Scenarios for Model II

We use the attributes defined in Table 3.3 to identify the scenario set for our experimental analysis for the appointment windows (AW) for each patient class. Note that, in all of the scenarios, we kept daily appointment requests of different patient classes, patient behavior functions, and daily resource capacities the same as they are in the current system (as explained in Section 3.2.1).

Table 3.3: Appointment window choices

Patient Types	**Appointment Window Choices**	**Current System Appointment Windows**
Internal Patients (INT)	1 day, 2 days, 1 week	12 weeks for all patient types
New External Patients (EXT)	1, 3, 6 weeks	
Subsequent Visit Patients (SV)	1, 3, 6 weeks	
Established Patients (EST)	6, 9, 12 weeks	

The clinic feels that because internal patients are already in the clinic, they need to be seen as soon as possible. The main concern with the external patients is that they have not been seen previously in this clinic, and thus they do not have a commitment yet. For this reason they need to be seen quickly, but not necessary as early as the internal patients. For the subsequent visit patients, since they are usually following up on their previous visit, they need to be seen in a reasonable amount of time but this can be longer than the appointment windows for the external patients. Finally, established patients can be given the longest appointment window, as their appointment tends to be less urgent compared to the other

types. Thus, when identifying the set of possible scenarios, we need to maintain the following appointment window (AW) order between the different patient classes AW(INT) ≤ AW(EXT) ≤ AW(SV) ≤ AW(EST) so that each scenario reflects the patient admission priorities. Based on this constraint we identified 54 scenarios to test in addition to the current/baseline model in which a 12-week appointment window is used for each patient class. A complete list of scenarios is given in Table 3.4.

Table 3.4: Scenarios for appointment windows (in days) *(Green for the internal, purple for the new external, blue for the subsequent visit, and orange for the established patients)*

Scenario	INT	EXT	SV	EST	Scenario	INT	EXT	SV	EST	Scenario	INT	EXT	SV	EST
1	5	30	30	60	**19**	5	30	30	45	**37**	5	30	30	30
2	2	30	30	60	**20**	2	30	30	45	**38**	2	30	30	30
3	1	30	30	60	**21**	1	30	30	45	**39**	1	30	30	30
4	5	15	30	60	**22**	5	15	30	45	**40**	5	15	30	30
5	2	15	30	60	**23**	2	15	30	45	**41**	2	15	30	30
6	1	15	30	60	**24**	1	15	30	45	**42**	1	15	30	30
7	5	5	30	60	**25**	5	5	30	45	**43**	5	5	30	30
8	2	5	30	60	**26**	2	5	30	45	**44**	2	5	30	30
9	1	5	30	60	**27**	1	5	30	45	**45**	1	5	30	30
10	5	15	15	60	**28**	5	15	15	45	**46**	5	15	15	30
11	2	15	15	60	**29**	2	15	15	45	**47**	2	15	15	30
12	1	15	15	60	**30**	1	15	15	45	**48**	1	15	15	30
13	5	5	15	60	**31**	5	5	15	45	**49**	5	5	15	30
14	2	5	15	60	**32**	2	5	15	45	**50**	2	5	15	30
15	1	5	15	60	**33**	1	5	15	45	**51**	1	5	15	30
16	5	5	5	60	**34**	5	5	5	45	**52**	5	5	5	30
17	2	5	5	60	**35**	2	5	5	45	**53**	2	5	5	30
18	1	5	5	60	**36**	1	5	5	45	**54**	1	5	5	30

3.4.4 Results Based on Model II

For each scenario, the net profit (as a function of the relative value unit), the seen patient percentages, and the total utilization are shown in Figure 3.10a, b, c, respectively. The first observation that we can make with this experimental analysis is that, all of the proposed scenarios perform significantly better than the baseline in terms of all performance indicators with 95% confidence.

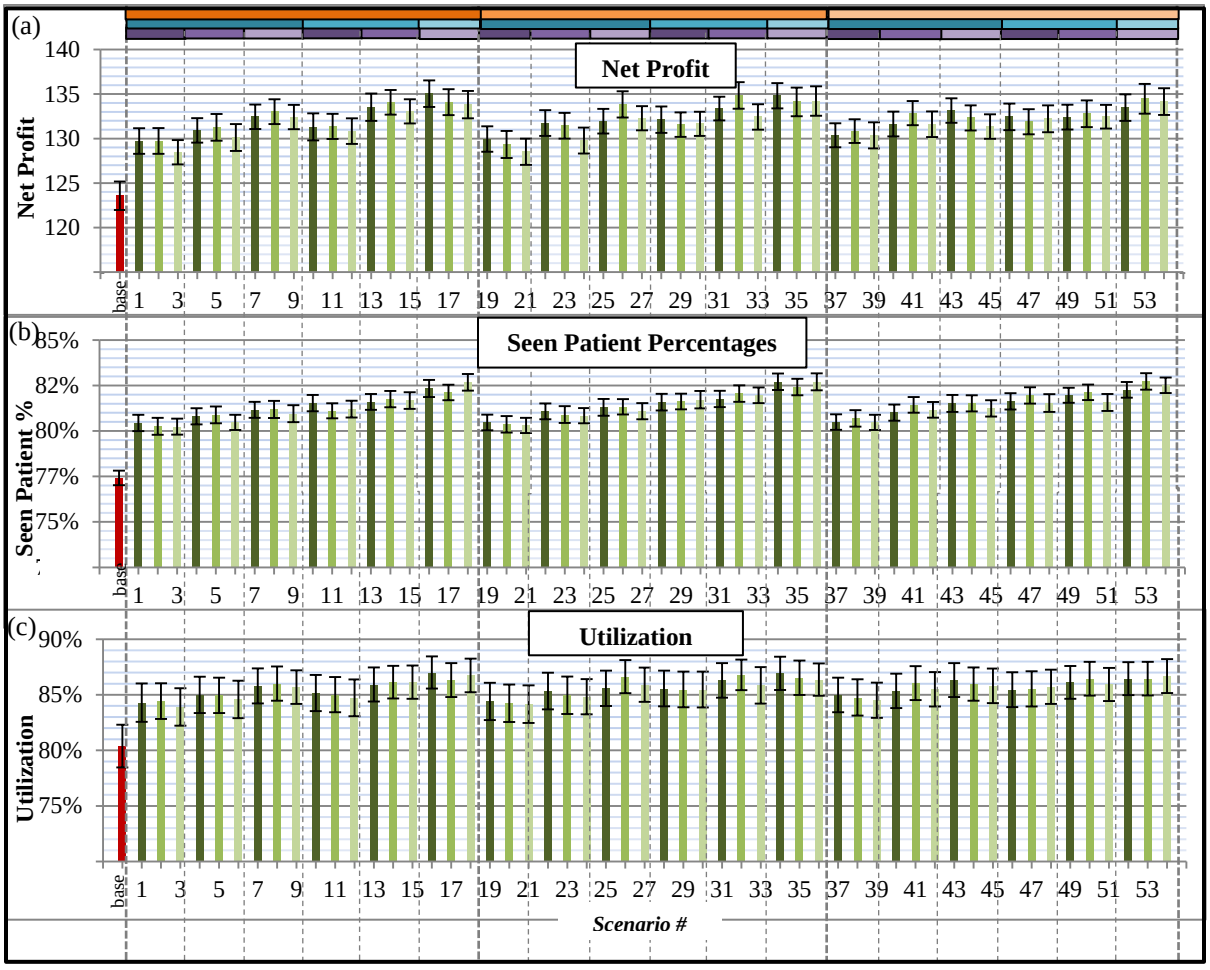

Figure 3.10: (a) Net profit per day, (b) Percentage of the seen patients per day, (c) Total utilization per day

From Figure 3.10a, we can see that the best performing scenarios, based on the Net Profit performance indicator, are scenarios 16, 32, 34, and 53 in which we choose the shortest appointment window, i.e., 5 days, for the external patients. If we look at the seen-patient percentages, there are other scenarios that perform close to these or better (see Figure 3.10b). However, as we discuss in the following subsections the increase in the overall seen-patient percentages does not necessarily indicate that those scenarios perform well in terms of the net profit. In terms of utilization, all scenarios perform well, i.e., there is no significant difference in performance, between the scenarios (see Figure 3.10c). Next we will analyze the scenarios in subsets to see the effect of changing appointment windows for each patient type.

Effects of the Internal Patients:

In order to see the effects of reducing the appointment window for the Internal Patients, we evaluate the change in the performance indicators within subsets of three scenarios in Figure 3.10a, b, and c. In other words, by comparing scenarios 1, 2, and 3, scenarios 4, 5, and 6, and scenarios 7, 8, and 9, and so on, we can assess the change in the performance indicators as we set the appointment window for internal patients to 5, 2, and 1 days, respectively, while keeping the other parameters fixed.

This analysis suggests that reducing the appointment window for the internal patients does not always increase the daily net profit, indeed the net profit decreases in most of the scenario subsets as the appointment window is decreased from 2 days to 1 day (see Figure 3.10a). The reason for this reduction is that reducing the appointment window to 1 day, forces the model to give next day appointments to all internal patients and since we observe

the same number of internal appointment requests each day, this creates a potential block in the system. Recall from Section 3.3.2 that for each patient type, after the initial appointment day is identified considering the appointment window, if that day is not available, the simulation schedules the patient to the first available day. Thus, filling up the days with the internal patients prevents other patients from being given appointments in a timely manner. In addition, as we know that the internal patients have the least revenue, filling the next day with them causes other patients, who have potentially higher revenue, to get later appointments (due to blocking) reducing the daily net profit. This is also the reason for the higher seen-patient percentages for some cases of the 1-day appointment windows compared with the 2-day appointment windows (see Figure 3.10b). Although in some cases we have higher seen-patient percentages for the 1-day case, the net profit is not higher for those cases as we increase the internal patients in the daily mix and prevent other types from having timely access to the system. A comparison of the utilization in Figure 3.10c for these subsets shows a similar pattern. Although the difference is not significant, utilization decreases in most cases as we decrease the appointment window for the internal patients.

On the other hand, although it is not significant, we note that for some cases reducing the appointment window from 5 days to 2 days improves the system in terms of daily net profit. The reason for this is that for those scenarios the capacity used for accommodating internal patients within two days is sufficient and does not cause blocking in the system, so that the other patients are not postponed further than their initially identified appointment days. Thus, the 2-day window obtains higher profit than the 5-day window by seeing more internal patients without blocking. The profit is not significantly different for seeing other types of

patients in both cases. This suggests that we can increase the daily net profit by choosing a 2-day instead of a 5-day appointment window for those settings.

Finally, we note one limitation of our model is that the net profit associated with an internal patient is more complex than we have been able to incorporate. Internal patients are referred by other departments, so the revenue that is obtained from these patients and accounted towards this specific department is low because it is being shared between different departments and we only consider a specific department's yield management problem based on that department's potential revenues. However, when the clinic-level revenues are considered, internal patients can have relatively higher revenues than they are at the department-level due to this sharing.

Effects of the New External Patients:

To evaluate the effect of decreasing the appointment window for the new external patients, we compare the scenarios within subsets of 2-3 scenarios, depending on the available appointment window choices for the new external patients. We evaluate the following subsets of scenarios where we compare scenarios within the same subset, considering the reduction of AW only for new external patients: {1, 4, 7}, {2, 5, 8}, {3, 6, 9}, {10, 13}, {11, 14}, {12, 15}, {19, 22, 25}, {20, 23, 26}, {21, 24, 27}, {28, 31}, {29, 32}, {30, 33}, {37, 40, 43}, {38, 41, 44}, {39, 42, 45}, {46, 49}, {47, 50}, and {48, 51}.

After comparing the scenarios within the above subsets, we observe that reducing the appointment window for the new external patients increases the net profit significantly in most cases, along with the seen patient percentages and utilization, although the improvements in the latter two indicators are not significant. Unlike the internal patients, the

new external patients have the highest revenue, thus although reducing their appointment window may cause blocking for other types of patients, the revenue gained by giving these patients priority is worthwhile in terms of the net profit. However, the significance of the profit increase, reduces for the subsets of scenarios in which all patient types have shorter appointment windows. This causes the earlier days to fill up rapidly, and the system pushes the new appointments to later days until it finds an available day. Similar to our discussion in the previous subsection for the internal patients, this might create blocking in the system and this is the reason those latter scenario subsets which reduce new external patients even more, do not improve the system significantly.

Effects of the Subsequent Visit Patients:

Subsequent visit patients have higher revenues (2 units) compared with the established (1 unit) and internal patients (0.8 unit), and revenues close to new external patients (3 units). Thus we can expect to have similar results to those for the new external patients, but the significance in the improvement is expected to be a bit lower than for the external patients. Because as shown in the Figure 3.1, external patients have a steeper decrease in their seen probabilities as the appointment delay increases, compared with subsequent-visit patients. In order to assess the effect of decreasing the appointment window for the subsequent visit patient, we compare the scenarios within the following subsets of scenarios: {4, 10}, {5, 11}, {6, 12}, {7, 13, 16}, {8, 14, 17}, {9, 15, 18}, {22, 28}, {23, 29}, {24, 30}, {25, 31, 34}, {26, 32, 35}, {27, 33, 36}, {40, 46}, {41, 47}, {42, 48}, {43, 49, 52}, {44, 50, 53}, {45, 51, 54}. Similar to the previous case, these subsets are formed based on the AW reduction of SV patient class. As shown in Figure 3.10a, the net profit increases significantly with 95%

confidence in most cases, as the appointment window decreases for subsequent visit patients. The seen patient percentages and the utilization have a similar pattern, although the change in the utilization is not significant with 95% confidence (see Figure 3.10b and c).

Effects of the Established Patients:

Finally, to assess the effect of decreasing the appointment window for established patients we compare the scenarios within the following subsets: {1, 19, 37}, {2, 20, 38}, {3, 21, 39}, {4, 22, 40}, {5, 23, 41}, {6, 24, 42}, {7, 25, 43}, {8, 26, 44}, {9, 27, 45}, {10, 28, 46}, {11, 29, 47}, {12, 30, 48}, {13, 31, 49}, {14, 32, 50}, {15, 33, 51}, {16, 34, 52}, {17, 35, 53}, and {18, 36, 54}.

In this case, we cannot directly conclude that the performance indicators always increase or decrease as we reduce the appointment window. This can be explained by the fact that the established patient type has lower revenue compared with the new external and subsequent visit patients, but higher revenue than the internal patients. Moreover, when we look at the behavioral functions for established patients, we observe that shape of curve for the probability of being seen is flatter than all other patient types. For this reason, reducing the appointment window for this patient class does not necessarily increase the net profit. In some cases it can cause blocking in the system which prevents other patients from being admitted in a timely manner, reducing the net profit. Figure 3.10a, b, and c suggest if external patients have 5-day appointment window, reducing the appointment window for established patients actually reduces the net profit, although the reduction not significant. By having appointment windows for the external patients at the lowest value, the early days fill up rapidly and if the established patients' appointment window is reduced, this creates blocking

due to capacity restrictions, so that the patients with higher revenues (compared to established patients) are delayed until an available slot is found. For the other scenarios there is not a significant increase in the performance indicators. Thus, given the lower revenue and the flatter seen probability curve, reducing the appointment window for this patient type is not worthwhile.

3.4.5 Discussion about the Model II

In the Model II, we consider an outpatient clinic with multiple physicians and multiple patients each of which has different revenue, different appointment request rate, different service time and different behavioral functions (cancellation, reschedule, no-show, seen) that are appointment delay dependent. Under this setting, we evaluate the effects of changing appointment windows for different patient types with the performance indicators: net profit, seen patient percentages and overall utilization of physicians. Due to the complexity of the system we use simulation with Arena in order to understand the system and explore the advantages and disadvantages of using different appointment windows for each patient class.

Based on our results we conclude that it is possible to significantly improve all of our performance indicators by using the scenarios, in which the largest appointment window is 1 week for internal patients (Scenario 2), 6 weeks for new external patients, 6 weeks for subsequent visit patients, and 12 weeks for established patients. By making these changes in the current system we can improve the clinic performance indicators significantly. However, although all scenarios perform better than the current system, if we compare the proposed scenarios, we cannot directly conclude that reducing appointment windows always increases

the performance indicators. For example, for the internal patients reducing the appointment windows to 1 day is actually worse than 2 days. We can explain this by the fact that filling the next day with internal patients would cause other patients to be unable to enter the system in a timely manner, and given the lower revenue obtained from internal patients, it is not worthwhile to give them next day appointments. However, with new external patients we observe an increase in the net profit in most cases as we reduce the appointment window. This is due to high revenue that these patients bring into the clinic which compensates for delaying patients of other types. It is important for the clinic to prioritize their performance indicators while considering the trade-off between scheduling a patient of a certain class to an earlier versus a later day by taking their blocking effect on the system into account.

An interesting area for future research is to incorporate the fact that established patients are also generated by new external patients and internal patients. In this chapter we have not explored the relationship between established patients and new external patients and internal patients due to data limitations. However, by defining established patients as functions of other patient types we can explore the effect of the dependency when increasing the load on the system.

Based on our findings from these simulation models, next we build mathematical programming models to find optimal appointment windows and allocation policies for different patient classes.

Chapter 4

Mathematical Programming Models

In this chapter analytical models, to find the optimal scheduling and capacity allocation policies, are presented. After the introduction in Section 4.1, in Section 4.2 we formulate an initial mathematical programming model for which we find structural properties to enable ease of solutions. In Section 4.3 we extend the model presented in the Section 4.2 by relaxing the assumption related to reschedules. Section 4.4 includes a simulation comparison of the current policy with the optimal policy. In Section 4.5 we consider the structures of the actual behavior functions used in Section 4.4, and Section 4.6 concludes this chapter.

4.1 Introduction

As we discussed in Chapter 2, there is a wealth of literature in the field of patient scheduling and capacity allocation. However, to the best of our knowledge none of these papers consider patient reschedule behavior (with either fixed or delay-based rates), or delay-based no-show and cancellation behaviors. Several papers model the patient appointment system as MDPs, which suffer from the curse of dimensionality and as a result these papers develop heuristics to approximate policies. These models are large even without considering reschedules or other delay-based behaviors (e.g., cancellation or no-show). To address this problem, in this chapter we develop an analytical approach for finding the optimal capacity allocation policy while taking delay-based behaviors into account. Our aim with these models is to make decisions at a more strategic level, for this reason we do not consider the appointment times but instead we only consider the days to which we assign an appointment.

The models developed in this chapter are based on an outpatient setting in which there are multiple classes of patients with different appointment durations, different delay-dependent cancellation, reschedule, no-show, and seen probability functions, different revenues for the seen patients, and different penalties for no-shows, cancellations and reschedules. In this context, an appointment delay refers to the time between the appointment request date and the actual appointment date. We assume that the subspecialties in the outpatient clinic are independent, and based on this assumption we focus on one subspecialty. This model considers a single physician pool with the same subspecialty with a fixed daily capacity, regardless of the day of the week. The objective is to find the optimal

daily capacity allocation to each class of patients and the optimal patient scheduling policy so that the net revenue is maximized. Note also that, in this chapter we assume stationary demand (different for each patient class) over different days and focus on the effects of the delay dependent patient behaviors. Thus in the models presented, we consider a fixed rate of new daily appointment requests, and delay-based probabilities of canceling, rescheduling and no-show.

In this problem setting we assume the patients are punctual, i.e., if a patient shows up on the appointment day, he/she is always on time. We also assume that the service times are deterministic. Since we only consider appointment days, not the appointment times, based on the daily capacities we are filling the days with a certain number of appointments for a certain patient class, ignoring the sequence of appointments within a day. We also assume that the patients do not have preferences regarding their initial appointment day, however we incorporate patient preferences for the appointment days, if patient decides to reschedule his/her appointment.

At the time an appointment is being scheduled, we know the likelihood that the patient will show up (or not), cancel or reschedule his/her appointment as a function of the appointment delay. Therefore, we define the state as the "expected number of patients that will not make any changes in their appointments until the appointment day, once the appointment is scheduled".

In this chapter we present two mathematical programming models for the above described outpatient clinic setting. The models differ in the way the rescheduled appointments are handled.

4.2 Model I

In our first model, we assume that the reschedule requests arriving on Day t are from patients who have been scheduled for an appointment on Day t, but want to reschedule the appointment to another day. Figure 4.1 shows a timeline characterizing the relationship between reschedule requests, the original (initial) appointment and the rescheduled appointment. As shown in this figure, initially new appointment requests are observed on Day t and appointments are assigned to future dates (shown by the dashed lines in blue) within a $[t, t+L]$ range. This constraint ensures that the latest appointment is not after the maximum allowed days in the future $(t+L)$, where t is the current day and L is the maximum allowed delay. At this point there is a chance, the assigned appointment will be requested to be rescheduled as a function of appointment delay (shown by solid lines in orange). Then the rescheduled appointments are assigned to days that are in the original range $[t, t+L]$ (shown by dotted lines in green) based on the patient's preferences for the L days. If the patient has no preference, then we assume that all the following L days are equally likely to be chosen.

The optimal policy for assigning incoming appointment requests to the future days is identified using Model 4.1. This policy provides the distribution for the percentage of the daily capacity that is to be allocated to patient class i. Although we define the daily demand in terms of days, we assume that the demand is the same for each day of the week, so $N_{i,t} = N_i$ in this formulation. However, the index t is kept in the model formulation for the reschedule calculations. Because we assume that demand is the same for each day of the

week, the optimal policy does not depend on the day. We define the variable a_{il} as the percentage of the initial appointment requests by patients of type i that should be scheduled l days (where $0 \leq l \leq L_i$) into the future. For example, for a two patient class setting with $L_1 = L_2 = 4$ if the optimal solution is $a_{11} = 1, a_{21} = 0.2, a_{22} = 0, a_{23} = 0, a_{24} = 0.8$, then according to this policy, for the requests coming from type 1 patients next day appointments should be scheduled, and for the requests coming from type 2 patients 20% of appointments should be scheduled to next day and 80% of the appointments should be scheduled 4 days out.

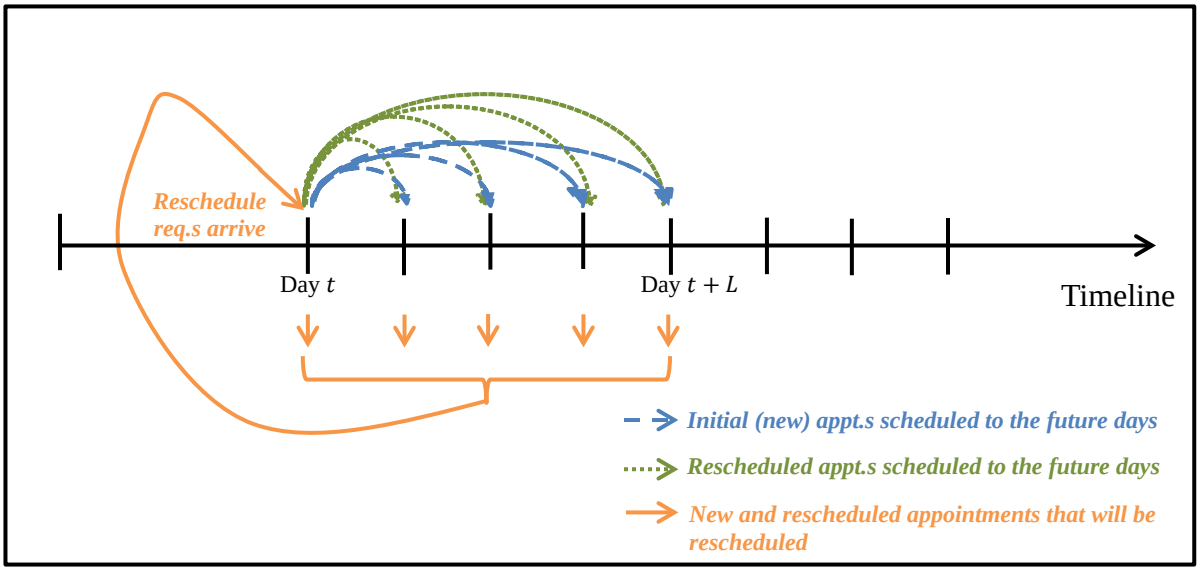

Figure 4.1: Characterization of the Reschedules

Table 4.1: Notation for Model 4.1

Decision variables:	
a_{il}	Percentage of daily appointment requests to be scheduled l days later for patient type i ($l = 0, .., L$ where $L \leq T$, $i = 1, .., M$)
Model parameters:	
α_i^{se}	Revenue obtained from each seen patient of type i ($\alpha_i^{se} \geq 0$)
β_i^{no}	Penalty for no-show per patient of type i ($\beta_i^{no} \geq 0$)
β_i^{ca}	Penalty for cancellation per patient of type i ($\beta_i^{ca} \geq 0$)
β_i^{re}	Penalty for rescheduling per patient of type i ($\beta_i^{re} \geq 0$)
$p_i^{nc}(l)$	Probability of not making any changes for patient of type i that scheduled an appt. with a delay of l days
$p_i^{ca}(l)$	Probability of cancellation for patient of type i that scheduled an appt. with a delay of l days
$p_i^{re}(l)$	Probability of rescheduling for patient of type i that scheduled an appt. with a delay of l days
$p_i^{no}(l)$	Probability of not showing up for patient of type i that scheduled an appt. with a delay of l days
$p_i^{se}(l)$	Probability of being seen for patient of type i that scheduled an appt. with a delay of l days
e_{il}	Probability that a type i patient wants an appointment l days later than the initial date of appointment by rescheduling it, where $\sum_{l=0}^{L_i} e_{il} = 1 \; \forall i \in \{1, \dots, M\}$
K	Total capacity (in terms of time units) available for all types of patients in each day
s_i	Service time per patient of type i ($s_i > 0$)
$N_{i,t}$	New appointment requests per patient of type i on day t ($N_{it} \geq 0$)
L_i	Maximum allowable appointment delay for patients of type i
T	Planning horizon (a multiple of L_i)
M	Total number of patient types
Other notations:	
$R_{i,t}$	Number of type i patients that requested their appointment be rescheduled on day t

Model 4.1:

$$\max \sum_{i=1}^{M} \sum_{t=1}^{T} \Big(\sum_{l=0}^{L_i} \big(R_{i,t-l} e_{il} + N_{i,t-l} a_{il} \big) \big(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l) \big) \Big) \tag{4.1}$$

$s.t.$

$$\sum_{l=0}^{L_i} \big(N_{i,t} a_{il} + R_{i,t} e_{il} \big) p_i^{re}(l) - R_{it} = 0 \qquad \forall i = 1, \dots, M, t = 1, \dots, T \tag{4.2}$$

$$\sum_{i=1}^{M} \sum_{l=0}^{L_i} p_i^{nc}(l) \big(N_{i,t-l} a_{il} + R_{i,t-l} e_{il} \big) s_i - K_t \leq 0 \qquad \forall t = 1, \dots, T \tag{4.3}$$

$$R_{i,t} + N_{i,t} = R_{i,t+T} + N_{i,t+T} \qquad \forall i = 1, \dots, M, t = 1 - L_i, \dots, 0 \tag{4.4}$$

$$\sum_{l=1}^{L_i} a_{il} = 1 \qquad \forall i = 1, \dots, M \tag{4.5}$$

$$a_{i0} = 0 \qquad \forall i = 1, \dots, M \tag{4.6}$$

$$1 \geq a_{il} \geq 0 \qquad \forall i = 1, \dots, M, l = 1, \dots, L_i \tag{4.7}$$

Our objective (4.1) is to maximize the expected net revenue which consists of the revenue gained for each seen patient, penalty cost associated with each no-show, cancellation and reschedule. The second part of the objective function, i.e, $\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l)$, represents the expected net profit per patient of type i given that his/her appointment delay is l, which is assumed to be always nonnegative. It is assumed that, if an appointment that was scheduled for day t is cancelled, rescheduled, or became a no-show, then the penalty associated with this event is reflected on day t. The penalty cost

associated with each no-show is due to the idle time of the physician. When a patient does not show up for his/her appointment then because that slot was reserved for that patient, the physician cannot fill that slot with another patient. The penalty cost associated with canceling and rescheduling are related to the opportunity lost due to the time the cancelled or rescheduled patient occupies on the calendar and the effort associated with changing the appointment. Note that rescheduled patients are not lost but scheduled to other days, thus the opportunity lost will be less than that of cancelled patients.

Equation (4.2) is used to calculate the reschedule requests on day t. Because the probability, $p_i^{re}(l)$, the patient will reschedule his appointment (either an initial or a rescheduled appointment) to another day within $[t, t + L_i]$ is known with certainty, we can calculate the expected number of reschedule requests on day t by Equation (4.2). Equation (4.2) ensures that sum of: (i) the expected number of type i appointments that were initially requested on day t (and scheduled to day $t + l$) then rescheduled to another day and (ii) the expected number of previously rescheduled appointments of type i which were requested to be rescheduled to another day, is equal to the expected number of reschedule requests of type i on day t. Equation (4.3) is the capacity constraint which ensures that the expected number of patients, who have appointments on day t and do not make any changes prior to their appointment, multiplied by their respective service times is less than the capacity of day t. This constraint takes both the rescheduled appointments for day t and the new scheduled appointments for day t into account. Equation (4.4) restricts the range of $R_{i,t}$ values such that t must be within the bounds $[1, T]$. In the next paragraph, we show that there are identical cycles of length T. This makes it possible to represent the out of bound values using Equation

(4.4) and remove Equation (4.4) from the formulation. Equation (4.5) ensures that sum of the percentage of the daily appointment requests of each type that are assigned to the future days is equal to 1. This ensures that all appointment requests are satisfied by assigning the appointments to the future days. Equation (4.6) prevents the assignment of a same day appointment. Equation (4.7) restricts the a_{il} variables to be within the range of $[0,1]$ because a_{il} represent the percentage of demand assigned to future days.

We can show that there are identical cycles of length T, which corresponds to the planning horizon (see Figure 4.2). In Figure 4.2, the appointment requests (whether for an initial appointment or a rescheduled appointment) coming in within the dashed area affect the number of scheduled appointments on the day that is identified by the arrow in the figure. In Figure 4.2a the dashed area on Cycle 1 affects the number of scheduled appointments of the first day of the Cycle 2. However, if there are identical cycles of length T, then based on Figure 4.2a′, the dashed area at the end of the cycle actually affects the number of scheduled appointments on the first day of the cycle. Similarly, in Figure 4.2b, the number of appointment requests observed during the last part of Cycle 1 and the first day of Cycle 2 affect the number of scheduled appointments on the second day of the Cycle 2. If it is assumed that there are identical cycles, we actually have the movement shown by Figure 4.2b′, in which the dashed area affects the number of scheduled appointments on the second day of the cycle.

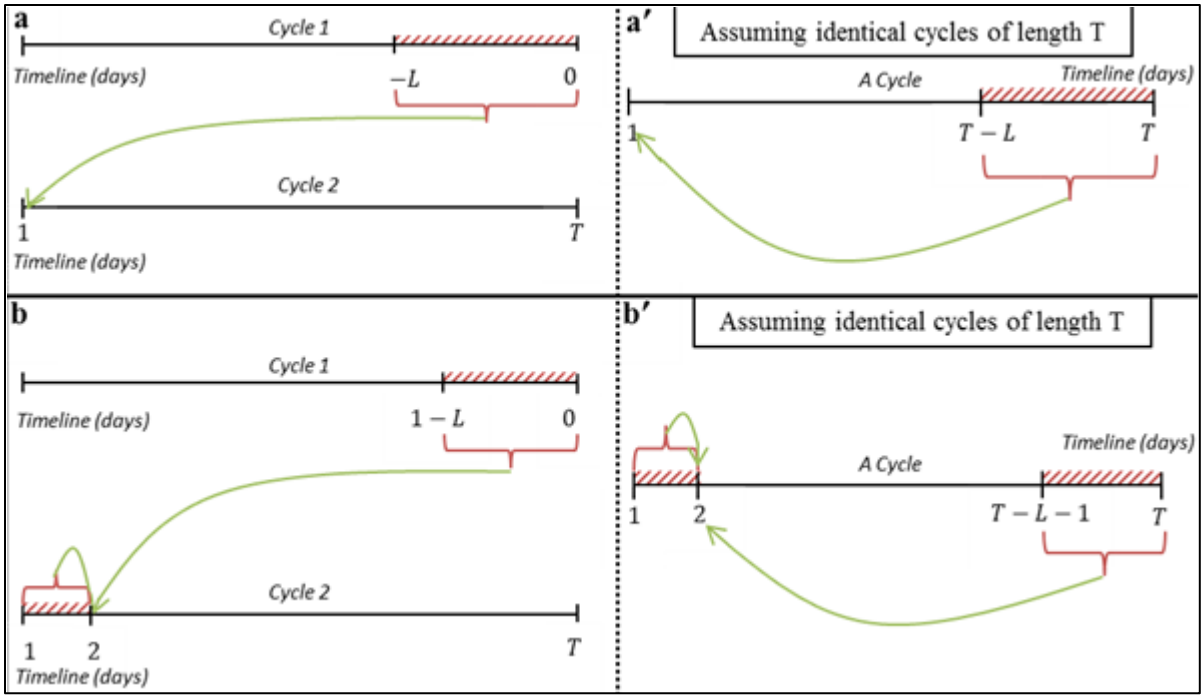

Figure 4.2: Identical Cycles

Figure 4.2a′ is equivalent to Figure 4.2a and Figure 4.2b′ is equivalent to Figure 4.2b, under the assumption of identical cycles of length T

In order to show that the cycles are identical, we use the expected number of appointments per day and show that this value is the same for both day t and day $t + T$ for any given t. For this model, it is sufficient to show this, to suggest identical cycles, because the decision variables as well as the behavior functions are same for all t and the only component that depends on t is the new appointment request. In other words, when an appointment is set to a given date, the later movements only depend on the appointment delay. For this reason, if we can show that the only components (based on t) that affects the number of appointments assigned to a given day t, are the appointment requests occurring

over a period, the only condition that we need to satisfy would be that the new appointment requests are the same in both periods.

Before calculating the expected number of appointments on day t, we first derive $R_{i,t}$ from the Equation (4.2) and obtain:

$$R_{i,t} = \sum_{l=0}^{L_i}\left(R_{i,t}e_{il} + N_{i,t}a_{il}\right)p_i^{re}(l)$$

$$R_{i,t} = \frac{N_{i,t}\left(\sum_{l=0}^{L_i} a_{il}p_i^{re}(l)\right)}{1-\sum_{l=0}^{L_i} e_{il}p_i^{re}(l)} \tag{4.8}$$

Next using Equation (4.8), the expected number of appointments on day t is:

$$E[Number\ of\ appointments\ on\ day\ t] = \sum_{i=1}^{M}\sum_{l=0}^{L_i}\left(R_{i,t-l}e_{il} + N_{i,t-l}a_{il}\right)$$

$$= \sum_{i=1}^{M}\sum_{l=0}^{L_i}\left(\frac{N_{i,t-l}\left(\sum_{k=0}^{L_i} a_{ik}p_i^{re}(k)\right)}{1-\sum_{k=0}^{L_i} e_{ik}p_i^{re}(k)}e_{il} + N_{i,t-l}a_{il}\right) \tag{4.9}$$

Similarly the expected number of appointments on day $t+T$ is:

$$E[Number\ of\ appointments\ on\ day\ t+T] = \sum_{i=1}^{M}\sum_{l=0}^{L_i}\left(R_{i,t+T-l}e_{il} + N_{i,t+T-l}a_{il}\right)$$

$$= \sum_{i=1}^{M}\sum_{l=0}^{L_i}\left(\frac{N_{i,t+T-l}\left(\sum_{k=0}^{L_i} a_{ik}p_i^{re}(k)\right)}{1-\sum_{k=0}^{L_i} e_{ik}p_i^{re}(k)}e_{il} + N_{i,t+T-l}a_{il}\right) \tag{4.10}$$

Equations (4.9) and (4.10) suggest that, if the number of new appointment requests is the same for day t and $t+T$, then there will be identical cycles. Thus, instead of dealing with an

infinite horizon problem, by defining identical cycles, we will solve a problem with a horizon length of T.

Model 4.1 can be simplified by using (4.8) and removing constraint (4.2) to obtain the following model:

$$\max \overbrace{\sum_{t=1}^{T}\sum_{i=1}^{M}\left(\sum_{l=0}^{L_i}\left(\frac{N_{i,t-l}\left(\sum_{l=0}^{L_i}a_{il}p_i^{re}(l)\right)}{1-\sum_{l=0}^{L_i}e_{il}p_i^{re}(l)}e_{il}+N_{i,t-l}a_{il}\right)\left(\alpha_i^{se}p_i^{se}(l)-\beta_i^{no}p_i^{no}(l)-\beta_i^{ca}p_i^{ca}(l)-\beta_i^{re}p_i^{re}(l)\right)\right)}^{Z} \tag{4.11}$$

$$s.t.$$

$$\sum_{i=1}^{M}\sum_{l=0}^{L_i}p_i^{nc}(l)\left(\frac{N_{i,t-l}\left(\sum_{l=0}^{L_i}a_{il}p_i^{re}(l)\right)}{1-\sum_{l=0}^{L_i}e_{il}p_i^{re}(l)}e_{il}+N_{i,t-l}a_{il}\right)s_i-K_t\leq 0 \qquad \forall t=1,\dots,T \tag{4.12}$$

$$\frac{N_{i,t}\left(\sum_{l=0}^{L_i}a_{il}p_i^{re}(l)\right)}{1-\sum_{l=0}^{L_i}e_{il}p_i^{re}(l)}+N_{i,t}=\frac{N_{i,t+T}\left(\sum_{l=0}^{L_i}a_{il}p_i^{re}(l)\right)}{1-\sum_{l=0}^{L_i}e_{il}p_i^{re}(l)}+N_{i,t+T} \qquad \forall i=1,\dots,M, t=1-L_i,\dots,L_i \tag{4.13}$$

$$\sum_{l=1}^{L_i}a_{il}=1 \qquad \forall i=1,\dots,M \tag{4.14}$$

$$a_{i0}=0 \qquad \forall i=1,\dots,M \tag{4.15}$$

$$1\geq a_{il}\geq 0 \qquad \forall i=1,\dots,M, l=1,\dots,L_i \tag{4.16}$$

Equation (4.13) can be removed from the model because it is redundant as we assume identical cycles of length T, the only variables that are different on each side of the equality

are $N_{i,t}$ and $N_{i,t+T}$ (which are actually equal). Under the assumption of stationary daily demands, $N_{i,t} = N_i$. Moreover, removing the index t from $N_{i,t}$ yields T times "Z", because under the assumption of stationary daily demand, the expected profit is the same for each day of the week, i.e., the objective function basically sums up the same expression T times. Finally, under the assumption of stationary daily capacity we can substitute $K_t = K \ \forall t = 1, \dots, T$. This reduces to Model 4.1′.

Model 4.1':

$$\max \ T \sum_{i=1}^{M} \left(\sum_{l=0}^{L_i} \left(\frac{N_i \left(\sum_{k=0}^{L_i} a_{ik} p_i^{re}(k) \right)}{1 - \sum_{k=0}^{L_i} e_{ik} p_i^{re}(k)} e_{il} + N_i a_{il} \right) \left(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l) \right) \right) \tag{4.17}$$

$$s.t.$$

$$\sum_{i=1}^{M} \sum_{l=0}^{L_i} p_i^{nc}(l) \left(\frac{N_i \left(\sum_{k=0}^{L_i} a_{ik} p_i^{re}(k) \right)}{1 - \sum_{k=0}^{L_i} e_{ik} p_i^{re}(k)} e_{il} + N_i a_{il} \right) s_i \leq K \tag{4.18}$$

$$\sum_{l=1}^{L_i} a_{il} = 1 \qquad \forall i = 1, \dots, M \tag{4.19}$$

$$a_{i0} = 0 \qquad \forall i = 1, \dots, M \tag{4.20}$$

$$1 \geq a_{il} \geq 0 \qquad \forall i = 1, \dots, M, l = 1, \dots, L_i$$

The above model explicitly shows that the coefficients for all a_{il}s are nonnegative both in the objective function (4.17) and the constraint (4.18). (Note that the expressions for these coefficients are presented later in this section in (4.32) and (4.33)). Thus, we have a variation

of the Knapsack Problem, called the Multiple Choice Knapsack Problem – MCKP (Kellerer et al., 2004 and Martello et al., 1990) and has the following general form:

$$\max \sum_{i=1}^{M} \sum_{j \in N_i} p_{ij} x_{ij} \tag{4.22}$$

$$s.t. \tag{4.23}$$

$$\sum_{i=1}^{M} \sum_{j \in N_i} w_{ij} x_{ij} \leq c \tag{4.24}$$

$$\sum_{j \in N_i} x_{ij} = 1 \qquad \forall i = 1, \dots, M$$

$$x_{ij} \in \{0,1\} \qquad \forall i = 1, \dots, M,\ j \in N_i \tag{4.25}$$

After relaxing the integrality constraint (4.25) on x_{ij} by changing the constraint to be $0 \leq x_{ij} \leq 1 \ \ \forall i, j$, the linear multiple choice knapsack problem (LMCKP) is obtained. Note that in the original formulation of LMCKP, each item j from the set N_i has a profit p_{ij} and a weight w_{ij}, where c represents the total capacity. Intuitively, the MCKP model chooses exactly one item from each set, and the LMCKP model chooses items so that their fraction adds up to 1 for each set. Similarly our model in its simplest form is:

$$\max \sum_{i=1}^{M} \sum_{l=1}^{L_i} p_{il} a_{il} \tag{4.26}$$

$$s.t. \tag{4.27}$$

$$\sum_{i=1}^{M} \sum_{l=1}^{L_i} w_{il} a_{il} \leq K$$

$$\sum_{l=1}^{L_i} a_{il} = 1 \qquad \forall i = 1, \dots, M \tag{4.28}$$

$$0 \leq a_{il} \leq 1 \qquad \forall i = 1, \dots, M,\ l \in L_i \tag{4.29}$$

In this model, p_{il} stands for the expected net revenue gained, by assigning an appointment for a patient of type i, to l days after the day the appointment request arrives; and w_{il} stands for the expected capacity required for a patient of type i, given that his/her appointment is scheduled l days out. Intuitively, this model determines how far out the appointment for a patient of type i should be scheduled by assigning percentages to the a_{il} variables, which must sum to one (when summed over l) in order to satisfy all appointment requests. We make the following assumptions about capacity which are also valid for the original LMCKP formulation.

Assumption 4.1:

(a) In order to eliminate trivial or infeasible cases we assume that

$$\sum_{i=1}^{M} \min_{l=1..L_i} w_{il} \leq K < \sum_{i=1}^{M} \max_{l=1..L_i} w_{il} \,, \tag{4.30}$$

(b) We also assume that $\exists l$ such that

$$w_{il} + \sum_{h=1..M, h\neq i} \min_{k=1..L_i} w_{hk} \leq K \quad \forall i = 1, \dots, M \,, \tag{4.31}$$

to ensure that there will be a feasible solution for $a_{il} = 1$.

By rewriting the expressions for p_{il}s and w_{il}s we obtain the following equations:

$$\begin{aligned} w_{il} &= s_i p_i^{nc}(l) N_i + s_i p_i^{re}(l) \sum_{k=0}^{L_i} p_i^{nc}(k) \left(\frac{e_{ik} N_i}{1 - \sum_{m=0}^{L_i} e_{im} p_i^{re}(m)} \right) \\ &= s_i p_i^{nc}(l) N_i + \left(\frac{s_i N_i p_i^{re}(l)}{1 - \sum_{m=0}^{L_i} e_{im} p_i^{re}(m)} \right) \sum_{k=0}^{L_i} p_i^{nc}(k) e_{ik} \end{aligned} \tag{4.32}$$

$$\begin{aligned} p_{il} = {} & T N_i \left(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l) \right) + \\ & T \sum_{k=0}^{L_i} \left(\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k) \right) \left(\frac{e_{ik} N_i p_i^{re}(l)}{1 - \sum_{m=0}^{L_i} e_{im} p_i^{re}(m)} \right) \end{aligned}$$

$$= TN_i\left(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l)\right) + \tag{4.33}$$

$$\left(\frac{TN_i p_i^{re}(l)}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\sum_{k=0}^{L_i}\left(\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k)\right) e_{ik}$$

Using these expressions structural properties for the feasible and optimal solutions can be obtained. First we define the conditions under which there are LP-dominated options for the appointment lag (l). The following proposition follows from the LP-dominance definition given by Kellerer et al. (2004).

Proposition 4.1:

(a) For the two appointment lag options $k, l \in \{1, \dots, L_i\}$ for a patient of type i, if $w_{ik} \leq w_{il}$ and $p_{ik} \geq p_{il}$ then the appointment lag of l is dominated by the appointment lag of k for the patient of type i. This relation is extended for the three options case in such a way that if appointment lags of $k, l, m \in \{1, \dots, L_i\}$ with the weights $w_{ik} \leq w_{il} \leq w_{im}$ and the net revenues $p_{ik} \leq p_{il} \leq p_{im}$ satisfy

$$\frac{p_{il} - p_{ik}}{w_{il} - w_{ik}} \leq \frac{p_{im} - p_{il}}{w_{im} - w_{il}} \tag{4.34}$$

then the option l is LP-dominated by the options k and m.

(b) Moreover, based on Proposition 11.2.2 (Kellerer et al., 2004), if option l is LP-dominated by the options $k, m \in \{1, \dots, L_i\}$ then there exists an optimal solution to Model 2 with $a_{il} = 0$.

For different appointment lags for a patient of type i, the only elements that are different are $p_i^{nc}(l)$ in the first term and $p_i^{re}(l)$ in the second term of w_{il}. Similarly for p_{il}, the net revenue per patient of type i, i.e., $\alpha_i^{se}p_i^{se}(l) - \beta_i^{no}p_i^{no}(l) - \beta_i^{ca}p_i^{ca}(l) - \beta_i^{re}p_i^{re}(l)$ and $p_i^{re}(l)$ are the only terms that are affected by the appointment lags. Based on this, expressions (4.32) and (4.33) can be rearranged as follows, where the parameters that are a function of the appointment lag of l are written in bold:

$$
\begin{aligned}
w_{il} &= \boldsymbol{p_i^{nc}(l)}s_iN_i + \boldsymbol{p_i^{re}(l)}\left(\frac{s_iN_i}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)\sum_{k=0}^{L_i} p_i^{nc}(k)e_{ik} \\
&= \left(1 - \boldsymbol{p_i^{re}(l)} - \boldsymbol{p_i^{ca}(l)}\right)s_iN_i + \boldsymbol{p_i^{re}(l)}\left(\frac{s_iN_i}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)\sum_{k=0}^{L_i} p_i^{nc}(k)e_{ik} \\
&= \left(1 - \boldsymbol{p_i^{ca}(l)}\right)s_iN_i + \boldsymbol{p_i^{re}(l)}s_iN_i\left(\left(\frac{\sum_{k=0}^{L_i}\left(1-p_i^{re}(k)-p_i^{ca}(k)\right)e_{ik}}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right) - 1\right) \\
&= \left(1 - \boldsymbol{p_i^{ca}(l)}\right)s_iN_i + \boldsymbol{p_i^{re}(l)}s_iN_i\left(\frac{\sum_{k=0}^{L_i} e_{ik} - \sum_{k=0}^{L_i}\left(p_i^{re}(k)+p_i^{ca}(k)\right)e_{ik} - 1 + \sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right) \\
&= \left(1 - \boldsymbol{p_i^{ca}(l)}\right)s_iN_i + \boldsymbol{p_i^{re}(l)}s_iN_i\left(\frac{-\sum_{k=0}^{L_i}\left(p_i^{re}(k)+p_i^{ca}(k)\right)e_{ik} + \sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right) \\
&= \left(1 - \boldsymbol{p_i^{ca}(l)}\right)s_iN_i + \boldsymbol{p_i^{re}(l)}s_iN_i\left(\frac{-\sum_{k=0}^{L_i} e_{ik}p_i^{re}(k) - \sum_{k=0}^{L_i} e_{ik}p_i^{ca}(k) + \sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right) \\
&= \left(1 - \boldsymbol{p_i^{ca}(l)}\right)s_iN_i + \boldsymbol{p_i^{re}(l)}s_iN_i\left(\frac{-\sum_{k=0}^{L_i} e_{ik}p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)
\end{aligned}
$$

$$
w_{il} = s_iN_i - \boldsymbol{p_i^{ca}(l)}s_iN_i + \boldsymbol{p_i^{re}(l)}s_iN_i\left(\frac{-\sum_{k=0}^{L_i} e_{ik}p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right) \tag{4.35}
$$

$$
\begin{aligned}
p_{il} = &\left(\alpha_i^{se}\boldsymbol{p_i^{se}(l)} - \beta_i^{no}\boldsymbol{p_i^{no}(l)} - \beta_i^{ca}\boldsymbol{p_i^{ca}(l)} - \beta_i^{re}\boldsymbol{p_i^{re}(l)}\right)TN_i + \\
&\boldsymbol{p_i^{re}(l)}\left(\frac{TN_i}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)\sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k) - \beta_i^{no}p_i^{no}(k) - \beta_i^{ca}p_i^{ca}(k) - \beta_i^{re}p_i^{re}(k)\right)e_{ik}
\end{aligned} \tag{4.36}
$$

In order to obtain structural properties we make the following assumption regarding the properties of the behavioral functions, $p_i^{se}(l)$, $p_i^{no}(l)$, $p_i^{ca}(l)$, $p_i^{re}(l)$, and $p_i^{nc}(l)$, which depend on the appointment lag l, where $p_i^{nc}(l) = 1 - p_i^{re}(l) - p_i^{ca}(l) = p_i^{se}(l) + p_i^{no}(l)$.

Assumption 4.2: The behavioral functions are assumed to have the following properties:

(a) All behavioral functions are differentiable functions within the range $[1, L_i]$ and defined as follows for $l = 0$: $p_i^{se}(0) = 1$, $p_i^{no}(0) = 0$, $p_i^{ca}(0) = 0$, $p_i^{re}(0) = 0$, $p_i^{nc}(0) = 1$

(b) For $l \geq 1$, $p_i^{se}(l)$ and $p_i^{nc}(l)$ are decreasing in l, whereas $p_i^{ca}(l)$, $p_i^{re}(l)$, and $p_i^{no}(l)$ are increasing in l. Here we also assume that these functions take values within the range of (0,1) $\forall i \in \{1, \dots, M\}, l \in [1, L_i]$

(c) The "expected net revenue per patient of type i given that the appointment lag is l", i.e., $\left(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l)\right)$, is a nonnegative function of l.

(d) The revenues and policies have following order: $\alpha_i^{se} > \beta_i^{no} > \beta_i^{ca} > \beta_i^{re}$.

Because $-p_i^{ca}(l)$ is a decreasing function while $p_i^{re}(l)$ is an increasing function in l, we cannot directly conclude whether w_{il} is increasing or decreasing.

Proposition 4.2: w_{il} is a decreasing function in $l \in [1, L_i]$, $\forall i \in \{1, \dots, M\}$.

Proof: In order to prove that w_{il} is a decreasing function, we need to show that it satisfies $\frac{dw_{il}}{dl} < 0$, where $\frac{dw_{il}}{dl} = -\frac{dp_i^{ca}(l)}{dl} s_i N_i + \frac{dp_i^{re}(l)}{dl} s_i N_i \left(\left(\frac{-\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k)}{1 - \sum_{m=0}^{L_i} e_{im} p_i^{re}(m)} \right) \right)$ (4.37)

First, $\frac{dp_i^{ca}(l)}{dl} > 0$ since $p_i^{ca}(l)$ is an increasing function in l, and $-\frac{dp_i^{ca}(l)}{dl} s_i N_i < 0$ because $-s_i N_i < 0$, so the first term in (4.37) is negative.

Next, $\frac{dp_i^{re}(l)}{dl} s_i N_i > 0$ because $p_i^{re}(l)$ is an increasing function in l and $s_i N_i > 0$, and since $0 < 1 - \sum_{m=0}^{L_i} e_{im} p_i^{re}(m) < 1$ and $-1 < -\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k) < 0$, the fraction in the second term is negative; thus the second term in (4.37) is negative. Finally, since both terms in the summation in (4.37) are negative, $\frac{dw_{il}}{dl} < 0$ which means that w_{il} is a decreasing function in $l \in [1, L_i],\ \forall i \in \{1, \dots, M\}$. ■

Next in order to prove the properties of p_{il}, we first need to show that the expected net revenue per patient of type i for appointment lag l is a decreasing function of l.

Proposition 4.3: The "expected net revenue per patient of type i given appointment lag l", i.e., $\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l)$, is a decreasing function of l.

Proof: We can show that this function is decreasing in l by taking its derivative with respect to l:

$\alpha_i^{se} \frac{dp_i^{se}(l)}{dl} - \beta_i^{no} \frac{dp_i^{no}(l)}{dl} - \beta_i^{ca} \frac{dp_i^{ca}(l)}{dl} - \beta_i^{re} \frac{dp_i^{re}(l)}{dl}$ which we know that $\alpha_i^{se} \frac{dp_i^{se}(l)}{dl} < 0$ because $\alpha_i^{se} \geq 0$ and $\frac{dp_i^{se}(l)}{dl} < 0$ because $p_i^{se}(l)$ is a decreasing function in l. We also know that $-\beta_i^{no} \frac{dp_i^{no}(l)}{dl} < 0,\ -\beta_i^{ca} \frac{dp_i^{ca}(l)}{dl} < 0,\ -\beta_i^{re} \frac{dp_i^{re}(l)}{dl} < 0$ because β_i^{no}, β_i^{ca}, and β_i^{re} are all nonnegative, and $\frac{dp_i^{no}(l)}{dl}$, $\frac{dp_i^{ca}(l)}{dl}$, $\frac{dp_i^{re}(l)}{dl}$ are all positive since they are each increasing in l. Thus $\alpha_i^{se} \frac{dp_i^{se}(l)}{dl} - \beta_i^{no} \frac{dp_i^{no}(l)}{dl} - \beta_i^{ca} \frac{dp_i^{ca}(l)}{dl} - \beta_i^{re} \frac{dp_i^{re}(l)}{dl} < 0$ which means the expected net revenue per patient of type i given appointment lag l, is a decreasing function of l. ■

Proposition 4.4: p_{il} is a decreasing function in $l \in [1, L_i],\ \forall i \in \{1, \dots, M\}$.

Proof: In order to prove that p_{il} is a decreasing function, we need to show that $\frac{dp_{il}}{dl} < 0$, where $\frac{dp_{il}}{dl} = \left(\alpha_i^{se} \frac{dp_i^{se}(l)}{dl} - \beta_i^{no} \frac{dp_i^{no}(l)}{dl} - \beta_i^{ca} \frac{dp_i^{ca}(l)}{dl} - \beta_i^{re} \frac{dp_i^{re}(l)}{dl}\right) TN_i +$

$$\frac{dp_i^{re}(l)}{dl}\left(\frac{TN_i}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\sum_{k=0}^{L_i}\left(\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k)\right) e_{ik} \tag{4.38}$$

We know that the first term in the summation in Equation (4.38) is negative, from Proposition 4.3 and $TN_i \geq 0$.

For the second term, we know that $\frac{dp_i^{re}(l)}{dl} > 0$ because it is an increasing function in l and we also know that $\frac{TN_i}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)} > 0$ because $TN_i > 0$ and $0 < 1 - \sum_{m=0}^{L_i} e_{im} p_i^{re}(m) < 1$. The summation in the second term of the Equation (4.38) is also positive because by Assumption 4.2c $\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k) \geq 0$ and $0 \leq e_{ik} \leq 1\ \forall i, k$. Because the first term is negative but the second term is positive in Equation (4.38), we cannot directly determine if p_{il} is increasing or decreasing. In other words, we can only specify the conditions for which p_{il} is decreasing, i.e., p_{il} is decreasing if the following holds

$$\left|\left(\alpha_i^{se} \frac{dp_i^{se}(l)}{dl} - \beta_i^{no} \frac{dp_i^{no}(l)}{dl} - \beta_i^{ca} \frac{dp_i^{ca}(l)}{dl} - \beta_i^{re} \frac{dp_i^{re}(l)}{dl}\right) TN_i\right| >$$
$$\frac{dp_i^{re}(l)}{dl}\left(\frac{TN_i}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\sum_{k=0}^{L_i}\left(\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k)\right) e_{ik}$$

which can be simplified by removing TN_i terms from both sides because $TN_i > 0$:

$$\left|\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}-\beta_i^{no}\frac{dp_i^{no}(l)}{dl}-\beta_i^{ca}\frac{dp_i^{ca}(l)}{dl}-\beta_i^{re}\frac{dp_i^{re}(l)}{dl}\right|>$$

$$\frac{dp_i^{re}(l)}{dl}\left(\frac{1}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)\sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k)-\beta_i^{no}p_i^{no}(k)-\beta_i^{ca}p_i^{ca}(k)-\beta_i^{re}p_i^{re}(k)\right)e_{ik}$$

which is equal to

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}+\beta_i^{no}\frac{dp_i^{no}(l)}{dl}+\beta_i^{ca}\frac{dp_i^{ca}(l)}{dl}+\beta_i^{re}\frac{dp_i^{re}(l)}{dl}>$$

$$\frac{dp_i^{re}(l)}{dl}\frac{\sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k)-\beta_i^{no}p_i^{no}(k)-\beta_i^{ca}p_i^{ca}(k)-\beta_i^{re}p_i^{re}(k)\right)e_{ik}}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}$$

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}+\beta_i^{no}\frac{dp_i^{no}(l)}{dl}+\beta_i^{ca}\frac{dp_i^{ca}(l)}{dl}+\beta_i^{re}\frac{dp_i^{re}(l)}{dl}>$$

$$\frac{dp_i^{re}(l)}{dl}\left(\frac{\sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k)-\beta_i^{no}p_i^{no}(k)-\beta_i^{ca}p_i^{ca}(k)\right)e_{ik}}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}-\beta_i^{re}\frac{\sum_{k=0}^{L_i}e_{ik}p_i^{re}(k)}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)$$

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}+\beta_i^{no}\frac{dp_i^{no}(l)}{dl}+\beta_i^{ca}\frac{dp_i^{ca}(l)}{dl}+\beta_i^{re}\frac{dp_i^{re}(l)}{dl}+\beta_i^{re}\frac{dp_i^{re}(l)}{dl}\frac{\sum_{k=0}^{L_i}e_{ik}p_i^{re}(k)}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}>$$

$$\frac{dp_i^{re}(l)}{dl}\left(\frac{\sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k)-\beta_i^{no}p_i^{no}(k)-\beta_i^{ca}p_i^{ca}(k)\right)e_{ik}}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)$$

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}+\beta_i^{no}\frac{dp_i^{no}(l)}{dl}+\beta_i^{ca}\frac{dp_i^{ca}(l)}{dl}+\beta_i^{re}\frac{dp_i^{re}(l)}{dl}\frac{1}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}>$$

$$\frac{dp_i^{re}(l)}{dl}\left(\frac{\sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k)-\beta_i^{no}p_i^{no}(k)-\beta_i^{ca}p_i^{ca}(k)\right)e_{ik}}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)$$ from which we can remove the $\left(1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)\right)$ terms from the denominators because these terms are positive, and by reorganizing the terms we obtain the following expression:

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}\left(1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)\right)+\beta_i^{no}\frac{dp_i^{no}(l)}{dl}\left(1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)\right)+\beta_i^{ca}\frac{dp_i^{ca}(l)}{dl}\left(1-\right.$$

$$\left.\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)\right)+\beta_i^{re}\frac{dp_i^{re}(l)}{dl}-\alpha_i^{se}\frac{dp_i^{re}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{se}(m)+$$

$$\beta_i^{no}\frac{dp_i^{re}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{no}(m)+\beta_i^{ca}\frac{dp_i^{re}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{ca}(m)>0 \qquad (4.39)$$

For expression (4.39) we know the following always holds based on the assumptions for the behavioral functions:

$$\left(\beta_i^{no}\frac{dp_i^{no}(l)}{dl}\left(1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)\right)+\beta_i^{ca}\frac{dp_i^{ca}(l)}{dl}\left(1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)\right)+\beta_i^{re}\frac{dp_i^{re}(l)}{dl}+\right.$$

$$\left.\beta_i^{no}\frac{dp_i^{re}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{no}(m)+\beta_i^{ca}\frac{dp_i^{re}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{ca}(m)\right)>0.$$

Thus it would be sufficient to show that the remaining term is also nonnegative, in other words we claim that the following relationship shown by Lemma 4.1 also holds:

Lemma 4.1: The following relationship holds:

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}+\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)>\alpha_i^{se}\frac{dp_i^{re}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{se}(m) \qquad (4.40)$$

In order to prove this lemma we start with the relationship between the behavioral functions:

$$p_i^{se}(m)=1-p_i^{re}(m)-p_i^{ca}(m)-p_i^{no}(m)$$

$\frac{dp_i^{se}(l)}{dl}=-\frac{dp_i^{re}(l)}{dl}-\frac{dp_i^{ca}(l)}{dl}-\frac{dp_i^{no}(l)}{dl}$ which is equal to $-\frac{dp_i^{se}(l)}{dl}=\frac{dp_i^{re}(l)}{dl}+\frac{dp_i^{ca}(l)}{dl}+\frac{dp_i^{no}(l)}{dl}$.

By Assumption 4.2b, we know that $\frac{dp_i^{se}(l)}{dl}<0$, and $\frac{dp_i^{re}(l)}{dl}$, $\frac{dp_i^{ca}(l)}{dl}$, $\frac{dp_i^{no}(l)}{dl}>0$.

Thus we can conclude that $-\frac{dp_i^{se}(l)}{dl}>\frac{dp_i^{re}(l)}{dl}$.

Then because $-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{se}(m)>\alpha_i^{se}\frac{dp_i^{re}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{se}(m)$ if we can show that $-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}+\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)>-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}\sum_{m=0}^{L_i}e_{im}p_i^{se}(m)$

holds, we can conclude that Equation (4.40) holds and prove Lemma 4.1. Moving all terms to the left hand side yields:

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}+\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}\sum_{m=0}^{L_i}e_{im}\left(p_i^{re}(m)+p_i^{se}(m)\right)=$$

$$-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}\left(1-\sum_{m=0}^{L_i}e_{im}\left(p_i^{re}(m)+p_i^{se}(m)\right)\right)$$

which is nonnegative because $-\alpha_i^{se}\frac{dp_i^{se}(l)}{dl}>0$ and since $p_i^{re}(m)+p_i^{se}(m)\leq 1$, and $\sum_{m=0}^{L_i}e_{im}=1$, then $\sum_{m=0}^{L_i}e_{im}\left(p_i^{re}(m)+p_i^{se}(m)\right)\leq 1$, and $1-\sum_{m=0}^{L_i}e_{im}\left(p_i^{re}(m)+p_i^{se}(m)\right)\geq 0$.

Thus, we have proven Lemma 4.1, and therefore Equation (4.39) holds. This proves Proposition 4.4, that p_{il} are decreasing in l. ■

Proposition 4.5: a) w_{il} and p_{il} are decreasing differentiable functions in $l\in[1,L_i],\ \forall i\in\{1,\dots,M\}$. **b)** If the behavior functions are linear then w_{il} and p_{il} are decreasing linear functions in $l\in[1,L_i],\ \forall i\in\{1,\dots,M\}$.

Proof: For part a, by Proposition 4.2 and 4.4 we know that w_{il} and p_{il} are decreasing functions in $l\in[1,L_i],\ \forall i\in\{1,\dots,M\}$. By Assumption 4.2c each behavioral function is differentiable where $p_i^{se}(0)=1,\ p_i^{no}(0)=0,\ p_i^{ca}(0)=0,\ p_i^{re}(0)=0,$ and $p_i^{nc}(0)=1$. Thus for $l\in[1,L_i]$ w_{il} and p_{il} are decreasing differentiable functions. In part b, we assume linearity of the behavior functions and because adding or subtracting linear functions yield a linear function, we can conclude that in this case also w_{il} and p_{il} are decreasing linear functions. ■

Theorem 4.1: If the behavior functions are linear, then there exists an optimal solution in which $a_{il} = 0$ for $\forall i \in \{1, \dots, M\}\ l \in [2, L_i - 1]$.

Proof: Proposition 4.1a states that for $w_{ik} \leq w_{il} \leq w_{im}$ and $p_{ik} \leq p_{il} \leq p_{im}$ if $\frac{p_{il}-p_{ik}}{w_{il}-w_{ik}} \leq \frac{p_{im}-p_{il}}{w_{im}-w_{il}}$ is satisfied, then option l is LP-dominated by options k and m. Then, based on Proposition 4.1b, $a_{il} = 0$ is an optimal solution.

Using Proposition 4.5 we know that the following orderings $w_{ik} \leq w_{il} \leq w_{im}$ and $p_{ik} \leq p_{il} \leq p_{im}$ are satisfied. Assuming linearity of the behavior functions within the range $[1, L_i]$, the inequality $\frac{p_{il}-p_{ik}}{w_{il}-w_{ik}} \leq \frac{p_{im}-p_{il}}{w_{im}-w_{il}}$ is also satisfied with equality $\forall m \leq l \leq k \in [1, L_i]$. Thus, there exists an optimal solution in which $a_{il} = 0\ \forall i \in \{1, \dots, M\}$ and $l \in [2, L_i - 1]$. ■

If the behavior functions are nonlinear but differentiable, the dominated options can be identified using the LP-dominance definition to reduce the number of options for the appointment lag for a given patient type.

Next we introduce a small sample problem to illustrate Theorem 4.1.

Example: For a two patient setting with $T = 5,\ L = [4,4],\ N = [25,25],\ K = 48,\ s = [1,1],\ \alpha^{se} = [1200, 1000],\ \beta^{no} = [500,500],\ \beta^{ca} = [100,100],\ \beta^{re} = [25,25],\ e_i = [0.2,0.2,0.2,0.2,0.2]\ for\ i = 1,2$

$p_i^{no}(l) = 0.0013l + 0.0308\ for\ l \in [1, L_i], i = 1,2, p^{no}(0) = [0,0]$

$p_i^{ca}(l) = 0.0192l - 0.0033\ for\ l \in [1, L_i], i = 1,2, p^{ca}(0) = [0,0]$

$p_i^{re}(l) = 0.0195l - 0.0191\ for\ l \in [1, L_i], i = 1,2, p^{re}(0) = [0,0]$

$p_i^{se}(l) = -0.04l + 0.9916\ for\ l \in [1, L_i], i = 1,2, p^{se}(0) = [0,0]$

Solving Model 4.1 with ILOG/OPL the optimal solution is:

$a = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0.1937 & 0 & 0 & 0.8063 \end{bmatrix}$ with an objective value of $z = 249{,}767.75$.

Using Theorem 4.1 the size of the problem can be reduced significantly, because we can eliminate most of the dominated variables from the problem. By setting $a_{il} = 0\ \forall i \in \{1, \dots, M\},\ l \in [2, L_i - 1],\ a_{il}\ \forall i \in \{1, \dots, M\}$ is evaluated only for $l \in \{1, L_i\}$. After eliminating the dominated variables, a modified version of the MCKP-Greedy Algorithm (Kellerer, Pferschy, and Pisinger, 2004) can be used to find an optimal solution to Model 4.1′ without a solver. The LMCKP-Greedy Algorithm adapted for Model 4.1′ is shown below.

LMCKP-Greedy Algorithm for Model 4.1′:

1- Obtain $\tilde{p}_{i1} = p_{i1} - p_{iL_i}$ and $\tilde{w}_{i1} = w_{i1} - w_{iL_i}$ for each $i \in \{1, \dots, M\}$

2- Calculate the residual capacity $\bar{c} = K - \sum_{i=1}^{M} w_{iL_i}$

3- Sort the patient types in decreasing incremental efficiency order, where the incremental efficiencies are obtained by $\tilde{e}_{i1} = \frac{\tilde{p}_{i1}}{\tilde{w}_{i1}}$ for $i \in \{1, \dots, M\}$

4- Use a Greedy Algorithm to fill the remaining capacity of the knapsack ($\bar{c}$)

   a. Initialize $z = \sum_{i=1}^{M} p_{iL_i}$ and set $i = argmax_{i \in \{1,\dots,M\}}\{\tilde{e}_{i1}\}$

   b. If $\bar{c} \geq \tilde{w}_{i1}$, set $a_{i1} = 1$, $a_{iL_i} = 0$, $z = z + \tilde{p}_{i1}$ and $\bar{c} = \bar{c} - \tilde{w}_{i1}$, then go to 4c;

      Else, set $a_{i1} = \bar{c}/\tilde{w}_{i1}$, $a_{iL_i} = 1 - \bar{c}/\tilde{w}_{i1}$, and $z = z + \tilde{p}_{i1} a_{i1}$, then go to Step 5

   c. Add option i to the set $\bar{I}$.

      If $\bar{I} \neq \{1, \dots, M\}$, set $i = argmax_{i \in \{1,\dots,M\} \setminus \bar{I}}\{\tilde{e}_{i1}\}$, then go to Step 4b;

      Else go to Step 5

5- Return the LP solution a with an objective function value of z.

In order to illustrate this algorithm we apply it to the sample problem introduced. Using MATLAB we obtain the following w and p values:

$$p = \begin{bmatrix} 140{,}588 & 136{,}898 & 133{,}208 & 129{,}517 \\ 116{,}789 & 113{,}643 & 110{,}496 & 107{,}350 \end{bmatrix}$$

$$w = \begin{bmatrix} 24.6021 & 24.1043 & 23.6064 & 23.1086 \\ 24.6021 & 24.1043 & 23.6064 & 23.1086 \end{bmatrix}$$

Then again using MATLAB (without solvers), we follow the steps of the MCKP Greedy Algorithm for Model 4.1′:

1- We obtain $\tilde{p}_{11} = 11{,}071,\ \tilde{p}_{21} = 9{,}439, \tilde{w}_{11} = 1.4936, \tilde{w}_{21} = 1.4936$

2- Then we calculate $\bar{c} = 1.7829$

3- Incremental efficiencies are obtained as $\tilde{e}_{11} = 7{,}412.48,\ \tilde{e}_{21} = 6{,}319.88$ which are already in decreasing order

4-

   a. We first initialize $z = 236{,}868.33,$ and set $i = 1$

   b. Since $\bar{c} = 1.7829 \geq 1.4936 = \tilde{w}_{11},$ we set $a_{11} = 1, a_{14} = 0, z = 247{,}939.33$ and updated $\bar{c} = 0.2893$

   c. We add $i = 1$ to the set $\bar{I}$ and since $\bar{I} \neq \{1,2\}$ we set $i = 2$ and go back to 4b:

   4b. (for $i = 2$) Since $\bar{c} = 0.2893 < 1.4936 = \tilde{w}_{21},$ we set $a_{21} = \bar{c}/\tilde{w}_{21} = 0.1937, a_{24} = 1 - \bar{c}/\tilde{w}_{21} = 0.8063,\ z = 249{,}767.75.$

5- Thus, $a = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0.1937 & 0 & 0 & 0.8063 \end{bmatrix}$ is the optimal solution to this problem with an objective value $z = 249{,}767.75$ using the LMCKP Greedy Algorithm.

Note that if the setting has nonlinear but differentiable behavior functions, we can still use a modification of this algorithm in which more options are evaluated (although this

number can be reduced by using LP-dominance). The Greedy Algorithm can still be used to solve the problem as there are a finite number of days.

4.3 Model II

In this section we build a more comprehensive model where we no longer assume that the reschedule requests are known on the day that the appointment is assigned, instead the reschedule requests can arrive any time before the actual appointment day. For example, in the clinic that motivates this research, the patients can call the clinic a few days before their initial appointment, and request their appointment be rescheduled to another day. In this case, the timing of that reschedule request call is important, as the new appointment range would start from that day until the latest allowed day (considering the patient's allowances).

For this purpose, we introduce a reschedule request function $p_i^{rr}(r|l)$ to represent the probability that a reschedule request occurs r days prior to the appointment day given that the appointment will be rescheduled. We now have three parameters to represent reschedules: $p_i^{re}(l)$ which is the probability that type i patient will reschedule her appointment given that her appointment delay is l; $p_i^{rr}(r|l)$ which is the probability that a type i patient with an appointment calls r days prior to her appointment day given that her initial appointment delay was l days; and e_{il} which is the probability that a type i patient wants an appointment l days later than the time of the reschedule request.

Model 4.2:

$$\max \sum_{i=1}^{M} \sum_{t=1}^{T} \Big(\sum_{l=0}^{L_i} \big(R_{i,t-l} e_{il} + N_{i,t-l} a_{il} \big) \big(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l) \big) \Big) \quad (4.41)$$

$$s.t.$$

$$\sum_{l=0}^{L_i} \sum_{r=0}^{l} (R_{i,t+r-l} e_{il} + N_{i,t+r-l} a_{il}) p_i^{re}(l) p_i^{rr}(r|l) - R_{it} = 0 \quad \forall i = 1, \dots, M, t = 1, \dots, T \quad (4.42)$$

$$\sum_{i=1}^{M} \sum_{l=0}^{L_i} p_i^{nc}(l) \big(R_{i,t-l} e_{il} + N_{i,t-l} a_{il} \big) s_i - K_t \leq 0 \quad \forall t = 1, \dots, T \quad (4.43)$$

$$R_{i,t} + N_{i,t} = R_{i,t+T} + N_{i,t+T} \quad \forall i = 1, \dots, 4, t = 1 - L_i, \dots, 0 \quad (4.44)$$

$$\sum_{l=1}^{L_i} a_{il} = 1 \quad \forall i = 1, \dots, M \quad (4.45)$$

$$a_{i0} = 0 \quad \forall i = 1, \dots, M \quad (4.46)$$

$$1 \geq a_{il} \geq 0 \quad \forall i = 1, \dots, 4, l = 0, \dots, L_i \quad (4.47)$$

We modify the first constraint in Model 4.1 to get constraint (4.42) so that the reschedule requests can occur on any day before the appointment day. Model 4.1 is a special case of Model 4.2 with $r = l$. Figure 4.3 illustrates how constraint (4.42) works for $i = 1, t = 4$. Note that the arrows illustrate a few terms of the left hand side of constraint (4.42) for visualization purposes. For example, $(t = 4, l = 3, r = 1)$ is shown with the orange arrows where the forward arrow represents the sum of "the appointments that are rescheduled" and "the new appointments that are scheduled" on day 2 for an appointment three days later and the backwards arrow represents the proportion of those appointments that are requested to be rescheduled on day 4 (one day prior to their previously assigned appointments).

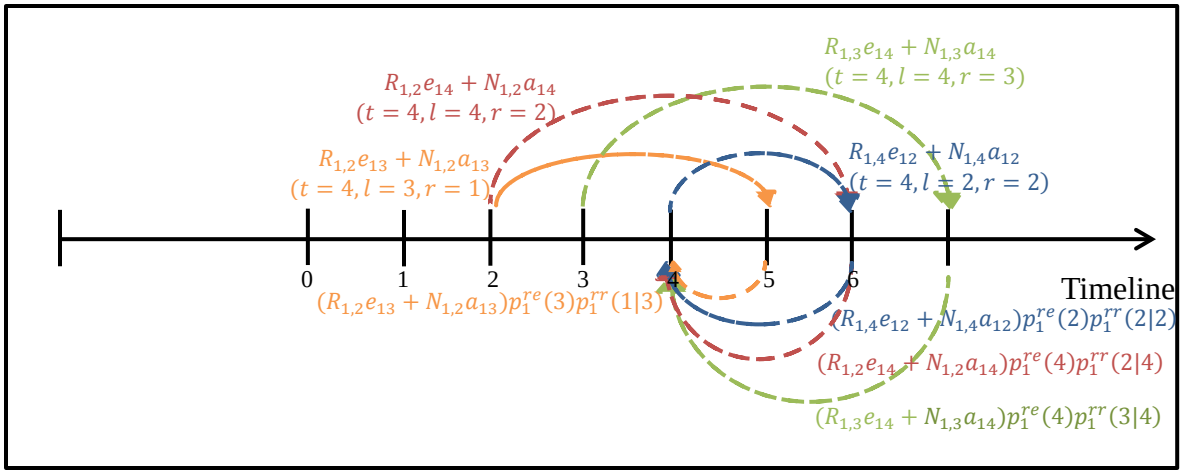

Figure 4.3: Illustration of the constraint (4.42) for $i = 1, t = 4$ *(Each color represents a possible movement of appointment requests from a day - Orange and Red from Day 2, Green from Day 3 and Blue from Day 4 - to Day 4 due to reschedule requests arriving certain number of days (1 for Orange, 2 for Red, 3 for Green, and 2 for Blue) prior to the appointment. The summation of all arrows coming into $t = 4$ will be equal to $R_{1,4}$)*

Constraint (4.44) defines $R_{i,t}$ values for the t indices that are outside of the $[1, T]$ bounds. Note in Model 4.1 we removed this constraint because the cycles of length T are identical. For the comprehensive model (Model 4.2), it is not easy to show that the cycles are identical. In order to allow patients to reschedule their appointments to an earlier or a later date any time before their appointment we created indices to keep track of this movement, making the model significantly more complicated. Due to the complex nature of this model, we assume that there are identical cycles of length T. From constraint (4.44) we have $R_{i,t} + N_{i,t} = R_{i,t+T} + N_{i,t+T} \quad \forall i = 1, \dots, M, t = 1 - L_i, \dots, 0$. Under the stationary demand assumption $N_{i,t} = N_i \forall t$ and for this case there are identical cycles. When we write Equation (4.42) explicitly we can show that there are T equations with T variables for each patient class i.

Because Equation (4.42) is more complicated than the corresponding equation (4.2) in Model 4.1, we cannot directly derive $R_{i,t}$. For this reason we take a different approach to represent $R_{i,t}$ in terms of other parameters and variables. We first examine this equation for small cases. First assume that $T = 2$ and $L_i = 2$ for a given patient class, then the following two constraints are generated by Equation (4.42):

$$\sum_{l=0}^{L_i} \sum_{r=0}^{l} (R_{i,1+r-l} e_{il} + N_i a_{il}) p_i^{re}(l) p_i^{rr}(r|l) - R_{i,1} = 0$$

$$\sum_{l=0}^{L_i} \sum_{r=0}^{l} (R_{i,2+r-l} e_{il} + N_i a_{il}) p_i^{re}(l) p_i^{rr}(r|l) - R_{i,2} = 0$$

We can write these equations explicitly using matrix form as follows:

$$\begin{pmatrix} \begin{bmatrix} e_{i0}p_i^{re}(0)p_i^{rr}(0|0) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(0|2) \\ +e_{i1}p_i^{re}(1)p_i^{rr}(1|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(2|2) - 1 \end{bmatrix} & \begin{bmatrix} e_{i1}p_i^{re}(1)p_i^{rr}(0|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} \\ \begin{bmatrix} e_{i1}p_i^{re}(1)p_i^{rr}(0|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} & \begin{bmatrix} e_{i0}p_i^{re}(0)p_i^{rr}(0|0) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(0|2) \\ +e_{i1}p_i^{re}(1)p_i^{rr}(1|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(2|2) - 1 \end{bmatrix} \end{pmatrix} \begin{pmatrix} R_{i1} \\ R_{i2} \end{pmatrix} = -\begin{pmatrix} \begin{bmatrix} a_{i0}N_i p_i^{re}(0)p_i^{rr}(0|0) \\ +a_{i2}N_i p_i^{re}(2)p_i^{rr}(0|2) \\ +a_{i1}N_i p_i^{re}(1)p_i^{rr}(1|1) \\ +a_{i2}N_i p_i^{re}(2)p_i^{rr}(2|2) \end{bmatrix} + \begin{bmatrix} a_{i1}N_i p_i^{re}(1)p_i^{rr}(0|1) \\ +a_{i2}N_i p_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} \\ \begin{bmatrix} a_{i0}N_i p_i^{re}(0)p_i^{rr}(0|0) \\ +a_{i2}N_i p_i^{re}(2)p_i^{rr}(0|2) \\ +a_{i1}N_i p_i^{re}(1)p_i^{rr}(1|1) \\ +a_{i2}N_i p_i^{re}(2)p_i^{rr}(2|2) \end{bmatrix} + \begin{bmatrix} a_{i1}N_i p_i^{re}(1)p_i^{rr}(0|1) \\ +a_{i2}N_i p_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} \end{pmatrix}$$

Note that we have a matrix operation with the following structure:

$\begin{bmatrix} a_1 & a_2 \\ a_2 & a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c \\ c \end{bmatrix}$ which means our equations look like:

$a_1 x_1 + a_2 x_2 = c$ and $a_2 x_1 + a_1 x_2 = c$. If we solve these the solutions for x_1 and x_2 are:

$x_1 = \frac{c}{a_1 + a_2}$ and $x_2 = \frac{c}{a_1 + a_2}$.

For $T = 3$ and $L_i = 2$ for a given patient class, then our equations will be as follows in matrix form:

$$\begin{pmatrix} \begin{bmatrix} e_{i0}p_i^{re}(0)p_i^{rr}(0|0) \\ +e_{i1}p_i^{re}(1)p_i^{rr}(1|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(2|2)-1 \end{bmatrix} & e_{i2}p_i^{re}(2)p_i^{rr}(0|2) & \begin{bmatrix} e_{i1}p_i^{re}(1)p_i^{rr}(0|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} \\ e_{i2}p_i^{re}(2)p_i^{rr}(0|2) & \begin{bmatrix} e_{i1}p_i^{re}(1)p_i^{rr}(0|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} & \begin{bmatrix} e_{i0}p_i^{re}(0)p_i^{rr}(0|0) \\ +e_{i1}p_i^{re}(1)p_i^{rr}(1|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(2|2)-1 \end{bmatrix} \\ \begin{bmatrix} e_{i1}p_i^{re}(1)p_i^{rr}(0|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} & \begin{matrix} e_{i0}p_i^{re}(0)p_i^{rr}(0|0) \\ +e_{i1}p_i^{re}(1)p_i^{rr}(1|1) \\ +e_{i2}p_i^{re}(2)p_i^{rr}(2|2)-1 \end{matrix} & e_{i2}p_i^{re}(2)p_i^{rr}(0|2) \end{pmatrix} \begin{matrix} R_{i1} \\ R_{i2} \\ R_{i3} \end{matrix} = -\begin{pmatrix} \begin{bmatrix} a_{i0}N_ip_i^{re}(0)p_i^{rr}(0|0) \\ +a_{i1}N_ip_i^{re}(1)p_i^{rr}(1|1) \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(2|2) \end{bmatrix} \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(0|2) \\ +\begin{bmatrix} a_{i1}N_ip_i^{re}(1)p_i^{rr}(0|1) \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} \\ \begin{bmatrix} a_{i0}N_ip_i^{re}(0)p_i^{rr}(0|0) \\ +a_{i1}N_ip_i^{re}(1)p_i^{rr}(1|1) \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(2|2) \end{bmatrix} \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(0|2) \\ +\begin{bmatrix} a_{i1}N_ip_i^{re}(1)p_i^{rr}(0|1) \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} \\ \begin{bmatrix} a_{i0}N_ip_i^{re}(0)p_i^{rr}(0|0) \\ +a_{i1}N_ip_i^{re}(1)p_i^{rr}(1|1) \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(2|2) \end{bmatrix} \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(0|2) \\ +\begin{bmatrix} a_{i1}N_ip_i^{re}(1)p_i^{rr}(0|1) \\ +a_{i2}N_ip_i^{re}(2)p_i^{rr}(1|2) \end{bmatrix} \end{pmatrix}$$

Here the matrix on the left handside is also a symmetric matrix and has the following form where c is nonzero:

$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_2 & a_3 & a_1 \\ a_3 & a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} c \\ c \\ c \end{bmatrix}$ which means our equations look like:

$a_1x_1 + a_2x_2 + a_3x_3 = c$

$a_2x_1 + a_3x_2 + a_1x_3 = c$

$a_3x_1 + a_1x_2 + a_2x_3 = c$, and the solution to these linear equations is:

$x_1 = \frac{c}{a_1+a_2+a_3}$, $x_2 = \frac{c}{a_1+a_2+a_3}$ and $x_3 = \frac{c}{a_1+a_2+a_3}$.

Proposition 4.6: For a set of linear equations if we have a symmetric TxT A matrix (which means $A = A^T$) and a C vector which has the same element c on each row, the unique solution to the linear equations will be in the form of $x_i = \frac{c}{\sum_{j=1}^{T} a_j}$ $\forall i$.

Proof: We can show that this argument is true for any $T = k$ size problems as follows:

For $T = k$ we have following set of equations:

$$a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_kx_k = c$$

$$a_2x_1 + a_3x_2 + a_4x_3 + \cdots + a_kx_{k-1} + a_1x_k = c$$

$$a_3x_1 + a_4x_2 + a_5x_3 + \cdots + a_kx_{k-2} + a_1x_{k-1} + a_2x_k = c$$

$$\vdots$$

$$a_kx_1 + a_1x_2 + a_2x_3 + \cdots + a_{k-2}x_{k-1} + a_{k-1}x_k = c$$

In each of the above equations if we plug in $x_i = \frac{c}{a_1+a_2+a_3+\cdots+a_k}\ \forall i = 1, \dots, k$, we can show that each equation is satisfied with equality, i.e.

$$a_1\frac{c}{a_1+a_2+a_3+\cdots+a_k} + a_2\frac{c}{a_1+a_2+a_3+\cdots+a_k} + a_3\frac{c}{a_1+a_2+a_3+\cdots+a_k} + \cdots + a_k\frac{c}{a_1+a_2+a_3+\cdots+a_k}$$

$$= \frac{c}{a_1+a_2+a_3+\cdots+a_k}(a_1 + a_2 + a_3 + \cdots + a_k) = c$$

$$\vdots$$

$$a_k\frac{c}{a_1+a_2+a_3+\cdots+a_k} + a_1\frac{c}{a_1+a_2+a_3+\cdots+a_k} + a_2\frac{c}{a_1+a_2+a_3+\cdots+a_k} + \cdots + a_{k-2}\frac{c}{a_1+a_2+a_3+\cdots+a_k} +$$

$$a_{k-1}\frac{c}{a_1+a_2+a_3+\cdots+a_k} = \frac{c}{a_1+a_2+a_3+\cdots+a_k}(a_k + a_1 + a_2 + \cdots + a_{k-1}) = c.$$

Moreover, $x_i = \frac{c}{a_1+a_2+a_3+\cdots+a_k}\ \forall i = 1, \dots, k$ is the unique solution for this set of linear equations due to the fact that each row in the matrix notation is independent. We can prove this by showing that $\sum_{j=1,\dots,k} \lambda_j A_j = 0$ is satisfied only if all λ_j are set to zero.

$$\lambda_1\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{k-1} \\ a_k \end{bmatrix} + \lambda_2\begin{bmatrix} a_2 \\ a_3 \\ \vdots \\ a_k \\ a_1 \end{bmatrix} + \cdots + \lambda_k\begin{bmatrix} a_k \\ a_1 \\ \vdots \\ a_{k-2} \\ a_{k-1} \end{bmatrix} = 0$$

$$a_1\lambda_1 + a_2\lambda_2 + a_3\lambda_3 + \cdots + a_k\lambda_k = 0$$

$$a_2\lambda_1 + a_3\lambda_2 + a_4\lambda_3 + \cdots + a_k\lambda_{k-1} + a_1\lambda_k = 0$$

$$a_3\lambda_1 + a_4\lambda_2 + a_5\lambda_3 + \cdots + a_k\lambda_{k-2} + a_1\lambda_{k-1} + a_2\lambda_k = 0$$

$$\vdots$$

$$a_k\lambda_1 + a_1\lambda_2 + a_2\lambda_3 + \cdots + a_{k-2}\lambda_{k-1} + a_{k-1}\lambda_k = 0$$

Solving this set of equation yields $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0$ which suggests that each row (equation) is independent of the other rows, which means we have a full rank matrix and therefore the solution to this set of equations is the unique solution. ■

By using these findings we can write the $R_{i,t}$ in terms of the other parameters and variables as:

$$R_{i,t} = \frac{\sum_{l=0}^{L_i}\sum_{r=0}^{l} N_i a_{il} p_i^{re}(l) p_i^{rr}(r|l)}{1-\sum_{l=0}^{L_i}\sum_{r=0}^{l} e_{il} p_i^{re}(l) p_i^{rr}(r|l)} \tag{4.48}$$

We can drop the index t as the values of R_{it} are equal for all t, i.e., they do not depend on day t, under the assumption of stationary demand. Also under the equal daily capacity assumption the index t can be removed from the daily capacity parameter K_t. Thus we can plug in the right hand side of the equation (4.48) for R_i in the Model 4.2. By making the above changes and rearranging to simplify Model 4.2, we obtain the following model:

$$\max T \sum_{i=1}^{M} \left(\sum_{l=0}^{L_i} \left(\frac{\sum_{k=0}^{L_i} \sum_{r=0}^{k} N_i a_{ik} p_i^{re}(k) p_i^{rr}(r|k)}{1-\sum_{k=0}^{L_i} \sum_{r=0}^{k} e_{ik} p_i^{re}(k) p_i^{rr}(r|k)} e_{il} + N_i a_{il} \right) \left(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \right. \right.$$

$$\left. \left. \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l) \right) \right) \tag{4.49}$$

$s.t.$

$$\sum_{i=1}^{M} \sum_{l=0}^{L_i} p_i^{nc}(l) \left(\frac{\sum_{k=0}^{L_i} \sum_{r=0}^{k} N_i a_{ik} p_i^{re}(k) p_i^{rr}(r|k)}{1-\sum_{k=0}^{L_i} \sum_{r=0}^{k} e_{ik} p_i^{re}(k) p_i^{rr}(r|k)} e_{il} + N_i a_{il} \right) s_i \leq K \tag{4.50}$$

$$\sum_{l=1}^{L_i} a_{il} = 1 \qquad \forall i = 1, \dots, M \tag{4.51}$$

$$a_{i0} = 0 \qquad \forall i = 1, \dots, M \tag{4.52}$$

$$1 \geq a_{il} \geq 0 \qquad \forall i = 1, \dots, 4, l = 0, \dots, L_i \tag{4.53}$$

We can simplify this model even further because $\sum_{r=0}^{k} p_i^{rr}(r|k) = 1$ in both the objective function and the first constraint. Thus the last version of Model 4.2 can be written as follows:

Model 4.2':

$$\max T \sum_{i=1}^{M} \left(\sum_{l=0}^{L_i} \left(\frac{\sum_{k=0}^{L_i} N_i a_{ik} p_i^{re}(k)}{1-\sum_{k=0}^{L_i} e_{ik} p_i^{re}(k)} e_{il} + N_i a_{il} \right) \left(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \right. \right.$$

$$\left. \left. \beta_i^{re} p_i^{re}(l) \right) \right)$$

$s.t.$

$$\sum_{i=1}^{M} \sum_{l=0}^{L_i} p_i^{nc}(l) \left(\frac{\sum_{k=0}^{L_i} N_i a_{ik} p_i^{re}(k)}{1-\sum_{k=0}^{L_i} e_{ik} p_i^{re}(k)} e_{il} + N_i a_{il} \right) s_i \leq K \tag{4.54}$$

$$\sum_{l=1}^{L_i} a_{il} = 1 \qquad \forall i = 1, \dots, M \tag{4.55}$$

$$a_{i0} = 0 \qquad \forall i = 1, \dots, M \tag{4.56}$$

$$1 \geq a_{il} \geq 0 \qquad \forall i = 1, \dots, 4, l = 0, \dots, L_i \tag{4.57}$$

Observe that Model 4.2 reduces to Model 4.1′, the multiple choice knapsack problem, where the variables a_{il} have the coefficients w_{il} and p_{il} for the capacity constraint and the objective function. The same w_{il} and p_{il} coefficients from Model 4.1′ can be used here and are given by equations (4.54) and (4.55) below where the terms that depend on l are shown bold.

$$w_{il} = s_i N_i - \boldsymbol{p_i^{ca}(l)} s_i N_i + \boldsymbol{p_i^{re}(l)} s_i N_i \left(\frac{-\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k)}{\left(1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)\right)} \right) \tag{4.58}$$

$$p_{il} = \left(\alpha_i^{se}\boldsymbol{p_i^{se}(l)} - \beta_i^{no}\boldsymbol{p_i^{no}(l)} - \beta_i^{ca}\boldsymbol{p_i^{ca}(l)} - \beta_i^{re}\boldsymbol{p_i^{re}(l)}\right) TN_i + \boldsymbol{p_i^{re}(l)} \left(\frac{TN_i}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)} \right) \sum_{k=0}^{L_i} \left(\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k)\right) e_{ik} \tag{4.59}$$

Since w_{il} and p_{il} are the same as for Model 4.1′, the model properties for Model 4.1′ also hold for Model 4.2′. This actually makes sense because our aim with these models is to find the optimal capacity allocation policy on each day by finding the optimal daily appointment assignment policy based on the expected revenue and expected number of unchanged appointments on a given day, assuming stationary demand on every day. Thus as far as this model is concerned, the earlier or the later the appointment reschedule request arrives should not have an effect on the result. This is primarily because we assume stationary demand and use the same decision variables a_{il} on each day. To obtain Model 4.2 we introduced the concept of reschedule request timing $p_i^{rr}(r|l)$. With constraint (4.42) we moved the time interval that we can schedule the rescheduled appointments. In other words, in Model 4.1 we were assigning the rescheduled appointments to another day within the range $[t, t + L_i]$ but

now with the Model 4.2 we are able to assign them to any day within the range $[t + l - r, t + l - r + L_i]$, where t is the initial appointment request date, l is the appointment delay, and r is the number of days prior to the appointment day when the patient made a reschedule request. Moreover since we assume stationary demand and the same appointment assignment policy over different days, this movement of the reschedule range does not affect the solution.

Based on the above discussion we would like to explore the effect of relaxing the stationary demand assumption under the identical cycles of length T assumption. We illustrate the effect of relaxing the stationary demand assumption with a simple example. For $T = 2$ and $L_i = 2$, Equation (4.42) will generate the following two equations:

$$\sum_{l=0}^{L_i} \sum_{r=0}^{l} (R_{i,1+r-l} e_{il} + N_{i,1+r-l} a_{il}) p_i^{re}(l) p_i^{rr}(r|l) - R_{i1} = 0$$

$$\sum_{l=0}^{L_i} \sum_{r=0}^{l} (R_{i,2+r-l} e_{il} + N_{i,2+r-l} a_{il}) p_i^{re}(l) p_i^{rr}(r|l) - R_{i2} = 0$$

Similar to the previous model, we write these equations explicitly in matrix form:

$$\begin{pmatrix} \begin{bmatrix} e_{i0} p_i^{re}(0) p_i^{rr}(0|0) \\ +e_{i2} p_i^{re}(2) p_i^{rr}(0|2) \\ +e_{i1} p_i^{re}(1) p_i^{rr}(1|1) \\ +e_{i2} p_i^{re}(2) p_i^{rr}(2|2) - 1 \end{bmatrix} & \begin{bmatrix} e_{i1} p_i^{re}(1) p_i^{rr}(0|1) \\ +e_{i2} p_i^{re}(2) p_i^{rr}(1|2) \end{bmatrix} \\ \begin{bmatrix} e_{i1} p_i^{re}(1) p_i^{rr}(0|1) \\ +e_{i2} p_i^{re}(2) p_i^{rr}(1|2) \end{bmatrix} & \begin{bmatrix} e_{i0} p_i^{re}(0) p_i^{rr}(0|0) \\ +e_{i2} p_i^{re}(2) p_i^{rr}(0|2) \\ +e_{i1} p_i^{re}(1) p_i^{rr}(1|1) \\ +e_{i2} p_i^{re}(2) p_i^{rr}(2|2) - 1 \end{bmatrix} \end{pmatrix} \begin{pmatrix} R_{i1} \\ R_{i2} \end{pmatrix} = - \begin{pmatrix} \begin{bmatrix} a_{i0} N_{i1} p_i^{re}(0) p_i^{rr}(0|0) \\ +a_{i2} N_{i1} p_i^{re}(2) p_i^{rr}(0|2) \\ +a_{i1} N_{i1} p_i^{re}(1) p_i^{rr}(1|1) \\ +a_{i2} N_{i1} p_i^{re}(2) p_i^{rr}(2|2) \end{bmatrix} + \begin{bmatrix} a_{i1} N_{i2} p_i^{re}(1) p_i^{rr}(0|1) \\ +a_{i2} N_{i2} p_i^{re}(2) p_i^{rr}(1|2) \end{bmatrix} \\ \begin{bmatrix} a_{i0} N_{i2} p_i^{re}(0) p_i^{rr}(0|0) \\ +a_{i2} N_{i2} p_i^{re}(2) p_i^{rr}(0|2) \\ +a_{i1} N_{i2} p_i^{re}(1) p_i^{rr}(1|1) \\ +a_{i2} N_{i2} p_i^{re}(2) p_i^{rr}(2|2) \end{bmatrix} + \begin{bmatrix} a_{i1} N_{i1} p_i^{re}(1) p_i^{rr}(0|1) \\ +a_{i2} N_{i1} p_i^{re}(2) p_i^{rr}(1|2) \end{bmatrix} \end{pmatrix}$$

In this case our matrices are in the form of:

$\begin{bmatrix} a_1 & a_2 \\ a_2 & a_1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$ which means we are solving $a_1x_1 + a_2x_2 = c_1$ and $a_2x_1 + a_1x_2 = c_2$ equations. When we solve these equations, the solutions for x_1 and x_2 are as:

$x_1 = \frac{a_1c_1 - a_2c_2}{a_1^2 - a_2^2}$ and $x_2 = \frac{a_1c_2 - a_2c_1}{a_1^2 - a_2^2}$.

For $T = 3$ and $L_i = 2$, we have $\begin{bmatrix} a_1 & a_2 & a_3 \\ a_2 & a_3 & a_1 \\ a_3 & a_1 & a_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$ the solutions for x_1, x_2 and x_3 will be $x_1 = \frac{a_1^2c_1 - a_2a_3c_1 + a_2^2c_2 - a_1a_3c_2 - a_1a_2c_3 + a_3^2c_3}{a_1^3 + a_2^3 + a_3^3 - 3a_1a_2a_3}$, $x_2 = \frac{a_1^2c_3 - a_1a_3c_1 + a_2^2c_1 - a_1a_3c_3 - a_1a_2c_2 + a_3^2c_2}{a_1^3 + a_2^3 + a_3^3 - 3a_1a_2a_3}$, and $x_3 = \frac{a_1^2c_2 - a_1a_2c_1 + a_2^2c_3 - a_1a_3c_3 - a_2a_3c_2 + a_3^2c_1}{a_1^3 + a_2^3 + a_3^3 - 3a_1a_2a_3}$.

Based on these equations if we substitute equivalent expressions for $R_{i,t}$ we can again obtain a multichoice knapsack problem. However, in this case we cannot provide a compact representation of $R_{i,t}$ as the structure of these expressions are unique to the problem size (T, L_i). Above we show the case for $T = 2, L_i = 2$ and $T = 3, L_i = 2$ and observe that these do not follow a pattern. Although we do not provide a compact representation of $R_{i,t}$, we observe that this set of equations is solvable. For a given setting, it is possible to derive $R_{i,t}$ by writing (4.42) as set of equations under the identical cycles assumptions and then replacing $R_{i,t}$ with the equivalent expressions. If the coefficients of a_{il} in both the objective function and the capacity constraints are decreasing, and if the behavior functions are linear then Theorem 4.1 will also be valid for this case. Then we can use a similar Modified MCKP Greedy Algorithm to find the optimal solution. On the other hand, if the coefficients are

decreasing but the functions are not linear, we can still reduce the size of the problem by eliminating the dominated variables and solve the MCKP. In this section we identified a general form of model which is based on $N_{i,t}$, for small problem instances. If the problem has seasonality or another relationship between $N_{i,t}$ as a function of t, it may be possible to find a compact representation of $R_{i,t}$ for larger problems. However, we leave this for future investigation.

4.4 Optimal Solution Performance Compared to the Current System via Simulation

In order to illustrate how an optimal solution would perform in the clinic that motivates this study, we use one of the simulation models presented in Chapter 3. For this illustration we use the following setting:

- 4 Patient Classes: Internal, New External, Subsequent Visit and Established patients
- Average new appointment requests per day (N_i): 17, 10, 44, and 36 respectively
- Average appointment lengths in minutes (s_i): 40, 60, 20, and 41.3 (calculated based on the frequencies of 20, 40, 60 minute appointments that Established patients request, i.e., $0.077 \times 20 + 0.781 \times 40 + 0.142 \times 60$), respectively
- Revenue for each seen patient (α_i^{se}): 0.8, 3, 2, and 1, respectively
- Penalties are set as follows for each behavior ($\beta_i^{no}, \beta_i^{ca}, \beta_i^{re}$): No-show 0.5, Cancellation 0.1, Reschedule 0.05 as a function of the relative value unit.

- Figure 4.4 shows the behavior functions (S: $p_i^{se}(l)$, NS: $p_i^{no}(l)$, C: $p_i^{ca}(l)$, R: $p_i^{re}(l)$) that are used in this setting (recall that these are the same behavior functions that we use in Figure 3.1. Note that because these functions are non-linear (but differentiable), we can solve for the coefficients of the MCKP using Equations (4.58) and (4.59) and solve the problem with an LP solver instead of using the algorithm from Section 4.2.
- We assume that after a reschedule request, a patient's preference for the new appointment date is equally likely within the maximum allowed delay range, that is $e_{il} = 1/L_i \ \forall l \in \{1, \dots, L_i\}$ for each patient class i.
- Maximum daily capacity of the clinic is 54 hours
- The maximum allowed appointment delay (L_i) for each group is set as: 2, 5, 15 and 30 days respectively

Using these inputs we obtain w_{il} and p_{il} values from Equations (4.35) and (4.36), then by plugging these into Model 4.2′ (or equivalently Model 4.1′) and solving in OPL we obtain the following optimal policy: $a_{12} = 1, a_{25} = 1, a_{31} = 0.835, a_{315} = 0.165, a_{430} = 1$. This means based on the optimal policy, among the appointment requests that we observe today, we should assign all internal patients to two days later, all new external patients to five days later, 83.5% of the subsequent visit patients to tomorrow, 16.5% of the subsequent visit patients to 15 days later, and all established patients to 30 days later. We apply this policy to the simulation model that we developed in Section 3.3. We obtain a daily net profit of

131.183 (as a function of the relative value unit), which has 95% CI lower bound of 129.873 and upper bound of 132.493.

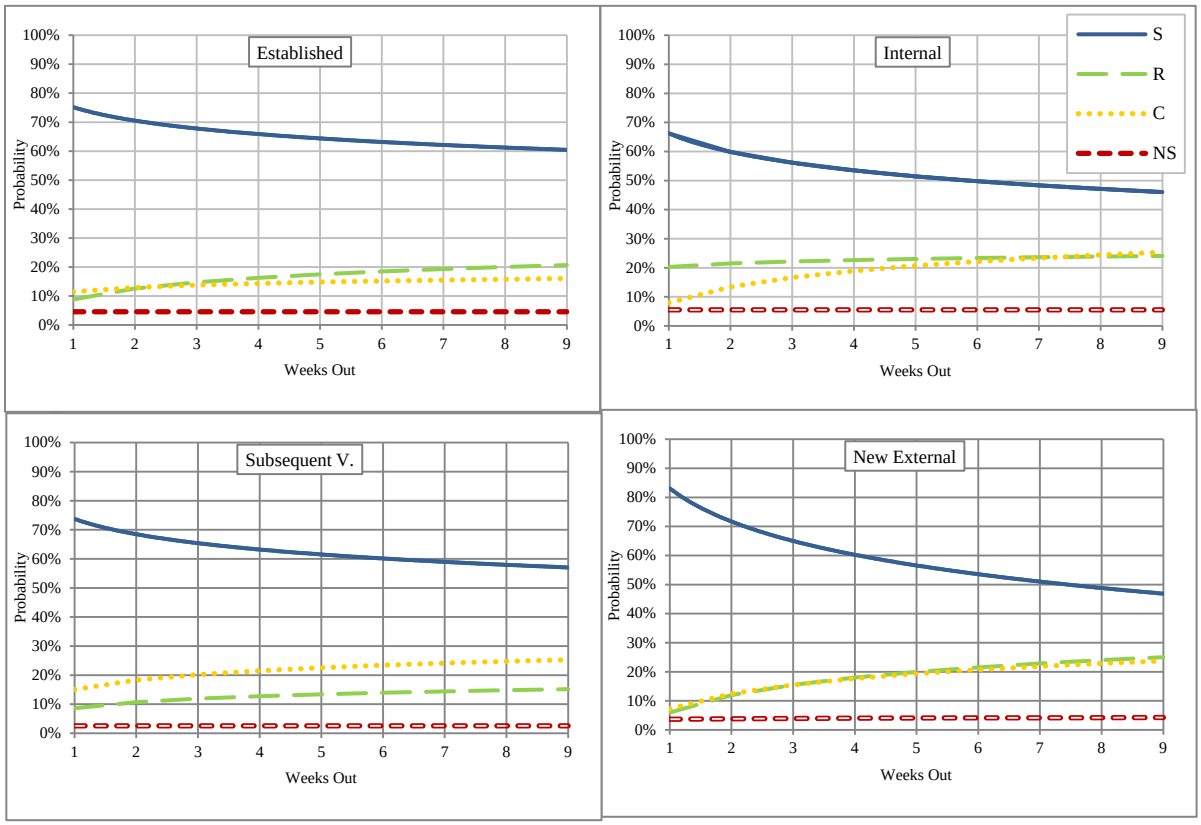

Figure 4.4: Behavior functions that are used for the performance illustration

We compare the optimal result with the result that we obtain by the using current allocation policy (using the days out function in the current system simulation). The value of the current system is 124.25 (a 95% CI of [122.51, 126.00]). Thus, compared to the current system, the optimal solution yields a significant improvement in terms of daily net profit.

Here we should note that these improvements may not seem large but in a large clinic a daily profit improvement of 5.6% can make a significant difference. Moreover, from the service level stand point, the optimal policy yields 80.42% of seen patient rates with 95% CI (79.99%, 80.86%) while the current system yields 77.47% of seen patient rates with 95% CI (77.11%, 77.84%) which is significant in terms of improving patients' access to the system.

4.5 Extended Properties of Model 4.1′ - the LMCKP Model

In this section, we present further analysis of the LMCKP model, Model 4.1′. Based on Sinha and Zoltners (1979), an optimal solution to the LMCKP has at most two fractional variables. Moreover, as they are used in the LMCKP-Greedy algorithm and explained in Section 4.2, the incremental efficiencies, $\tilde{e}_{il} = \frac{\tilde{p}_{il}}{\tilde{w}_{il}} = \frac{p_{il}-p_{il+1}}{w_{il}-w_{il+1}}$, can be used to understand the structure of the optimal solution. Intuitively, $\tilde{e}_{il}$ is the ratio of net profit gained, to the capacity utilized, by moving an appointment to the previous day.

Recall that p_{il} and w_{il} are defined as follows:

$$p_{il} = \left(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l)\right) TN_i +$$

$$p_i^{re}(l)\left(\frac{TN_i}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right) \sum_{k=0}^{L_i} \left(\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k)\right) e_{ik}$$

$$w_{il} = s_i N_i - p_i^{ca}(l) s_i N_i + p_i^{re}(l) s_i N_i \left(\frac{-\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right).$$

Thus we can obtain $\tilde{e}_{il} = \frac{\tilde{p}_{il}}{\tilde{w}_{il}}$ where:

$$\tilde{p}_{il} = (p_{il} - p_{il+1}) = TN_i \Bigg(\alpha_i^{se}\big(p_i^{se}(l) - p_i^{se}(l+1)\big) - \beta_i^{no}\big(p_i^{no}(l) - p_i^{no}(l+1)\big) -$$

$$\beta_i^{ca}\big(p_i^{ca}(l) - p_i^{ca}(l+1)\big) - \beta_i^{re}\big(p_i^{re}(l) - p_i^{re}(l+1)\big) + \big(p_i^{re}(l) - p_i^{re}(l+$$

$$1)\big)\left(\frac{1}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\sum_{k=0}^{L_i}\big(\alpha_i^{se} p_i^{se}(k) - \beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k)\big)\, e_{ik}\Bigg)$$

$$= TN_i \Bigg(\alpha_i^{se}\big(p_i^{se}(l) - p_i^{se}(l+1)\big) - \beta_i^{no}\big(p_i^{no}(l) - p_i^{no}(l+1)\big) - \beta_i^{ca}\big(p_i^{ca}(l) -$$

$$p_i^{ca}(l+1)\big) + \big(p_i^{re}(l) - p_i^{re}(l+1)\big)\left(-\beta_i^{re} + \left(\frac{1}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\sum_{k=0}^{L_i}\big(\alpha_i^{se} p_i^{se}(k) -$$

$$\beta_i^{no} p_i^{no}(k) - \beta_i^{ca} p_i^{ca}(k) - \beta_i^{re} p_i^{re}(k)\big)\, e_{ik}\right)\Bigg), \tag{4.60}$$

and

$$\tilde{w}_{il} = (w_{il} - w_{il+1}) =$$

$$s_i N_i \left(-\big(p_i^{ca}(l) - p_i^{ca}(l+1)\big) + \big(p_i^{re}(l) - p_i^{re}(l+1)\big)\left(\frac{-\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\right) \tag{4.61}$$

The ordering of the $\tilde{e}_{il}$ over all i and l are used to find the optimal solution via the LMCKP-Greedy Algorithm and we can also use $\tilde{e}_{il}$ to characterize the structure of the optimal solution. In the next proposition we show that the incremental efficiency ratios stay the same for all appointment delays for each patient class, if the behavioral functions are logarithmic, as they are in the outpatient clinic in this research.

Proposition 4.7: The incremental efficiency ratios $\tilde{e}_{il}$ and $\tilde{e}_{il+1}$ are equal, i.e., $\frac{\tilde{p}_{il}}{\widetilde{w}_{il}} = \frac{\tilde{p}_{il+1}}{\widetilde{w}_{il+1}}\ \forall l$ for each i, when the behavioral functions are logarithmic with the following form:

$$P\{Patient\ of\ type\ i\ will\ be\ seen\ (or\ cancel\ or\ reschedule\ or\ be\ no\ show)\ |\ The\ patient's\ appointment\ is\ l\ days\ out\} = {m_0}^x \log(l) + {m_1}^x,$$

where m_0 and m_1 are coefficients of the logarithmic functions, and $x \in \{se, ca, re, no\}$.

Proof: We can show this by obtaining $\tilde{p}_{il}$ and $\tilde{p}_{il+1}$, and $\widetilde{w}_{il}$ and $\widetilde{w}_{il+1}$ for any l. We seek to show that the following holds for any l:

$$\frac{\tilde{p}_{il}}{\widetilde{w}_{il}} = \frac{\tilde{p}_{il+1}}{\widetilde{w}_{il+1}}$$

which is equal to

$$\frac{(p_{il} - p_{il+1})}{(w_{il} - w_{il+1})} = \frac{(p_{il+1} - p_{il+2})}{(w_{il+1} - w_{il+2})}$$

or

$$(p_{il} - p_{il+1})(w_{il+1} - w_{il+2}) = (p_{il+1} - p_{il+2})(w_{il} - w_{il+1}).$$

$\tilde{p}_{il}$ given in equation (4.60) can be rewritten as follows:

$$\tilde{p}_{il} = (p_{il} - p_{il+1}) = TN_i \Bigg(\alpha_i^{se}\big(p_i^{se}(l) - p_i^{se}(l+1)\big) - \beta_i^{no}\big(p_i^{no}(l) - p_i^{no}(l+1)\big) - \beta_i^{ca}\big(p_i^{ca}(l) - p_i^{ca}(l+1)\big) - \beta_i^{re}\big(p_i^{re}(l) - p_i^{re}(l+1)\big) + \big(p_i^{re}(l) - p_i^{re}(l+1)\big)\left(\frac{1}{1 - \sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right) \mathrm{A} \Bigg)$$

$$= TN_i\left(\alpha_i^{se}\left(p_i^{se}(l) - p_i^{se}(l+1)\right) - \beta_i^{no}\left(p_i^{no}(l) - p_i^{no}(l+1)\right) - \beta_i^{ca}\left(p_i^{ca}(l) - p_i^{ca}(l+1)\right) + \left(p_i^{re}(l) - p_i^{re}(l+1)\right)\left(-\beta_i^{re} + \left(\frac{1}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)\mathrm{A}\right)\right), \quad (4.62)$$

where $A = \sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k) - \beta_i^{no}p_i^{no}(k) - \beta_i^{ca}p_i^{ca}(k) - \beta_i^{re}p_i^{re}(k)\right)e_{ik}$ is constant,

and $\widetilde{w}_{il}$ given in equation (4.61) can be reorganized as follows:

$\widetilde{w}_{il} = (w_{il} - w_{il+1}) =$

$$s_i N_i\left(-\left(p_i^{ca}(l) - p_i^{ca}(l+1)\right) + \left(p_i^{re}(l) - p_i^{re}(l+1)\right)\left(\frac{-\sum_{k=0}^{L_i} e_{ik}p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)\right). \quad (4.63)$$

Next, we can write $\tilde{p}_{il+1}$ as follows:

$\tilde{p}_{il+1} = (p_{il+1} - p_{il+2}) =$

$$= TN_i\left(\alpha_i^{se}\left(p_i^{se}(l+1) - p_i^{se}(l+2)\right) - \beta_i^{no}\left(p_i^{no}(l+1) - p_i^{no}(l+2)\right) - \beta_i^{ca}\left(p_i^{ca}(l+1) - p_i^{ca}(l+2)\right) + \left(p_i^{re}(l+1) - p_i^{re}(l+2)\right)\left(-\beta_i^{re} + \left(\frac{1}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)\mathrm{A}\right)\right) \quad (4.64)$$

where $A = \sum_{k=0}^{L_i}\left(\alpha_i^{se}p_i^{se}(k) - \beta_i^{no}p_i^{no}(k) - \beta_i^{ca}p_i^{ca}(k) - \beta_i^{re}p_i^{re}(k)\right)e_{ik}$ is constant,

and $\widetilde{w}_{il+1}$ can be written as follows:

$$\widetilde{w}_{il+1} = (w_{il+1} - w_{il+2}) = s_i N_i\left(-\left(p_i^{ca}(l+1) - p_i^{ca}(l+2)\right) + \left(p_i^{re}(l+1) - p_i^{re}(l+2)\right)\left(\frac{-\sum_{k=0}^{L_i} e_{ik}p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im}p_i^{re}(m)}\right)\right). \quad (4.65)$$

Based on equations (4.64) and (4.65), our goal is to show that the following relationship holds:

$$TN_i\Bigg(\alpha_i^{se}\big(p_i^{se}(l)-p_i^{se}(l+1)\big)-\beta_i^{no}\big(p_i^{no}(l)-p_i^{no}(l+1)\big)-\beta_i^{ca}\big(p_i^{ca}(l)-p_i^{ca}(l+1)\big)+$$

$$\big(p_i^{re}(l)-p_i^{re}(l+1)\big)\Bigg(-\beta_i^{re}+\left(\frac{1}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)\mathrm{A}\Bigg)\Bigg)s_iN_i\Bigg(-\big(p_i^{ca}(l+1)-$$

$$p_i^{ca}(l+2)\big)+\big(p_i^{re}(l+1)-p_i^{re}(l+2)\big)\left(\frac{-\sum_{k=0}^{L_i}e_{ik}p_i^{ca}(k)}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)\Bigg)$$

$$=\ TN_i\Bigg(\alpha_i^{se}\big(p_i^{se}(l+1)-p_i^{se}(l+2)\big)-\beta_i^{no}\big(p_i^{no}(l+1)-p_i^{no}(l+2)\big)-\beta_i^{ca}\big(p_i^{ca}(l+$$

$$1)-p_i^{ca}(l+2)\big)+$$

$$\big(p_i^{re}(l+1)-p_i^{re}(l+2)\big)\Bigg(-\beta_i^{re}+\left(\frac{1}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)\mathrm{A}\Bigg)\Bigg)s_iN_i\Bigg(-\big(p_i^{ca}(l)-$$

$$p_i^{ca}(l+1)\big)+\big(p_i^{re}(l)-p_i^{re}(l+1)\big)\left(\frac{-\sum_{k=0}^{L_i}e_{ik}p_i^{ca}(k)}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)\Bigg)\ .$$

Above equation can be further simplified and rewritten as follows:

$$\Bigg(\alpha_i^{se}\big(p_i^{se}(l)-p_i^{se}(l+1)\big)-\beta_i^{no}\big(p_i^{no}(l)-p_i^{no}(l+1)\big)-\beta_i^{ca}\big(p_i^{ca}(l)-p_i^{ca}(l+1)\big)+$$

$$\big(p_i^{re}(l)-p_i^{re}(l+1)\big)\Bigg(-\beta_i^{re}+\left(\frac{1}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)\mathrm{A}\Bigg)\Bigg)\Bigg(-\big(p_i^{ca}(l+1)-p_i^{ca}(l+$$

$$2)\big)+\big(p_i^{re}(l+1)-p_i^{re}(l+2)\big)\left(\frac{-\sum_{k=0}^{L_i}e_{ik}p_i^{ca}(k)}{1-\sum_{m=0}^{L_i}e_{im}p_i^{re}(m)}\right)\Bigg)$$

$$= \left(\alpha_i^{se}\left(p_i^{se}(l+1) - p_i^{se}(l+2)\right) - \beta_i^{no}\left(p_i^{no}(l+1) - p_i^{no}(l+2)\right) - \beta_i^{ca}\left(p_i^{ca}(l+1) - p_i^{ca}(l+2)\right) + \left(p_i^{re}(l+1) - p_i^{re}(l+2)\right)\left(-\beta_i^{re} + \left(\frac{1}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\mathrm{A}\right)\right)\left(-\left(p_i^{ca}(l) - p_i^{ca}(l+1)\right) + \left(p_i^{re}(l) - p_i^{re}(l+1)\right)\left(\frac{-\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\right)$$

And with some additional manipulation we obtain the following equation (4.66):

$$\left(\alpha_i^{se}\left(p_i^{se}(l) - p_i^{se}(l+1)\right) - \beta_i^{no}\left(p_i^{no}(l) - p_i^{no}(l+1)\right) - \beta_i^{ca}\left(p_i^{ca}(l) - p_i^{ca}(l+1)\right) - \beta_i^{re}\left(p_i^{re}(l) - p_i^{re}(l+1)\right) + \left(p_i^{re}(l) - p_i^{re}(l+1)\right)\left(\frac{1}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\mathrm{A}\right)\left(-\left(p_i^{ca}(l+1) - p_i^{ca}(l+2)\right) + \left(p_i^{re}(l+1) - p_i^{re}(l+2)\right)\left(\frac{-\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\right)$$

$$= \left(\alpha_i^{se}\left(p_i^{se}(l+1) - p_i^{se}(l+2)\right) - \beta_i^{no}\left(p_i^{no}(l+1) - p_i^{no}(l+2)\right) - \beta_i^{ca}\left(p_i^{ca}(l+1) - p_i^{ca}(l+2)\right) + \left(p_i^{re}(l+1) - p_i^{re}(l+2)\right)\left(-\beta_i^{re} + \left(\frac{1}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\mathrm{A}\right)\right)\left(-\left(p_i^{ca}(l) - p_i^{ca}(l+1)\right) + \left(p_i^{re}(l) - p_i^{re}(l+1)\right)\left(\frac{-\sum_{k=0}^{L_i} e_{ik} p_i^{ca}(k)}{1-\sum_{m=0}^{L_i} e_{im} p_i^{re}(m)}\right)\right) . \quad (4.66)$$

If we substitute the generic logarithmic behavior functions; $p_i^{se}(l), p_i^{ca}(l)$, $p_i^{re}(l)$, and $p_i^{no}(l)$ with the form of ${m_0}^x \log(l) + {m_1}^x$, we see that the left hand side and right hand side are equal in equation (4.66). This is due to the relationship given in the following Lemma:

Lemma 4.2: $\left(p_i^x(l) - p_i^x(l+1)\right)\left(p_i^y(l+1) - p_i^y(l+2)\right)$

$$= \left(p_i^y(l) - p_i^y(l+1)\right)\left(p_i^x(l+1) - p_i^x(l+2)\right),\ \forall x, y \in \{se, ca, re, no\}, x \neq y, \quad (4.67)$$

where $p_i^x(l) = {m_0}^x \log(l) + {m_1}^x$ and $p_i^y(l) = {m_0}^y \log(l) + {m_1}^y$.

Proof of Lemma 4.2: If we substitute $p_i^x(l)$ and $p_i^y(l)$ functions into the equality (4.67) given in Lemma 4.2, and rearrange the terms, we obtain the following expression for the left hand side of (4.67):

$$({m_0}^x \log(l) - {m_0}^x \log(l+1))({m_0}^y \log(l+1) - {m_0}^y \log(l+2)), \quad (4.68)$$

and for the right hand side of (4.67) the expression becomes:

$$({m_0}^y \log(l) - {m_0}^y \log(l+1))({m_0}^x \log(l+1) - {m_0}^x \log(l+2)). \quad (4.69)$$

It follows directly that the left hand side, (4.68), and right hand side, (4.69), are equal as they are both equal to:

$${m_0}^x {m_0}^y \log(l)\log(l+1) - {m_0}^x {m_0}^y \log(l)\log(l+2) - {m_0}^x {m_0}^y \log(l+1)\log(l+1) + {m_0}^x {m_0}^y \log(l+1)\log(l+2).$$

This completes the proof of Lemma 4.2.

Thus, using Lemma 4.2, we can now conclude that equation (4.66) holds, when the generic logarithmic behavior functions with the form of $p_i^x(l) = {m_0}^x \log(l) + {m_1}^x$ are used, and thus we conclude that $\tilde{e}_{il}$ values for each i are constant for all l. ■

Proposition 4.7 means that although the behavioral functions are nonlinear, the ratio of the increase in the expected net profit gained, to the increase in the utilization of the daily capacity, is the same as moving an appointment to any prior day. This enables us to eliminate the appointment options between the first day and the last allowed day, based on Proposition 4.1a which is related to LP-dominance. Thus, we have the following theorem for the optimal solution structure for the logarithmic case.

Theorem 4.2: If the behavior functions are logarithmic, then there exists an optimal solution in which $a_{il} = 0$ for $\forall i \in \{1, \dots, M\}\ l \in [2, L_i - 1]$.

Proof: This can be proven in a similar manner to Theorem 4.1. Recall that based on Proposition 4.1a, for $w_{ik} \leq w_{il} \leq w_{im}$ and $p_{ik} \leq p_{il} \leq p_{im}$ if $\frac{p_{il}-p_{ik}}{w_{il}-w_{ik}} \leq \frac{p_{im}-p_{il}}{w_{im}-w_{il}}$ is satisfied, then the options k and m dominate option l. Under these conditions, based on Proposition 4.1b, $a_{il} = 0$ is an optimal solution.

From Proposition 4.5a, w_{il} and p_{il} are decreasing in l, i.e., $w_{ik} \leq w_{il} \leq w_{im}$ and $p_{ik} \leq p_{il} \leq p_{im}$ are satisfied for any $k \geq l \geq m$. Moreover, from Proposition 4.7, we know that the inequality $\frac{p_{il}-p_{ik}}{w_{il}-w_{ik}} \leq \frac{p_{im}-p_{il}}{w_{im}-w_{il}}$ is satisfied with equality for $m \leq l \leq k \in [1, L_i]$ such that $k - l = 1$ and $l - m = 1$. For this reason, for a given patient class i, all the appointment

delay options within $[2, L_i - 1]$ range, are being dominated by their lower and upper neighbor values and we can eliminate those options. Thus, we can conclude that there exists an optimal solution in which $a_{il} = 0 \; \forall i \in \{1, \dots, M\}$ and $l \in [2, L_i - 1]$. ■

Recall from Section 4.4 that the optimal solution obtained using the behavior functions, derived from clinic data (which are all logarithmic) are as follows: $a_{1,2} = 1, a_{2,5} = 1, a_{3,1} = 0.835, a_{3,15} = 0.165, a_{4,30} = 1$. These results suggest that, the optimization model prioritizes patient class 3 and tries to assign that patient to the next day, and as proven in Theorem 4.2 it assigns the rest of that patient class to the furthest day (15 for that class). Because there is sufficient capacity the other patient classes are assigned to the furthest days possible, $a_{1,2} = a_{2,5} = a_{4,30} = 1$. We note here that because the $\tilde{e}_{il} = \frac{\tilde{p}_{il}}{\tilde{w}_{il}} = \frac{\tilde{p}_i}{\tilde{w}_i}$ ratios are constant for all patient classes, based on Proposition 4.7 we could use the Greedy Algorithm for the MCKP, explained in Section 4.2 to obtain the same solution without solving the LP. For that purpose we could calculate $\frac{\tilde{p}_i}{\tilde{w}_i}$ for each patient class, which are 39.884, 103.023, 194.815, and 46.952 for patient class 1, 2, 3, and 4, respectively. Thus, based on Theorem 4.2, since we only need to consider either the next day or the last allowed day options for assigning appointments, the Greedy Algorithm would first pick patient class 3 and allocate as much of the next day appointment capacity as possible to class 3 patients. And if, as is this case here, the capacity is filled by the first patient class, then the remaining appointment requests for that patient class will be allocated to the furthest possible day (as defined by the patient class maximum allowed delay). All remaining appointment requests from other

patient classes will be assigned to the furthest day based on their own maximum allowed delay.

4.6 Discussion and Conclusion

In this chapter we focused on the problem of capacity allocation and patient scheduling in an outpatient clinic at the strategic level using mathematical programming. Our aim was to find the optimal policy by which to assign incoming patient appointment requests to future days while maximizing the expected net profit considering delay-based cancellations, no-shows and reschedules under daily capacity constraints. We showed that this LP model is actually a form of the Linear Multiple Choice Knapsack Problem (LMCKP) and using the properties of the behavior functions, we obtain structural properties for the model and its solution. We show that under certain conditions, the optimal policy is to assign the appointments either to the next day or the latest allowed day for a patient class. Based on this idea, we also provide a greedy algorithm by which an optimal solution can be easily found for any T.

In Model 4.1, we assumed that the reschedule probability is known at the time an appointment is initially assigned. In Model 4.2, we introduce a probability distribution for assigning reschedule requests on a given day to model the reschedules more accurately. Our findings suggest that since demand is stationary on each day, decision policies are required to be the same on different days. Because the expected number of unchanged appointments is used to check the capacity, Model 4.2 is actually an equivalent to Model 4.1. Thus, the same properties and propositions that obtained for Model 4.1 are also valid for Model 4.2.

In reality, the timing of the reschedule requests is important, because the earlier it is known that an occupied slot will be available, the more likely the slot be filled. Moreover, there is variability in the daily appointment requests in the data that should be taken into account, to allocate sufficient capacity to a patient class and to make more accurate decisions. Although the solution to the models, which we present in this chapter, can be used as initial approximate policies for the actual system; we can further improve the system if we use a multi-stage or time-varying approach to assign incoming appointment requests to future days based on the information that we have on a given day. In order to capture the importance of the timing of the reschedule requests, as a next step we propose to model this problem as a time-varying linear program in which we have a different decision variable set on each day. This will enable to the model to respond to changes in the appointments and varying daily demands better.

Chapter 5

Time-varying Optimization Model

In this chapter we extend our work from Chapter 4. We introduce a time-varying optimization model which can be used to incorporate the timing of reschedule requests more accurately. Moreover, thus far we have assumed stationary and deterministic demand; we have only focused on the stochasticity of the patient behaviors as characterized by the delay-based cancellation, reschedule and no-show probabilities. In this chapter we also relax this assumption and use randomly generated appointment requests on each day. In addition, we allow for a different decision variable set for each day.

5.1 Introduction

In this chapter, we formulate and solve the appointment scheduling problem with a time-varying optimization model. Most of the model components are the same as in Section 4.3. In Chapter 4, we assumed that the appointment requests arrive according to a pre-determined constant rate. There, our purpose was to focus on the stochasticity related to the patient behaviors and movements of the appointments based on appointment delays. In this chapter, we generate the daily demands from the appointment request distributions.

In the current literature there are some papers in which, stochastic or time-varying appointment requests are used. Most of these papers assumed the number of appointment requests received on a given day was Poisson distributed, specifically, Patrick, Puterman and Queyranne (2008), Gupta and Wang (2008), Liu, Ziya and Kulkarni (2010) (used Poisson for numerical analysis), Ayvaz and Huh (2010), Feldman et al. (2012), Patrick (2012), and Schutz and Kolisch (2013). In these papers Poisson distributed appointment requests are used, in order to formulate the problem as an MDP. As we explained in Chapter 2, given the size of the problems, most of these papers use heuristics or other methods to solve the MDP models. These papers do not consider the rescheduling behavior of the patients, which adds another level of complexity. This complexity motivated our use of a linear program in Chapter 4. The flexibility of the linear programming formulation means that we do not need to assume Poisson arrivals, instead we fit distributions to the clinic data to generate the appointment requests to be used in the model.

There are few papers, which consider time-varying demand without the Poisson distribution assumption. Stanciu (2009) and Stanciu et al. (2010) use Normally-distributed daily appointment requests, by truncating the negative tail of the Normal density function. They mention that they could have used Weibull distribution since it was another good fit for their demand data, but they preferred to use Normal for the simplicity. Cayirli, Veral and Rosen (2008) use an empirical arrival pattern obtained from their clinic data, while Samorani and LaGanga (2011) use forecasted arrivals for the future appointment requests.

In this chapter, we use the Weibull distribution for generating the number of daily appointment requests, because it was a good fit for the daily demand data obtained from the clinic compared to the other distributions. Moreover, the Weibull distribution is a good option for generating demand in this setting due to its flexibility (with α and β parameters) enabling the representation of the actual daily demand pattern, and its non-negativity (hence no truncation needed).

In order to handle randomly generated appointment requests better, we incorporate daily decisions in the linear programming formulation to allow for a different appointment assignment policy on each day. We explain the details on this model in the following sections.

5.2 Model III: Time-varying Optimization Model to Handle Reschedule Requests and Time-dependent Demands

As discussed in Chapter 4, a better way to model reschedule requests is to incorporate the importance of the timing of the reschedule requests. The solution for Model 4.2′ (which

reduces to Model 4.1′) uses average weights for a_{il}, each of which is calculated based on the expected number of unchanged appointments, considering the expected time of reschedule requests, using the patient behavior functions. This model is like an expected value model. In an expected value model, from the stochastic programming perspective, decisions are made at the beginning of the horizon and it is assumed that all patient behavior probabilities are realized at their expected values. We develop a new optimization model to find an optimal solution in a random environment defined by the probability of a reschedule request for a given day t. The earlier we know that an appointment will change, the more accurately we can allocate the remaining capacity to the other appointment requests, and move the rescheduled appointment to an appropriate date. Therefore, this model must take the timing of the reschedule requests into account.

In this model we introduce a_{ilt} to allow for different decisions at different time epochs, t (stages). In other words, because this problem could be modeled as a multistage stochastic programming model, at each stage t the model would assign appointments based on the decision variables a_{ilt}.

The order of events that are observed at a stage t is as follows:

1. New appointment requests arrive according to N_{it} (the random process)
2. Our knowledge about the available capacity on days $t, t+1, \dots, t+L_i$ is updated as we learn about cancellations and reschedules which create empty slots on the corresponding days, prior to the scheduled appointments.
3. We observe reschedule requests that are made at time (stage) t and need to be assigned to future days.

This structure is not common in multiple stage stochastic programming models, because here the decision that is made at stage t affects the capacity calculations for stages $t, t+1, \ldots, t+L_i$. In addition, the belief that is updated at stage t about stages $t, t+1, \ldots, t+L_i$ is also affected by the decisions that were made in previous stages $(t-L_i, \ldots, t-1)$ and the decision made at stage t. A basic assumption that has been made in studies that use the standard stochastic programming formulation, is exogenous uncertainty which means the realization of uncertainty does not depend on the optimization decisions (Jonsbraten, 1998). However, the main contribution of our research is related to the effect of appointment delays on patient's behaviors. In other words, the optimization decisions (i.e., how far out an appointment should be assigned) impact the probabilities that the patient will be seen on that day, will not show up, or will cancel his/her appointment prior to the appointment. Furthermore, the number of days prior to the appointment, that is the time at which the request to cancel or reschedule the appointment will be made, also depends on the delay that is a function of the optimization decisions. Thus, the exogenous uncertainty assumption does not hold for this problem. For this problem we would need to solve a stochastic programming model with endogenous uncertainty where the optimization decisions affect the realization of uncertainty. Unfortunately, stochastic programming studies that use endogenous uncertainty are very limited in the literature as explained in Gupta and Grossmann (2010). The main reason for this is that as the number of uncertain parameters and their realizations increase, the number of non-anticipativity constraints increases exponentially. Due to the complexity and size of our problem, a stochastic programming approach is not a realistic option and is left for a future exploration. Instead we use a time-varying optimization model to incorporate

the timing of reschedules and randomly generated appointment requests, by allowing for a different set of decision variables for each time period. The notation in this model is similar to Model 4.2, refer to Table 5.1. We used bold font to identify the variables and parameters with a new or different definition.

In this model a_{ilt} represents the percentage of incoming appointment requests of type i at stage/day t (or $N_{it}a_{ilt}$) that are assigned to an appointment l days in the future. The behavior functions are defined similarly to those in Chapter 4. Figure 5.1 illustrates the model. For illustration purposes, we assume that L_i is 2, which means the maximum appointment delay for a patient is 2 days. We will illustrate the case for a single patient type, but in the actual model, we make decisions for each patient type. As shown in Figure 5.1, on day 1 (i.e., stage 1), we assign a percentage of the incoming appointment requests to day 2 and the remaining percentage to day 3 considering the capacities of days 2 and 3 (K_2 and K_3, respectively). Each distribution of these requests represents a node in stages 2 and 3 of the tree. On day 2 (or stage 2), we consider the available capacity on days 2 and 3, based on the expected number of cancellations and reschedule requests that arrive on day 2. We assign the reschedules to future days based on the patients' preferences (e_{il}) and decide how to distribute the incoming new appointment requests to future days (i.e., day 3 and day 4).

Table 5.1: Notation for the time-varying optimization model

Decision variables:	
$\boldsymbol{a_{ilt}}$	Percentage of daily appointment requests that are observed on day t and to be scheduled l days later for patient type i ($l = 0,..,L$ where $L \leq T$, $i = 1,..,M$)
Model parameters:	
α_i^{se}	Revenue obtained from each seen patient of type i ($\alpha_i^{se} \geq 0$)
β_i^{no}	Penalty for no-show per patient of type i ($\beta_i^{no} \geq 0$)
β_i^{ca}	Penalty for cancellation per patient of type i ($\beta_i^{ca} \geq 0$)
β_i^{re}	Penalty for rescheduling per patient of type i ($\beta_i^{re} \geq 0$)
$\boldsymbol{\theta}$	Penalty for overtime per hour
$p_i^{ca}(l)$	Probability of cancellation for patient of type i that scheduled an appt. with a delay of l days
$p_i^{re}(l)$	Probability of rescheduling for patient of type i that scheduled an appt. with a delay of l days
$p_i^{no}(l)$	Probability of not showing up for patient of type i that scheduled an appt. with a delay of l days
$p_i^{se}(l)$	Probability of being seen for patient of type i that scheduled an appt. with a delay of l days
$\boldsymbol{p_i^{rr}(r\|l)}$	Probability that a type i patient with an appointment calls r days prior to her appointment day to reschedule given that her initial appointment delay was l days
e_{il}	Probability that a type i patient wants an appointment l days later than the time of the reschedule request, where $\sum_{l=0}^{L_i} e_{il} = 1 \; \forall i \in \{1, \dots, M\}$
$\boldsymbol{K_t}$	Total capacity (in terms of hours) available for all types of patients on day t
s_i	Service time per patient of type i ($s_i > 0$)
$\boldsymbol{N_{i,t}}$	Number of new appointment requests per patient of type i on day t ($N_{it} \geq 0$)
L_i	Maximum allowable appointment delay for patients of type i
T	Planning horizon
M	Total number of patient types
Variables:	
$\boldsymbol{R_{it}}$	Number of type i patients that requested their appointment be rescheduled on day t
$\boldsymbol{O_t}$	Number of overtime hours on day t

We build the time-varying optimization model, Model 5.1, based on the following logic. In order to find the global optimal solution we include each stage's constraints and decision variables in the one model. This is done to ensure that the global optimum is found as one stage's decision variables depend on the others. This dependence is because of the randomly generated appointment requests on each day and the appointment movements between days based on the patient behaviors. In other words, decisions that we make on stage (day) t actually depend on the decisions that were made at least L_i days prior and L_i days later because an appointment can be assigned a maximum of L_i days later.

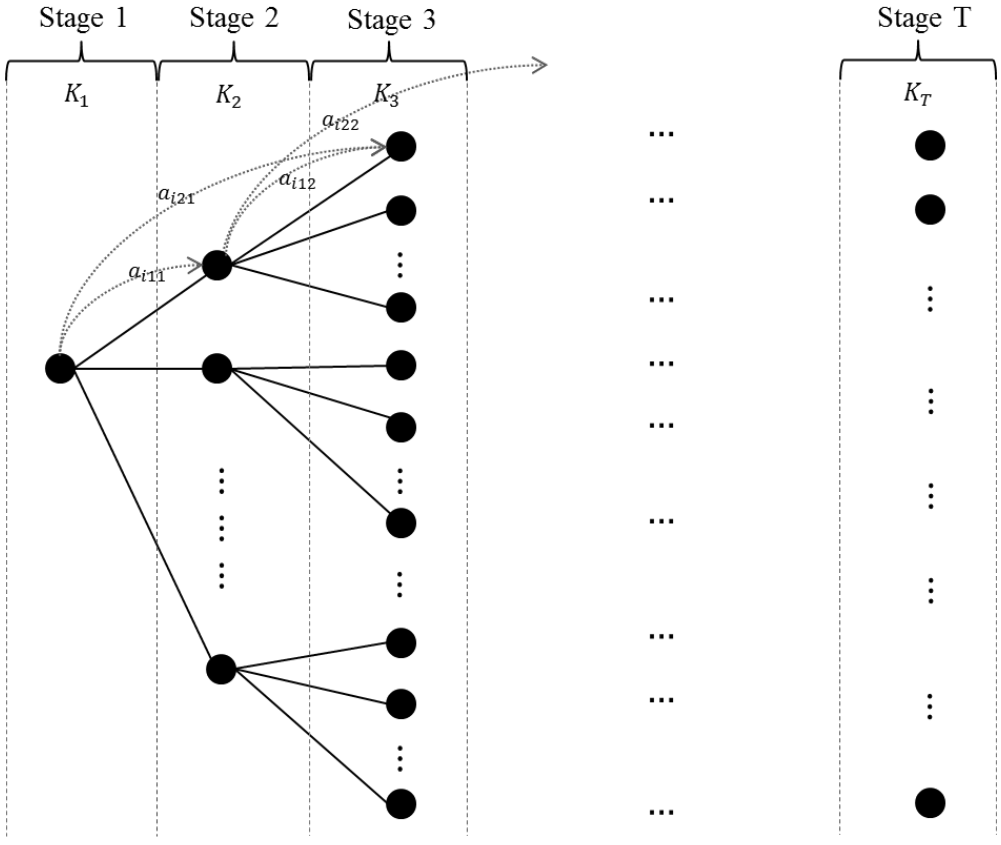

Figure 5.1: Illustration of time-varying optimization model with T-stages

Model 5.1

$$\max \sum_{t=1}^{T} \sum_{i=1}^{M} \Big(\sum_{l=0}^{L_i} \big(R_{i,t-l} e_{il} + N_{i,t-l} a_{i,l,t-l} \big) \big(\alpha_i^{se} p_i^{se}(l) - \beta_i^{no} p_i^{no}(l) - \beta_i^{ca} p_i^{ca}(l) - \beta_i^{re} p_i^{re}(l) \big) - \theta O_t \Big) \quad (5.1)$$

$$s.t.$$

$$\sum_{l=0}^{L_i} \sum_{r=0}^{l} (R_{i,t+r-l} e_{il} + N_{i,t+r-l} a_{i,l,t+r-l}) p_i^{re}(l) p_i^{rr}(r|l) - R_{i,t} = 0 \quad \forall i = 1, \dots, M, t = 1, \dots, T \quad (5.2)$$

$$\sum_{i=1}^{M} \sum_{l=0}^{L_i} \big(p_i^{se}(l) + p_i^{no}(l) \big) \big(R_{i,t-l} e_{il} + N_{i,t-l} a_{i,l,t-l} \big) s_i - K_t - O_t \leq 0 \; \forall t = 1, \dots, T \quad (5.3)$$

$$\sum_{l=1}^{L_i} a_{i,l,t} = 1 \qquad \forall i = 1, \dots, M, t = 1, \dots, T \quad (5.4)$$

$$a_{i,0,t} = 0 \qquad \forall i = 1, \dots, M, t = 1, \dots, T \quad (5.5)$$

$$0 \leq a_{i,l,t} \leq 1 \qquad \forall i = 1, \dots, M, l = 0, \dots, L_i, t = 1, \dots, T \quad (5.6)$$

$$R_{i,t} = 0 \qquad \forall i = 1, \dots, M, t = 1 - L_i, \dots, 0 \quad (5.7)$$

$$a_{i,l,t} = A_{i,l,t} \qquad \forall i = 1, \dots, M, l = 1, \dots, L_i, t = 1 - L_i, \dots, 0 \quad (5.8)$$

Moreover, the decisions that were made more than L_i days earlier can also have an effect on the decision that we make on day t because of the rescheduled appointments that are being moved into future days. An appointment can be assigned to L_i days out, and then it can be requested to be rescheduled any time prior to the appointment. Assume that an appointment is requested to be rescheduled on the day before the appointment day, and the new appointment can be assigned to any future days within the next L_i days assuming that each day is equally likely to be chosen. As we mentioned earlier, we do not make an assignment decision when rescheduling an appointment, thus the furthest day in the future

that our current decision can affect is day $t + 2L_i$. Equivalently, the decision we made on day $t - 2L_i$ can affect our decision on day t. Thus, if we assume that we have enough independent capacity reserved for each patient class, we can say that the decision we are making on day t depends on the decisions that we make within the time window $[t - 2L_i, t + 2L_i]$. However, since in our setting we assume shared capacity is allocated to different patient classes, the decisions that we are making for each patient class also affect the other patient classes. Based on this idea, we can say that the decision on day t actually depends directly on the decisions that we make within the time window $[t - 2\max_i L_i, t + 2\max_i L_i]$. The maximum time range that the decision on day t can depend on is defined by the patient class with the "maximum allowed delay." However, note that it is possible for the dependencies to extend beyond this range due to the chain effect of the appointment requests on different days, i.e., an appointment request that arrives on day $t + 2\max_i L_i$ also affects the appointment assignment decisions on the future days. Thus, although the impact of these future appointments is much less than the appointments within the above mentioned range, they are not independent of each other. For this reason, we analyze our system to identify obtain the bounds under which the significance of this dependency reduces.

Note that for stages $1, .., L_i$ the decision depends on the decisions that are made at stages $1 - L_i, ..., 0$ (which is outside of the horizon defined by the model). In order to solve this problem, we initialize the values that correspond to these stages. We propose using the solution that is obtained by solving Model 4.2′ in which we assume identical cycles and the same demand for each day, to initialize these undefined stage variables. Moreover, we verify with our numerical analysis that the initial values do not significantly affect the decision

variables if we trim some periods from the beginning and end of the planning horizon in order to remove the warm-up and end-of the horizon effects due the variable dependencies within the range $[t - 2\max_i L_i , t + 2\max_i L_i]$ of any stage t.

In Model 5.1, the objective function represents the expected net revenue obtained by assigning incoming appointment requests on day t to future days and the expected net revenue from the appointments that are rescheduled. An overtime option is also added to the model to prevent infeasibility due to high demand in order to keep the appointment assignment policies within the maximum allowed delay ranges for each patient type. To incorporate overtime, the decision variable O_t, the overtime hours on day t, and θ, the penalty for the overtime, are introduced. Constraint (5.2), ensures that the total number of reschedule requests that are received on day t is equal to all previously scheduled appointments with a reschedule request. Constraint (5.3) is the capacity constraint for each day, allowing for the overtime if needed. This constraint ensures the available capacity on each day t is not exceeded considering the seen and no-show patients that were assigned to that day. Constraint (5.4) guarantees that all incoming appointment requests are assigned to the future days, and Constraint (5.5) prevents same day appointments. Constraint (5.6) restricts a_{ilt} to be within $[0,1]$. Constraints (5.7) and (5.8) are the initialization constraints that set R_{it} to 0 and define the range values for the variables a_{ilt} (where A_{ilt} represents the solution from Model 4.2′) for the days within $[1 - L_i, 0]$. As stated previously, we trim the initial periods as well as the final periods in order to prevent the effect of the initialization and the end of horizon, because in reality there is no end or beginning in the appointment

scheduling system. Therefore we would like to obtain the solutions for the stable or steady state system similar to our simulation model.

5.3 Solving Model III

In this section we present the optimal solution for Model III (Model 5.1). However, in order to solve this model requires additional data analysis, to estimate the daily appointment request distributions and the reschedule request timing distributions.

5.3.1 Data Analysis

Number of Appointment Requests per day:

We use JMP Pro 11 to fit distributions for the daily appointment requests for each patient type. The Weibull distribution fits fairly well for all patient types, this is reasonable because the Weibull distribution forces the number of appointments per day to be nonnegative and incorporates the single peak that can be seen in the daily appointment request histograms, see Figures 5.2, 5.3, 5.4, and 5.5. Table 5.2 provides details about the distributions fit to the number of appointment requests per day, including the dataset sample sizes (N) and p-values based on Cramér-von Mises test that JMP uses for Weibull distribution fit, the mean and standard deviation for the number of appointment requests per day for each patient class. Note that although Weibull distribution is ranked as the best based on JMP "fit all" option for two patient classes (new external and internal patients), p-values corresponding to those distributions are still low (0.01 and 0.25 respectively). For other classes (subsequent visits and established patients) Normal 2 or Johnson distribution seems to fit better however both

distributions have low p-values. We use Weibull distribution (although its fit is not identified as the best), because other distributions do not fit significantly better, and we prefer to be consistent for all patient classes. Moreover, the Weibull distribution is easier to use and considering the literature in this area (in which Normal or Poisson distributed demand being used), it is more reasonable to use the Weibull distribution than the Normal or Poisson distributions based on our data analysis.

Table 5.2: Number of Appointments per day Distributions for each Patient Class

	Distribution	**N**	**p-value**	**Mean**	**Std. Dev.**
Internal Patients	Weibull (21.828, 2.948)	144	0.25	19.479	7.206
External Patients	Weibull (11.617, 1.951)	173	0.01	10.335	5.443
Subsequent Visit Patients	Weibull (59.308, 3.646)	128	0.01	53.765	16.533
Established Patients	Weibull (45.344, 2.335)	160	0.01	40.362	18.242

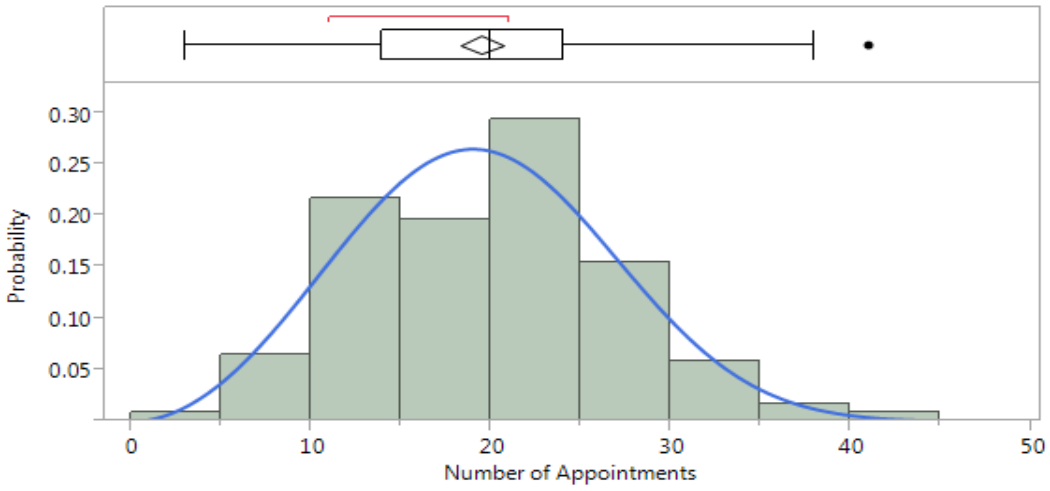

Figure 5.2: Number of Appointment Requests for Internal Patients per day, fit by Weibull Distribution - Weibull(21.828, 2.948)

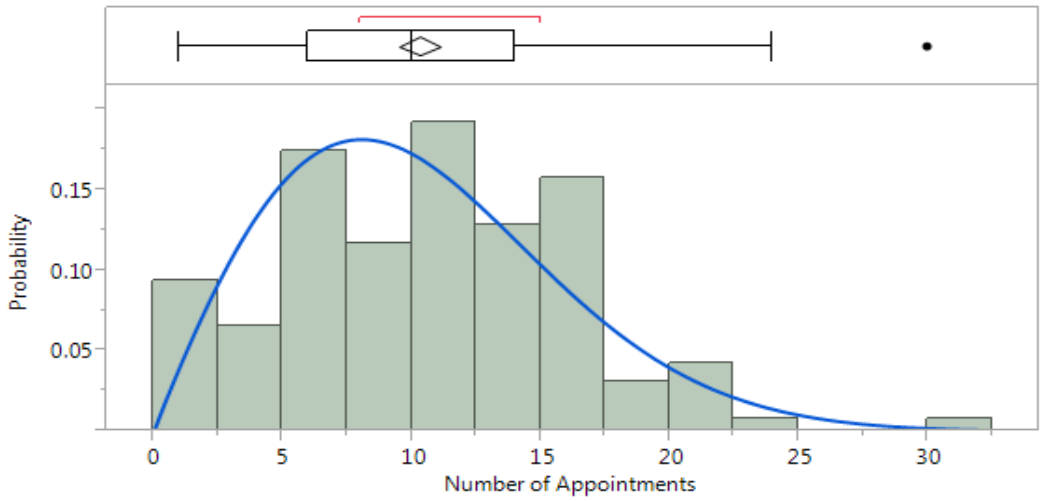

Figure 5.3: Number of Appointment Requests for External Patients per day, fit by Weibull Distribution - Weibull(11.617, 1.951)

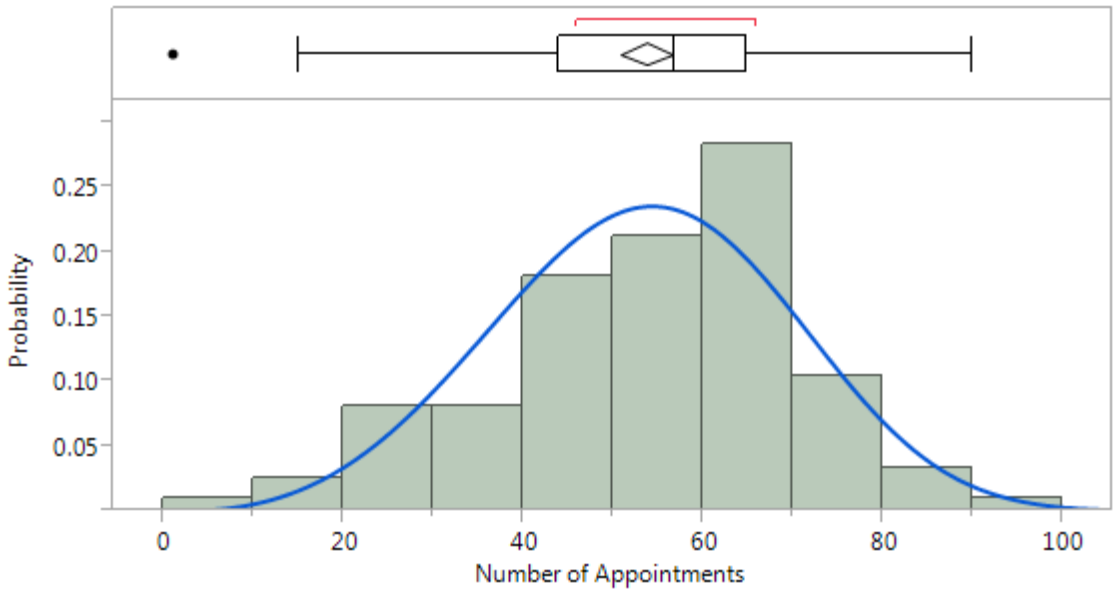

Figure 5.4: Number of Appointment Requests for Subsequent Visit Patients per day, fit by Weibull Distribution - Weibull (59.308, 3.646)

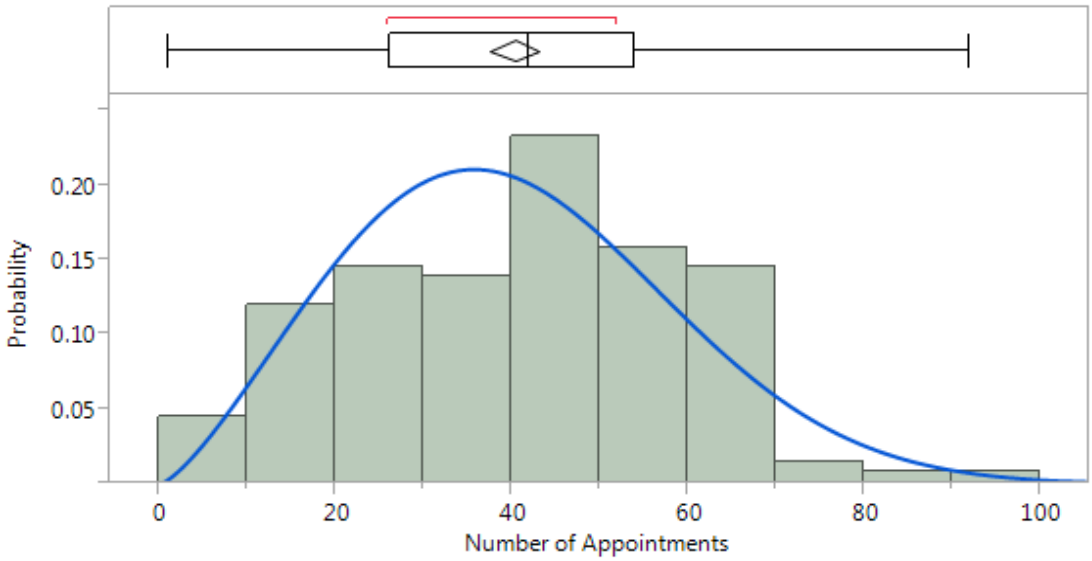

Figure 5.5: Number of Appointment Requests for Established Patients per day, fit by Weibull Distribution - Weibull $(45.344, 2.333)$

Reschedule Request Timing Distributions:

In order to solve Model 5.1 we also need to estimate $p_i^{rr}(r|l)$ - the probability that a reschedule request will arrive r days prior to the appointment, given that the initial appointment delay was l days. We observe that the timing of reschedule requests is characterized by a U shaped function for each patient type. That is, patients either tend to reschedule shortly after an appointment is assigned or wait until close to the appointment date to change their appointment. We plot the function $p_i^{rr}(r|l)$ for each patient type in the Figures 5.6-5.9. Note that each figure only includes the functions within the range of 0 to the maximum allowed delay (i.e., L is 2, 5, 15, and 30 days for Internal, External, Subsequent Visit and Established Patients, respectively). In Figures 5.6-5.9, each line represents the delay (l) in days, the x-axis represents number of days (r) that a reschedule request is

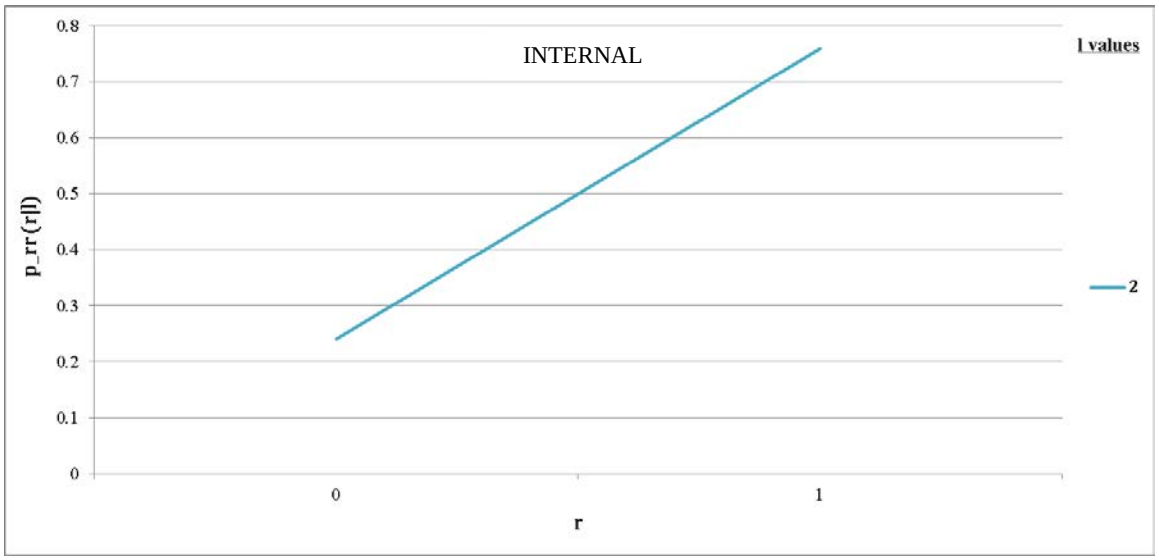

Figure 5.6: $p_i^{rr}(r|l)$ Function for the Internal Patients

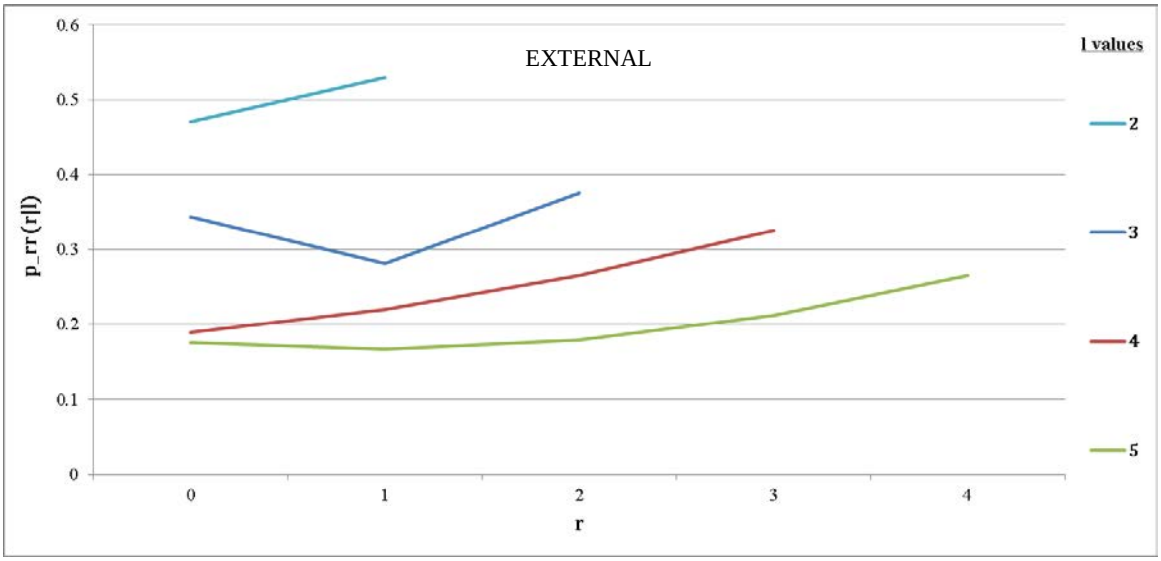

Figure 5.7: $p_i^{rr}(r|l)$ Function for the External Patients

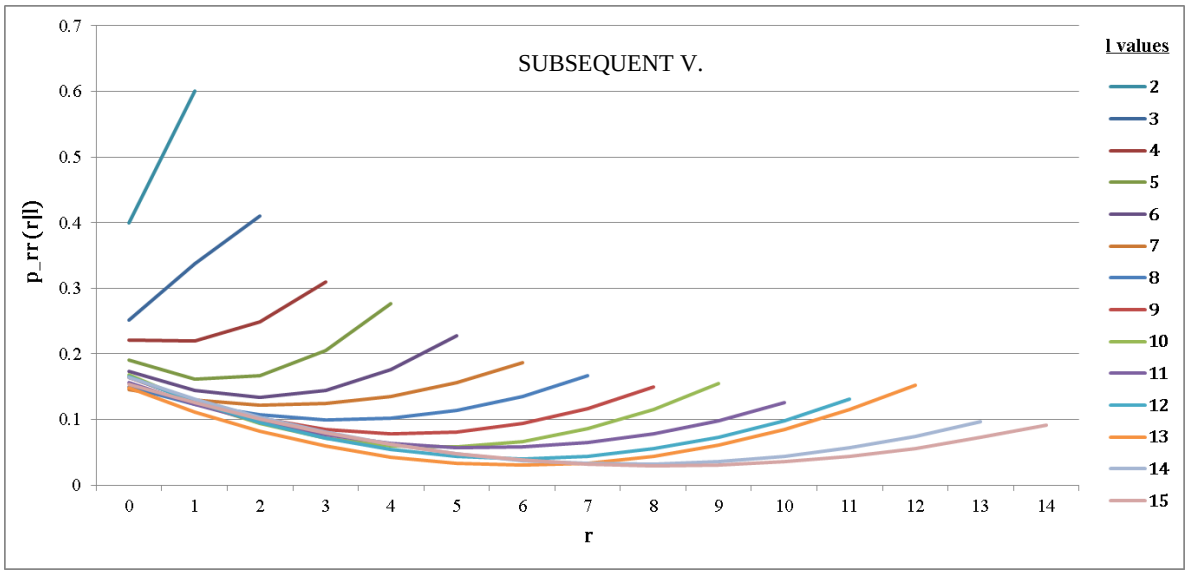

Figure 5.8: $p_i^{rr}(r|l)$ Function for the Subsequent Visit Patients

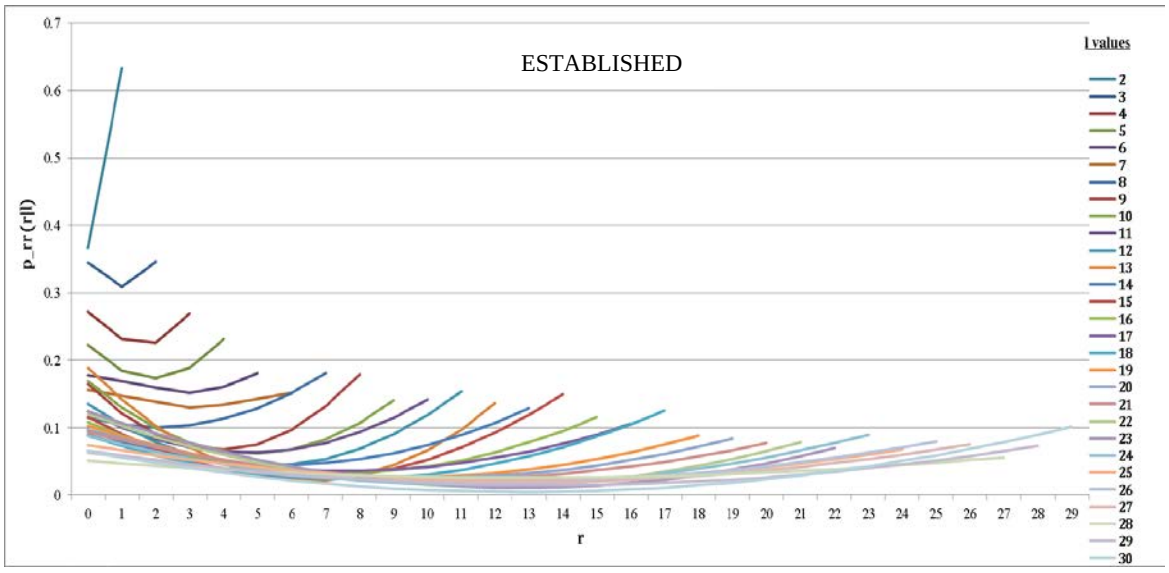

Figure 5.9: $p_i^{rr}(r|l)$ Function for the Established Patients

received prior to appointment, and y-axis represents that probability of that l and r combination. Although we consider integer values for l and r, we use lines in the following graphs to represent their relationship clearly.

5.3.2 Optimal Solution

To solve the Model 5.1, we generate 20 demand samples for 600 days from the Weibull distributions described in Section 5.3.1. As we discuss previously, we eliminate the initial and final periods from the solution to remove the effect of initialization and the end of horizon. For this reason, we initially solve the model using the same mean demand on each day and observe the stable period as defined by the decision variables. We see that the optimization model gives the same solution from day 120 to day 480. This suggests that the initialization and end of horizon effects reduce and become insignificant, 120 days after the planning horizon start day and 120 days before the planning horizon end day. In order to test this stabilization, we added the following constraint to the model:

$$a_{ilt} - a_{ilt-1} = 0 \; \forall t = 120, \dots, 480 \tag{5.9}$$

When we solve the model without Constraint (5.9) the total expected net profit for 600 days is 90,938 units and when we solve the model including Constraint (5.9) the objective value is still 90,938 units. The total net profit within the range of days [120, 480], is 54,656 units for the initial case and 54,655 units for the restricted case; that is 1.8E-5 reduction in the objective value within the periods that we are planning to consider. This suggests that the model, which uses the average demands on each day, stabilizes between days 120 and 480.

Based on this idea, we will solve the model for the 600 day period, but we use the decision variables and the expected net profit obtained for the days between 120 and 480 to minimize the effect of the instable initial and final periods.

As mentioned before, we solve this model for 20 samples of demand, and then we obtain the appointment assignment policy for each sample $s \in [1,20]$, i.e., $a_{ilt}^{s} \in [0,1]$ and $\sum_{l=1}^{L_i} a_{ilt}^{s} = 1$ for each patient type i and day t. The solution obtained for each sample contains a daily assignment policy for each day within the period of [120,480], i.e., 360 daily appointment assignment policies. For a given sample s and patient type i, we take the average of a_{ilt}^{s} over t to obtain the average appointment assignment policy that suggests the percentage of patient type s to be assigned l days out. We can represent this average appointment assignment policy for each sample and patient type with $\boldsymbol{A_i^s}$ vector of size L_i, and each component that is an element of this vector is as $\sum_{t=120}^{480} \frac{a_{ilt}^{s}}{360}$. Based on this approach we obtain 20 average appointment assignment policies for each patient type. We take the mean of these samples to obtain a single appointment assignment policy for each patient type i and we represent this with an $\boldsymbol{A_i}$ vector of size L_i, where $L_1 = 2, L_2 = 5, L_3 = 15, L_4 = 30$, and $\boldsymbol{A_i} = \sum_{s=1}^{20} \boldsymbol{A_i^s} /20$. The solution vectors are as follows:

$\boldsymbol{A_1} = [0,1]$

$\boldsymbol{A_2} = [0.0502, 0.0228, 0.0261, 0.0245, 0.8764]$

$\boldsymbol{A_3} = [0.8991, 0.0161, 0.0053, 0.0035, 0.0019, 0.0019, 0.0016,$

$0.0026, 0.002, 0.003, 0.0028, 0.0041, 0.006, 0.0103, 0.0398]$

$\boldsymbol{A_4} = [0, \dots, 0, 0.0003, 0.002, 0.006, 0.0121, 0.0189, 0.0264, 0.0339, 0.046,$

$0.0611, 0.0812, 0.1075, 0.1551, 0.4495]$

Each element of a vector $\boldsymbol{A_i}$ is the mean of the 20 samples, for this reason we also present the standard deviation for each element as follows:

$\boldsymbol{\sigma_{A_1}} = [0,0]$,

$\boldsymbol{\sigma_{A_2}} = [0.0889, 0.035, 0.0394, 0.035, 0.1893]$

$\boldsymbol{\sigma_{A_3}} = [0.0724, 0.0078, 0.0042, 0.0026, 0.0015, 0.0018, 0.0012,$

$0.0022, 0.0019, 0.0024, 0.0019, 0.0029, 0.005, 0.0076, 0.0417]$

$\boldsymbol{\sigma_{A_4}} = [0, \dots, 0, 0.0013, 0.0072, 0.0147, 0.0217, 0.0222, 0.02015, 0.02523, 0.0387,$

$0.0411, 0.0477, 0.0444, 0.0505, 0.1746]$

This overall average appointment assignment policy suggests the model prioritizes the patients in the following order: $3 > 2 > 4 > 1$. In other words, the model tries to bring the Subsequent Visit Patients (type 3), to the earliest possible day, External Patients (type 2) are next most likely to be assigned to an early day, then the Established Patients and finally the Internal Patients. Note that this order is relative to each patient type's maximum allowed delay L_i. For example, Internal Patients are assigned to their maximum allowed day 2, meaning that they are pushed forward into the future as much as they are allowed. Recall the same order of patient priorities was observed in the solution we obtained in Chapter 4. This is due to the ratio of the expected net profit to the expected physician time usage for a given patient type, considering each patient type's maximum allowed delays and behavior functions.

This optimal policy seems to be a bit complicated to apply from a clinical implementation perspective. For this reason, we need to explore an easy to apply policy as explained in the next section.

5.3.3 Alternative "Easy-to-Apply" Solution

We propose an easy-to-apply policy that has a similar structure to the solutions observed in Chapter 4, that is the appointments are assigned either to the next day or the furthest day. This is a reasonable option, based on what we observe in the current solution of Model 5.1, the assignment policies tend to assign either to the earlier days or later days, and the assignment percentages in the middle tend to be lower than either earlier days or later days. The optimal policy for the Multiple Choice Knapsack problem (Model 4.2′) has this structure. However, for Model 5.1 to have this form of solution the following constraint must be added:

$$a_{ilt} = 0 \; \forall i = 1, \dots, M, l = 2, \dots, L_i - 1, t = 1, \dots, T \tag{5.10}$$

From now on we refer to the model with this additional constraint as Model 5.1′. When Constraint (5.10) is added and Model 5.1′ is solved, the following average appointment solution vectors result:

$\boldsymbol{A_1} = [0.2146, 0.7854]$

$\boldsymbol{A_2} = [0.19, 0, 0, 0, 0.81]$

$\boldsymbol{A_3} = [0.811, 0, \dots, 0, 0.189]$

$\boldsymbol{A_4} = [0.0273, 0, \dots, 0, 0.9727]$.

The standard deviations associated with these vectors are as follows:

$\boldsymbol{\sigma_{A_1}} = [0.0231, 0.0231]$

$\boldsymbol{\sigma_{A_2}} = [0.068, 0, 0, 0, 0.068]$

$\boldsymbol{\sigma_{A_3}} = [0.0709, 0, \ldots, 0, 0.0709]$

$\boldsymbol{\sigma_{A_4}} = [0.0119, 0, \ldots, 0, 0.0119]$.

As shown in Table 5.3, we compare the objective values for each demand sample in order to understand the effect of this constraint, i.e., the decrease in the expected net profit due to Constraint (5.10). The first column represents the demand for each sample, the second column represents the objective values for the original Model 5.1 for each demand sample, the third column represents the objective values for Model 5.1′ after adding Constraint (5.10), and the fourth column represents the percentage change in the net profit when Constraint (5.10) is used.

The maximum reduction observed by solving Model 5.1′ is 1.63%, and the overall average reduction is 1.23% with a 95% CI [1.09%, 1.37%]. This suggests that, the "easy-to-apply" policy performs fairly well and would be a good assignment policy. In the next section, we compare these two solutions with each other, the solution for Model 4.2′ and the current system using the discrete event simulation model described in Chapter 3.

5.4 Optimal Solution Performance of Model III Using Simulation

In this section we use a simulation model similar to the one described in Chapter 3 to compare the performance of the current system at the outpatient clinic, Model II (4.2′), Model III (5.1) and Model 5.1′ with constraint (5.10). This simulation model uses the

appointment request distributions from Section 5.3.1, i.e., Internal Patients with Weibull $(21.828, 2.948)$, External Patients with Weibull $(11.617, 1.951)$, Subsequent Visit Patients with Weibull $(59.3085, 3.64631)$, and Established Patients with Weibull $(45.344, 2.333)$.

Table 5.3: Net Profit Comparison based on Model 5.1 and Model 5.1′

Sample #	**Model 5.1 Net Profit**	**Model 5.1′ Net Profit**	**% Change in the Net Profit if Model 5.1′ is used**
0 (Mean)	54,656	54,655	0.00 %
1	55,037	54,521	-0.94 %
2	53,443	53,234	-0.39 %
3	54,006	53,126	-1.63 %
4	54,503	53,914	-1.08 %
5	53,623	52,755	-1.62 %
6	54,233	53,831	-0.74 %
7	53,876	53,081	-1.48 %
8	53,520	52,776	-1.39 %
9	54,183	53,366	-1.51 %
10	53,356	52,935	-0.79 %
11	54,368	53,575	-1.46 %
12	54,768	54,107	-1.21 %
13	54,871	54,248	-1.14 %
14	53,422	52,725	-1.30 %
15	53,506	52,793	-1.33 %
16	54,750	54,171	-1.06 %
17	53,907	53,221	-1.27 %
18	53,156	52,445	-1.34 %
19	54,072	53,318	-1.39 %

20	54,609	53,759	-1.56 %

We also resolve Model 4.2′ with the updated mean demand based on the analysis in Section 5.3.1. The resulting optimal policy is $a_{1,2} = 1, a_{2,5} = 1, a_{3,1} = 0.3228, a_{3,15} = 0.6772, a_{4,30} = 1$ for all days (recall that the decision variables do not have day t dimension in Model 4.2′).

As we did in Chapter 3, for each policy we use 1000 days to warm up the system, and then we collect results for 260 weekdays (52 weeks). In this case, since the days are not identical (due to the randomly generated appointment requests on each day) we use 10 replications to obtain 95% confidence intervals for the performance indicators (average daily net profit and percentage of seen patients).

Tables 5.4 and 5.5 compare the models in terms of the daily net profit and percentage of the seen patients, respectively. As it can be seen from the tables, the optimal policy obtained using Model 4.2′ still performs significantly better than the current system, with a 2.29% (with a 95% CI [1.86%, 2.72%]) increase in the average daily net profit and a 1.36% (with a 95% CI [1.31%, 1.42%]) increase in the seen patient percentage. The optimal policy for Model 5.1, increases the average daily net profit by 5.98% (with a 95% CI [5.68%, 6.30%]) and the seen-patient percentage increases by 4.01% (with a 95% CI [3.99%, 4.03%]) compared with the current system. Finally, we evaluate the performance of the easy to apply solution that we obtain by using Model 5.1′, in which we have Constraint (5.10). In this case, the average daily net profit increases by 5.06% (with a 95% CI [4.39%, 5.74%]) and the

seen-patient percentage increases by 3.82% (with a 95% CI [3.78%, 3.89%]) compared with the current system.

Table 5.4: Net profit comparison using the simulation model

	Current System	**Model 4.2′**	**Model 5.1**	**Model 5.1′**
Mean	140.57	143.78	148.98	147.68
95% CI Lower Bound	138.69	142.46	147.43	146.65
95% CI Upper Bound	142.45	145.11	150.54	148.70
Compared to the current system:		**2.29%**	**5.98%**	**5.06%**

Table 5.5: Seen patient percentage comparison using the simulation model

	Current System	**Model 4.2′**	**Model 5.1**	**Model 5.1′**
Mean	77.14%	78.19%	80.23%	80.09%
95% CI Lower Bound	76.95%	77.96%	80.06%	79.94%
95% CI Upper Bound	77.32%	78.42%	80.41%	80.24%
Compared to current system:		**1.36%**	**4.01%**	**3.82%**

Recall that Model 4.2′ uses the same stationary demand on each day and provides a single assignment policy, while Model 5.1 uses randomly generated demand on each day an provides different assignment policies on each day. Model 5.1 has the flexibility to balance

higher demand that could be observed on a given day (due to high variance), by adjusting the appointment assignment policy around those days. This is seen in the policies that the model generates for each of 480 days. In fact, in the setting that has Weibull demand arrivals, the solution for Model 5.1 performs better than the Model 4.2′ solution.

Moreover, if we compare the results with the policy obtained by Model 5.1 without Constraint (5.10) with the policy from Model 5.1′ with Constraint (5.10), we see that the differences in the performance indicators are not significant at the 5% level. Using the easy-to-apply policy from Model 5.1′, the clinic will observe significant improvement both in terms of the net profit and seen patient percentages.

5.5 Conclusion

In this chapter we extended the models developed in Chapter 4 to incorporate time-varying demand and the timing of reschedule requests. We also added a time (day) dimension to the decision variables in order to have the flexibility to adjust the optimal policy based on each day's randomly generated demand. Model 5.1 has a different set of decision variables, which provide an appointment assignment policy, for each day. Although this increases the model size significantly, the model can still be solved quickly because it is a linear program. In order to evaluate the appointment assignment policy obtained by this model, we used data from the outpatient clinic data that motivates this research.

For each patient class, we use daily appointment requests and fit a probability distribution, to model the appointment requests more accurately. For all patient classes we fit the Weibull distribution to the actual data. It is reasonable to use the Weibull distribution to

generate demand in this setting, due to its flexibility and non-negativity (unlike the Normal distribution for example).

To characterize the timing of the reschedule requests, we derive functions that estimate the probability an appointment rescheduling request arrives r days prior to the appointment, given that the initial appointment delay is l days. We plot a function for each given delay l, for each patient class i. These functions are U-shaped, because patients either request their appointments be rescheduled shortly after they make the appointment, or they wait until just before the appointment to reschedule. This affects the distribution of the capacity usage associated with rescheduled appointments, because how early we know about the appointment reschedule requests determines the day of new appointment. This is particularly true because demand differs each day. Therefore, incorporating the timing of appointment reschedule requests is important for this setting.

Model 5.1 was solved for 20 samples of demand from the appointment request distributions. Using the average of daily optimal appointment distribution policies, we obtain a single approximate appointment assignment policy. Observing that this solution schedules appointments on several days, we decided to evaluate a heuristic that assigns either next day appointments or last day (based on maximum allowed delay) appointments. For this purpose we added a constraint that forces the model to assign appointments only to the next day, and the last day (today plus maximum allowed delay). For the 20 samples, the average reduction in the objective value was 1.23% with a 95% CI [1.09%, 1.37%].

Finally, we compared the performances of each optimization model to the current system. In order to compare the performance of these policies, we used a discrete event simulation

model similar to the one in Chapter 3. We changed the simulation model to use Weibull distributed appointment requests and applied these policies as well as the current system appointment assignment policy. All three (Model 4.2′, Model 5.1, and Model 5.1′) models performed significantly better than the current system at the outpatient clinic, in terms of the net profit and the percentage of seen patients. While Model 5.1 performed the best, Model 5.1′ which has the additional restriction also performed well. Moreover, since the latter policy is easier to implement, we recommend the outpatient clinic use this policy based on anticipated gain in the net profit of 3.82% with a 95% CI $[3.78\%, 3.89\%]$ and in the seen patient percentage of 5.06% with a 95% CI $[4.39\%, 5.74\%]$ as compared to the current system.

5.6 Future Work

The optimization model presented in this chapter can be extended in a variety of ways. First, similar to reschedule requests, the model can be extended to consider cancellation requests. The earlier we know about a cancellation the more likely that we can fill that cancelled appointment's slot. By including a few parameters related to the timing of cancellations, this can be incorporated in the model. However, due to data limitations, we have left this aspect for future work.

Another extension of the optimization model would be the consideration of appointment slots within a day. In this dissertation, our focus was to find the optimal policy to assign the patients to days and allocate capacity for that patient based on the assignment policies, without considering the time within a day. However, we could introduce another index for

time slots in addition to days. Obviously this will significantly increase the size of the problem. Moreover, in order to obtain the optimal assignment policy at the slot level, we need to use integer programming instead of linear programming, because only one patient can be assigned to a given slot and the solution cannot be in terms of the distribution of the patient classes. Solving this larger integer programming model may require special solution methods.

Finally, as explained in Section 5.2, a stochastic programming approach could be used to incorporate the knowledge of the appointment request, reschedule or cancellation request timing. However, as discussed before, our problem does not satisfy the exogenous uncertainty assumption, which is a basic assumption of one of the commonly used stochastic programming approaches. Because appointment scheduling decisions, affect the patient's probability of being seen, this problem must be solved with a stochastic programming model with endogenous uncertainty. Due to the increase in the size of the problem caused by endogenous uncertainty, this is potentially difficult to solve.

Chapter 6

Conclusion and Future Work

6.1 Summary

In this dissertation we focus on patient scheduling and capacity allocation in outpatient clinics. We incorporate multiple patient classes with the possibility of cancelling, rescheduling or not showing up for appointments. Each of these behaviors depend on how far out the appointment was assigned, i.e., appointment delay and the patient class.

To the best of our knowledge, this is the first work to consider the rescheduling of appointments. The solution without considering reschedules can be infeasible for the actual system, because reschedules are not like cancellations, they involve the moving of appointments between days. When a patient reschedules his/her appointment, a slot on his

initial appointment day will be available, but that appointment will be moved to another day, and, there has to be available capacity for that rescheduled appointment.

Another contribution of this study is related to handling the delay-based behavior functions. Our data analysis indicated that whether a patient is assigned to an earlier or a later date significantly affects the probability that the patient shows up. In fact, we observed that the further in the future the appointment is, the less likely it is that the patient will be seen and the more likely it is that the patient will cancel, reschedule or not show up for his/her appointment. However, until this study, it had been assumed that the probability the patient will be seen is a fixed value, regardless of the appointment delay. In this study, we used logarithmic functions to represent the decrease in the seen probabilities and the increase in the other behavior probabilities as the delay increases.

Using this setting, we built a discrete event simulation model to understand how the appointment windows (i.e., maximum allowed delays) and the slot types affect the system. Under this setting we observed that if we reduce the appointment windows for all patient classes at the same time, the seen patient percentages increase significantly. Moreover, our analysis suggested that 20-minute slot length is the most appropriate appointment slot option to use. For this reason, in the following simulation models we use 20-minute slots and allow for multiple slot usage for the appointment types that require 40 or 60-minute appointments.

As a next step we built a more comprehensive simulation model, in which we evaluated different appointment windows for different patient classes. We also incorporated the dependency relationship between the subsequent visit patients and the other patient classes, i.e., each visit by any other patient class creates a certain number of subsequent visit patient

appointment requests. Under this setting, we tested 54 policy scenarios in addition to the current system. We concluded that it is possible to significantly improve all of the clinic performance indicators by using the scenarios, in which the largest appointment window is 1 week for internal patients, 6 weeks for new external patients, 6 weeks for subsequent visit patients, and 12 weeks for established patients. By only restricting the appointments to be within the abovementioned range, we can improve the clinic performance indicators significantly. Our more detailed analysis with different combinations of appointment windows for different patient classes showed that reducing appointment windows for all patient classes does not always improve the performance indicators (net profit and seen patient percentages). This is due to the dependency between the patient classes, the number of subsequent visit appointments created by each class and the potential revenue gained by each class.

Based on the insights that we gained from the simulation models, we built mathematical programming models to identify optimal appointment scheduling policies. Model 4.1 focused on the stochasticity of the patient behaviors under stationary demand in that model. We were able to show that this model can be reduced to the multi-choice knapsack problem and derive properties related to the optimal policy. For example, the optimal policy is either to assign a next-day appointment or the furthest possible day appointment, based on the maximum allowed delay for each patient class. As a next step, we sought to improve this model by introducing distributions related to the timing of the reschedules, and formulated Model 4.2. However, because of the stationary demand assumption and we used the same decision variables on each day, Model 4.2 also reduced to the same multi-choice knapsack problem.

We solved Model 4.2′ (equivalently Model 4.1) using the actual behavior functions and the mean appointment requests to obtain the optimal solution. We used the simulation model discussed above to compare the optimal solution to the current system. The results indicated that the new policy significantly improves the performance of the clinic in terms of performance indicators, i.e., net profit and the seen patient percentages.

In our last model, we allowed for the number of appointment requests to vary each day according to the distributions derived from clinic data. We built a time-varying optimization model, Model 5.1, to solve this problem. In this case, we handled the timing of reschedule requests more accurately as we allowed for a different set of decision variables each day. We observed that the average appointment distribution policy is difficult to apply as it had appointment assignments on several days. Thus, we forced the model to either assign a next day appointment or a last possible day appointment, as this was the optimal policy structure was observed for Model 4.2′, and called this Model 5.1′. When we solved Model 5.1′, the objective values are not significantly different from Model 5.1, the unrestricted model.

In order to evaluate these policies, we modified the Simulation Model I from Chapter 3, to generate daily appointment requests based on the stochastic demand distributions instead of the stationary demand. We compared the performances of Model 5.1, Model 5.1′, Model 4.2′ and the current system. Based on this comparison, we concluded that Model 5.1 with or without the restriction yields significantly higher net profit and seen-patient percentages, based on the 95% confidence intervals. The optimal policy obtained by Model 5.1 increases the average daily net profit by 5.98% and the seen patient percentage by 4.01%; while the optimal policy obtained by Model 5.1′ increases the average daily net profit by 5.06% and

the seen patient percentage by 3.82% compared with the current system. Because it is easier to apply the policy obtained by Model 5.1′, we recommend this policy for the outpatient clinic.

6.2 Future Work

In this section we describe possible extensions of our work. In Section 6.2.1, we propose to model the dependencies between patient groups, in order to capture the transition of one patient type to another, i.e., a new external patient or an internal patient can become a subsequent visit patient and then an established patient based on the number of appointments the patient makes or the time that has passed after the previous appointment.

Additionally, in the outpatient clinics, physicians usually share specialties. In other words, any physician can take a general type of appointment in addition to his/her own specialties. For this reason there is actually a dependency between the physicians. Thus, another possible extension for our work is to incorporate the dependency between physicians, which we explain in Section 6.2.2.

6.2.1 Dependency of the Patient Classes

As we considered in Simulation Model II in Section 3.4, the patient classes are not independent. Subsequent visits and established patient visits are generated from internal and new external patients. Once a new external patient or an internal patient is admitted to the system, after the initial appointment the patient changes to a subsequent-visit patient; and after multiple visits, the patient changes to an established patient. From Simulation Model II,

we observe that this relationship between patient classes affects the optimal solution in terms of the capacity allocation. If the capacity allocation is assigned ignoring this relationship, then the subsequent visit and established patients, who are generated by other classes, may not enter the clinic in a timely manner, since the solution may defer those patients to later dates as their revenues may be lower than new patients. However, new external and internal patients will eventually turn into subsequent visit and established patients, thus we need to allocate enough capacity to those "generated" patient types in order to not block system with the initial patient classes.

For this extension we need to consider the movement of the patient classes over time or after a certain number of appointments, because a patient changing from one type to another affects the revenue stream. A finite horizon Markov decision process (MDP) model is a possible approach for addressing this problem. However, the main drawback of this approach would be the size of the state and action space.

6.2.2 Dependency of the Physicians

In the models presented in this dissertation the physicians within the seven subspecialties (ES (Esophageal), GF (Gut Failure), HB (Hepatobiliary), IBD (IBD), MO (Motility), NE (Neoplasia) and PA (Pancreas)) are assumed to be independent and have their own capacities. However, in reality a general (GE) appointment can be assigned to any physician. For example, a physician with subspecialty i, i = ES, GF, HB, IBD, MO, NE, PA can divide her available hours between subspecialty appointments and GE appointments. In this study, we assumed that the capacity allocated for GE patients is fixed. However, it is based on the

incoming patient demand, and a physician has the flexibility to have more GE patients and fewer patients of his own subspecialty on one day, and accept no GE patients on another day. This is true for all subspecialties.

This problem can be viewed as incorporating the physicians' preferences regarding how many GE patients to schedule relative to subspecialty patients. An approach for solving the problem of optimizing physician preferences would be to allow flexible capacity allocation for each subspecialty and allow GE type appointments to be assigned more dynamically. Simulation optimization could be one approach for modeling this problem.

6.3 Our Contribution

This dissertation contributes to the patient scheduling and capacity management literature. To the best of our knowledge this is the first work to consider an outpatient clinic with multiple patient classes with no-show and cancellation probabilities. In addition, we consider the rescheduling behavior of patients, which was not considered in the prior literature due to its complexity. Our data analysis reveals that reschedule rates can be significant, so ignoring them could result in suboptimal and potentially infeasible solutions because rescheduled appointments stay in the appointment system creating load on another day.

Another contribution of this research is the consideration of the delay-based behavioral functions, where delay is the number of days between the appointment request date and the actual appointment date. For each of the abovementioned behaviors (cancelling, rescheduling or not showing up), we observe that each patient class has different delay-based behavior functions and the probability of cancelling, rescheduling or not showing-up increases as the

appointment delay increases. In other words, the further in the future the appointment is, the less likely it is that the patient will show-up for the appointment without changing it. Incorporating these delay-based functions into the decision model is crucial, particularly under the multiple patient system where there are different delay-based behavioral functions and different revenue potentials for each patient class.

These factors add complexity to the appointment scheduling and capacity allocation problem. For this reason, before analytically modeling the problem, we developed two discrete event simulation models to understand how different elements in this system affect each other. Simulation Model II also incorporates the dependency between patient classes, which is another contribution to the literature. In the clinic, the subsequent visit appointments are generated by other patient classes with different rates. To the best of our knowledge, none of the multi-patient studies considered this aspect. We were able to capture this factor in our simulation model, and observed that this dependency affects the policy performances. When a patient class that generates more subsequent visit appointments is prioritized in the solution, there is a resulting future impact on capacity. Using the simulation model, we were able to demonstrate the importance of the appointment delay decision while scheduling different classes of patients.

Based on the insights gained from the simulation models, we developed three mathematical programming models. In Chapter 4, our aim was to determine the number of days out to schedule an appointment for a given patient class, to maximize the expected net profit, taking the clinic daily capacity constraints, different demand rates, revenues, and appointment delay-dependent behavioral functions for rescheduling, cancelling, or no-shows,

into account. We showed that under certain conditions this model reduces to a multiple choice knapsack problem, and this helped us to obtain the optimal policy structure. This approach to solve a patient scheduling and capacity allocation problem is new to the literature. Finally, we incorporated time-varying demand and the reschedule request timing in Chapter 5. In Model 5.1, instead of using stationary demand on each day, we used a Weibull distribution to generate different appointment requests on each day which is one step closer to the actual clinic setting. The timing of the reschedule requests is important in this time-dependent setting, because how early we learn about a reschedule affects how far out the rescheduled appointment will be assigned. The addition of these factors further contributes to the literature, as this allows for a different appointment allocation policy on each day in the optimal solution. We observed that these policies significantly improve the outpatient clinic performance in terms of both net profit and seen patient percentages.

In summary, this dissertation contributes to the patient scheduling and capacity management literature, not only in terms of the novel problem components that have been incorporated, but also the modeling approach. This made it possible to obtain nice optimal policy structures for the complex setting of an outpatient clinic.

REFERENCES

Akin, G., Ivy, J., Huschka, T. R., Rohleder, T. R., 2013a, "Simulation-Based Analysis of Scheduling Decisions in an Outpatient Clinic," Proceedings of the 2013 ISERC.

Akin, G., Ivy, J., Huschka, T. R., Rohleder, T. R., Marmor, Y., 2013b "Capacity Management and Patient Scheduling in an Outpatient Clinic Using Discrete Event Simulation," Proceedings of the 2013 Winter Simulation Conference.

Ayvaz, N., Huh, W., 2010, "Allocation of Hospital Capacity to Multiple Types of Patients," Journal of Revenue & Pricing Management, 9(5), 386–398.

Birge, J. R., Louveaux, F., 2011, *Introduction to Stochastic Programming*, New York: Springer Science and Business Media.

Cayirli, T., Veral, E., 2003, "Outpatient Scheduling in Health Care: A Review of Literature," Production and Operations Management, 12(4), 519-549.

Cayirli, T., Veral E., Rosen, H., 2006, "Designing Appointment Scheduling Systems for Ambulatory Care Services," Health Care Management Science, 9, 47-58.

Cayirli, T., Veral, E., Rosen, H., 2008, "Assessment of Patient Classification in Appointment System Design," Production and Operations Management, 17(3), 338-353.

Chakraborty, S., Muthuraman, K., Lawley, M., 2010, "Sequential Clinical Scheduling with Patient No-Shows and General Service Time Distributions," IIE Transactions, 42(5), 354-366.

Feldman, J., Liu, N., Topaloglu, H., Ziya, S., 2012, "Appointment Scheduling under Patient Preference and No-Show Behavior," Working Paper.

Fetter, R. B., Thompson, J. D., 1965, “The Simulation of Hospital Systems,” Operations Research, 13(5), 689-711.

Gallucci, G., Swartz, W., Hackerman, F., 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., Savin, S., Wang, B., 2006, “Managing Patient Service in a Diagnostic Medical Facility,” Operations Research, 54(1), 11-25.

Green, L. V., Savin, S., 2008, “Reducing Delays for Medical Appointments: A Queueing Approach,” Operations Research, 56(6), 1526-1538.

Gunal, M. M., Pidd, M., 2010, “Discrete Event Simulation for Performance Modelling in Health Care: A Review of the Literature,” Journal of Simulation, 4, 42-51.

Guo, M., Wagner, M., West, C., 2004, “Outpatient Clinic Scheduling – A Simulation Approach,” Proceedings of the 2004 Winter Simulation Conference, 1981-1987.

Gupta, D., Denton, B., 2008, “Appointment Scheduling in Health Care: Challenges and Opportunities,” IIE Transactions, 40(9), 800-819.

Gupta, D., Wang, L., 2008, “Revenue Management for a Primary-Care Clinic in the Presence of Patient Choice,” Operations Research, 56(3), 576-592.

Gupta, V., Grossmann, I. E., 2010, “Solution Strategies for Multistage Stochastic Programming with Endogenous Uncertainties,” Computers & Chemical Engineering, 35(11), 2235-2247.

Harper, P. R., Gamlin, H. M., 2003, "Reduced Outpatient Waiting Times with Improved Appointment Scheduling: A Simulation Modelling Approach," OR Spectrum, 25, 207-222.

Hashimoto, F., Bell, S., 1996, "Improving Outpatient Clinic Staffing and Scheduling with Computer Simulation," Journal of General Internal Medicine, 11(3), 182-184.

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

Huh, W. T., Liu, N., Truong, V. A., 2013, "Multi-resource Allocation Scheduling in Dynamic Environments," MSOM, in press.

Jonsbraten, T. W., Wets, R. J. B., Woodruff, D. L., 1998, "A class of stochastic programs with decision dependent random elements," Annals of Operations Research, 82, 83–106.

Jun, J. B., Jacobson, S. H., Swisher, J. R., 1999, "Application of Discrete-Event Simulation in Health Care Clinics: A Survey," Journal of the Operational Research Society, 50(2), 109-123.

Kaandorp, G. C., Koole, G., "Optimal Outpatient Appointment Scheduling," Health Care Management Science, 10, 217-229.

Kellerer, H., Pferschy, U., Pisinger, D., 2004, "The Multiple-Choice Knapsack Problems," *Knapsack Problems*, 317-340, Germany: Springer-Verlag.

Kim, S., Giachetti, R. E., 2006, "A Stochastic Mathematical Appointment Overbooking Model for Healthcare Providers to Improve Profits," IEEE Transactions on Systems, Man, and Cybernetics, Part A, 36(6), 1211-1219.

King, A. J., Wallace, S. W., 2012, *Modeling with Stochastic Programming*, New York: Springer Science and Business Media.

Klassen, K. J., Rohleder, T. R., 1996, "Scheduling Outpatient Appointments in a Dynamic Environment," Journal of Operations Management, 14(20), 83-101.

LaGanga, L. R., Lawrence, S. R., 2007a, "An Appointment Overbooking Model to Improve Client Access and Provider Productivity," POMS College of Service Operations, London.

LaGanga, L. R., Lawrence, S. R., 2007b, "Clinic Overbooking to Improve Patient Access and Increase Provider Productivity," Decision Sciences, 38(2), 251-276.

Liu, B., 2009, "Stochastic Programming," *Theory and Practice of Uncertain Programming*, 25-56, Berlin: Springer-Verlag.

Liu, N., Ziya, S., Kulkarni, V. G., 2010, "Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations," MSOM, 12(2), 347-364.

Martello, S., Toth, P., 1990, "0-1 Multiple Knapsack Problem" *Knapsack Problems: Algorithms and Computer Implementations*, 157-182, England: John Wiley & Sons Ltd.

Muthuraman, K., Lawley, M., 2008, "A Stochastic Overbooking Model for Outpatient Clinical Scheduling with No-Shows", IIE Transactions, 40(9), 820-837.

Patrick, J., Puterman, M. L, Queyranne, M., 2008, "Dynamic Multi-Priority Patient Scheduling for a Diagnostic Resource," Operations Research, 56(6), 1507-1525.

Patrick, J., 2012, "A Markov Decision Model for Determining Optimal Outpatient Scheduling," Health Care Management Science, 15, 91-102.

Ratcliffe, A., Gilland, W., Marucheck, A., 2012, “Revenue Management for Outpatient Appointments: Joint Capacity Control and Overbooking with Class-Dependent No-shows,” Flex Serv Manuf J., 24(4), 516-548.

Samorani, M., LaGanga, L., 2011, “Scheduling Appointments in a Multi-Day Scheduling Horizon Given Individual Show Probabilities,” Working Paper.

Schutz, H., Kolisch, R., 2013, “Capacity Allocation for Demand of Different Customer Product Combinations with Cancellations, No Shows, and Overbooking When There is a Sequential Delivery of Service,” Annals of Operations Research, 206(1), 401-423.

Shapiro, A., Dentcheva, D., Ruszczynski, A., 2009, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM.

Sinha, P., Zoltners, A. A., 1979, “The Multiple-Choice Knapsack Problem,” Operations Research, 27(3), 503-515.

Stanciu, A., 2009, Applications of Revenue Management in Healthcare, Thesis, Joseph M. Katz Graduate School of, Business University of Pittsburgh, Pittsburgh.

Stanciu, A., Vargas, L., May, J., “A Revenue Management Approach for Managing Operating Room Capacity,” Proceedings of the 2010 Winter Simulation Conference.

Talluri, K. T., Van Ryzin G. J., 2004, The Theory and Practice of Revenue Management, Kluwer Academic Publishers.

VanBerkel, P. T., Blake, J. T., 2007, “A Comprehensive Simulation for Wait Time Reduction and Capacity Planning Applied in General Surgery,” Health Care Management Science, 10, 373-385.

Vanden Bosch, P. M., Dietz, D. C., 2000, “Minimizing Expected Waiting in a Medical Appointment System,” IIE Transactions, 32(9), 841-848.

Vanden Bosch, P. M., Dietz, D. C., 2001, “Scheduling and Sequencing Arrivals to an Appointment System,” Journal of Service Research, 4(1), 15-25.

Vanden Bosch, P. M., Dietz, D. C., Simeoni, J. R., 1999, “Scheduling Customer Arrivals to an Appointment System,” Naval Research Logistics, 46(5), 549-559.

White, D. L., Froehle, C. M., Klassen, K. J., 2011, “The Effect of Integrated Scheduling and Capacity Policies on Clinical Efficiency,” Production and Operations Management, 20(3), 442-455.

Wijewickrama, A., Takakuwa, S., 2005, “Simulation Analysis of Appointment Scheduling in an Outpatient Department of Internal Medicine,” Proceedings of the 2005 Winter Simulation Conference, 2264-2273.

Wijewickrama, A., Takakuwa, S., 2008, “Outpatient Appointment Scheduling in a Multi Facility System,” Proceedings of the 2008 Winter Simulation Conference, 1563-1571.

Zeng, B., Turkcan, A., Lin, J., Lawley, M., 2010, “Clinic Scheduling Models with Overbooking for Patients with Heterogeneous No-Show Probabilities,” Annals of Operations Research, 178(1), 121– 144.