August 2015

Guidelines for Scheduling in Primary Care: An Empirically Driven Mathematical Programming Approach

Hyun Jung Alvarez Oh

GUIDELINES FOR SCHEDULING IN PRIMARY CARE:
AN EMPIRICALLY DRIVEN MATHEMATICAL PROGRAMMING APPROACH

A Dissertation Presented

by

HYUN JUNG ALVAREZ OH

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

MAY 2015

Industrial Engineering & Operations Research

GUIDELINES FOR SCHEDULING IN PRIMARY CARE:
AN EMPIRICALLY DRIVEN MATHEMATICAL PROGRAMMING APPROACH

A Dissertation Presented

by

HYUN JUNG ALVAREZ OH

Approved as to style and content by:

____

Ana Muriel, Co-Chair

____

Hari Balasubramanian, Co-Chair

____

Ahmed Ghoniem, Member

____

Donald Fisher, Department Head
Mechanical and Industrial Engineering

DEDICATION

This dissertation is dedicated to my husband, parents and brother who have supported in my whole life.

ACKNOWLEDGMENTS

The word *thank you* is not good enough to appreciate all people who have supported my Ph.D. journey. This journey has been the best because of the support and love from people who I have met during the Ph.D. years.

There is an old Korean saying, “once she/he is your teacher, always a teacher in your life” which means to respect your teacher in the present and future. Professor Ana Muriel and Professor Hari Balasubramanian have been/will be always my teacher with my deep, great respect. I appreciate their thoughtful guidance and support in many aspects through all these years. I would also like to extend my gratitude to my committee member, Ahmed Ghoniem, for his valuable comments and suggestions for this research and life.

I want to thank Atkinson Family Practice to help initiate and provide inspiration for this research. I can say for certain sure that Atkinson practice truly cares for patients. I am very grateful that I had an opportunity to work with them.

A special thank to all my great friends, Asli, Jocelyn, Michael, and Xiaoling who have shared the ups and downs and also warm friendship. I also would like to thank my friends, Bohyeong, Eunjin, Junghwan, Misun, Seoyoung and Yangsook, for lots of laughter and encouragement to continue when the going got rough.

I would like to express my most love and sincere appreciation to my parents, Jinsuk Oh and Yeongsun Bak, and brother, Hyeontaek Oh who provide me their unconditional love and great support in my whole life. Without them, this journey would not have even begun.

I also would like to thank my new family in PR who has given me family love and care. I am deeply grateful that I can share my happiness of life with my new family.

Last person I want to acknowledge is my soul mate, Jose Alvarez. He has made my life and a Ph.D. journey much enjoyable and has been providing me full of happiness. His great support made this journey complete!

ABSTRACT

GUIDELINES FOR SCHEDULING IN PRIMARY CARE:
AN EMPIRICALLY DRIVEN MATHEMATICAL PROGRAMMING APPROACH

MAY 2015

HYUN JUNG ALVAREZ OH, B.S., NAMSEOUL UNIVERSITY

M.S., CHUNGANG UNIVERSITY

M.S., LEHIGH UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ana Muriel and Hari Balasubramanian

Primary care practices play a vital role in healthcare delivery since they are the first point of contact for most patients, and provide health prevention, counseling, education, diagnosis and treatment. Practices, however, face a complex appointment scheduling problem because of the variety of patient conditions, the mix of appointment types, the uncertain service times with providers and non-provider staff (nurses/medical assistants), and no-show rates which all compound into a highly variable and unpredictable flow of patients. The end result is an imbalance between provider idle time and patient waiting time.

To understand the realities of the scheduling problem we analyze empirical data collected from a family medicine practice in Massachusetts. We study the complete chronology of patient flow on nine different workdays and identify the main patient types and sources of inefficiency. Our findings include an easy-to-identify patient classification, and the need to focus on the effective coordination between nurse and provider steps.

We incorporate these findings in an empirically driven stochastic integer programming model that optimizes appointment times and patient sequences given three well-differentiated appointment types. The model considers a session of consecutive appointments for a *single-*

provider primary care practice where one nurse and one provider see the patients. We then extend the integer programming model to account for multiple resources, two nurses and two providers, since we have observed that such *team primary care practices* are common in the course of our data collection study. In these practices, nurses prepare patients for the providers' appointments as a team, while providers are dedicated to their own patients to ensure continuity of care. Our analysis focuses on finding the value of nurse flexibility and understanding the interaction between the schedules of the two providers. The team practice leads us to a challenging and novel multi step multi-resource mixed integer stochastic scheduling formulation, as well as methods to tackle the ensuing computational challenge. We also develop an Excel scheduling tool for both single provider and team practices to explore the performance of different schedules in real time.

Overall, the main objective of the dissertation is to provide easy-to-implement scheduling guidelines for primary care practices using both an empirically driven stochastic optimization model and a simulation tool.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Introduction and Research Motivation

Primary care practices provide the first point of contact between patients and the health care system. They include family physicians, general internists, and pediatricians. Compared to specialty practices, family-focused primary care practices serve “health promotion, disease prevention, health maintenance, counseling, patient education, diagnosis and treatment of acute and chronic illness” according to the American Academy of Family Physicians. Thus, patients visiting the primary care practices have a wide variety of conditions under all ages and genders. In addition, patients can be served by not only providers – physicians, nurse practitioners, and physician assistants, but also non-provider staffs – such as nurses or medical assistants. Primary care practices, therefore, involve significant variability in many dimensions.

Another challenge of primary care practices is a shortage of providers which has been one of the main issues in the health care system. The U.S. government has attempted to increase number of primary care providers. According to U.S. Department of Health & Human Services, $250 million has invested in primary care in 2010, and more than 16,000 new providers will be trained and developed over the next five years. Although an increase in number of providers is one of significant factors that can improve efficiency of the healthcare system, doing so poses considerable challenges in terms of years to be completed, financial support and encouragement to medical students to be primary care providers. Patterson et al. (2012) project primary care workforce needs through 2025 and claim that the number of providers is still insufficient to meet population growth and aging. Thus, it is important to search a solution under current circumstances. The solution can be found by analyzing factors that lead to variability in the patient flow, adversely affecting the utilization of the system.

We collect data from a three-provider family medicine practice in Amherst, Massachusetts. This practice is a representative of primary care practices in the Unites States since many practices have similar patient conditions and a small size. In fact, there are 78% of the practices involving 5 or less providers in the U.S. (Bodenheimer and Pham, 2010).

Unlike a periodically sampled data, our data is for all patients during a workday, and provides a complete chronology of patient flow on a given workday. The patient flow at the practice follows: after notifying arrivals to the receptionist, a patient waits in the lobby until a nurse calls (wait time in the lobby); the nurse examines the patient in the exam room (service time with nurse); after nurse step, the patient begins to wait until her/his provider is ready to see her/him (wait time in the exam room); and finally a provider examines the patient (service time with provider). Among components in the patient flow, we examine factors leading to high variability, which negatively affects the utilization of the practice: patient wait time and provider idle time. Based on data analysis, we propose an easy-to-implement patient classification scheme and quantify the importance of effective coordination between nurse and provider steps.

We first formulate the appointment scheduling problem under our new patient classification and two service stages (nurse and provider) as a stochastic program for a single-provider primary care practice where one nurse and one provider see patients. We suggest robust scheduling guidelines found by optimal schedules and heuristic schedules which provide patient time-of-day preferences. Then, we develop an Excel scheduling tool to dynamically see the performance of schedules as the patients call in. Since it is not easy for small practices to access advanced simulation tools, we build the scheduling tool in Excel, which is widely accessible. Our original algorithm in visual basic for applications (VBA) was based on the model of a single-provider practice. We then further expand the scheduling tool to accommodate multiple nurses and providers. In the course of data collection, we have observed that multiple human resources may be used at the two sequential steps in the patient flow. More specifically, nurses flexibly see patients as a team while providers see their own panel appointments. We call this a team primary care practice.

While developing excel simulation tool for a team primary practice, we are interested in optimal schedules incorporating stochastic service time of both multiple nurses and providers and allowing flexibility in each step. These components make the scheduling problem computationally challenging. We restrict the optimal scheduling formulation for a single patient type which includes complex conditions requiring sufficient service time with both nurses and providers. There are practices scheduling only this type of patients. To improve the computational time, we develop tightening constraints and lower bounds.

Based on optimal and heuristic schedules generated by the stochastic models and Excel simulation tool, we provide broader guidelines for the adequate coordination of nurse and provider, and strategies for introducing slack to counter the effect of variability. We have also further examined sensitivity of no-show rates and coefficients of the objectives.

1.2 Dissertation Overview

The dissertation consists of five chapters.

In Chapter 2, we review literature related to data analysis and appointment scheduling in primary care practices. The review of data analysis includes literature which analyze the observational data and which study NAMCS data. The appointment scheduling literature contains scheduling guidelines regardless of the methodology and appointment scheduling using mathematical programming.

In Chapter 3, we analyze data collected from the primary care practice to understand various factors causing inefficiency of the patient flow – wait time in the lobby, service time with nurse, wait time in the exam room, service time with provider, and total time patient spent in the practice. In addition, we study more factors resulting in variability: service time by patient conditions/appointment types, each provider service time, wait time by hours, and practice utilization by days. Based on empirical study, we suggest new easy-to-identify patient classifications which can be easily employed in many primary care practices. Distinguishing from literature which only considers service time with provider, we suggest an effective coordination of service time with both nurse and provider.

In Chapter 4, we formulate a two-stage stochastic integer program for a single-provider primary care practice in which a nurse and a provider take care of all patients. The model is optimally scheduled and sequenced patient appointments using stochastic service time of two service steps, nurse and provider, and new patient classifications. The objective of the model is minimization of patient wait time and provider idle time. We assign patients into 15-min. incremental appointment slots which the practice currently uses. We suggest the scheduling guidelines obtained by the optimal schedules as well as heuristic schedules which provide patient time-of-day preferences and the practice financial viability. We also compare the performance between the only provider model versus the nurse and provider model since the literature mainly consider only the provider model. We apply the certain weight combination of patient wait time and provider idle time and zero no-show rates for our collaborating practice’s needs. For general practice settings, thus, we study sensitivity of coefficients of objective function and no-show rates.

In Chapter 5, we further study the scheduling problem for a team primary care practice where multiple nurses share patients as a team and multiple providers see their own patients. We develop a mixed integer program with stochastic service time of multiple resources at two sequential stages – nurse and provider. The structure of our model is similar to flexible flow shop (FFS) - multiple stages and machines at each stage. However, unlike FFS determining sequences of multiple job types with deterministic processing time, our model decides appointment times for a single patient type with stochastic service times. The unique structure of our model is flexibility among nurses in the nurse step and flexibility among patients in the provider step. This structure makes the problem challenging. Hence, we have developed constraint tightening and lower bounds to improve the running time. We propose the robust scheduling guidelines and also to investigate the impact between a single-provider versus team primary care practices. In the optimization model, we limit to explore schedules with a single appointment type; thus, we study scheduling problem of multiple appointment types in a team practice using an Excel scheduling tool we develop which can dynamically provide the performance of the schedules as patients call in.

The Chapter 3 and 4 in the dissertation is based on Oh *et al.* (2013) and the section 5.3 in Chapter 5 is based on Oh *et al*. (2014).

CHAPTER 2

LITERATURE REVIEW

The literature review has two sections: data analysis and appointment scheduling. The literature of data analysis in primary care practice consists of analysis of appointment durations reported in the National Ambulatory Medical Care Survey (NAMCS) data and other time/observational studies. In appointment scheduling literature, we review two issues: scheduling guidelines irrespective of the methodology and mathematical programming approaches to appointment scheduling.

2.1 Data Analysis

The analysis of our observational data collected from a family medicine practice is the primary motivating factor in this dissertation. Thus, we review literature which analyzes observational data in primary care practices. In addition, we study papers analyzing NAMCS data, a well known source on outpatient practices collected annually from 1973 to 1981, in 1985, and from 1989-present.

We first review literature with analysis of the observational data. Gottschalk *et al*. (2005) apply a cross-sectional observational study design and time-motion techniques. They study specific times for a physician's tasks, consisting of inside the exam room care and outside the exam room work related to ongoing case load management. The main results indicate: an average of 29.1 patients seen per day, an average 8.6 hour office-based workday, an average 55% of the day spent in the exam room, and an average 14% spent outside the exam room. They also compare service time of patient care from their data to time collected as part of NAMCS 2003. They found that their average service time was significantly less than that of the NAMCS (10.7 vs 18.7 min. $P < 0.01$). Although the total time spent in and out of the exam room was approximately that of NAMCS, their total time was still less than time from NAMCS (13.3 vs

18.7 min. *P* < 0.01). They claim that the results of the NAMCS are inclusive since the total physician work time was overestimated.

Also, Gilchrist *et al*. (2005) focus on the time inside the exam room as well as the time outside the exam room, using direct observation (20-sec. intervals recording system). The main results are: an average 20.1 patients seen per day, an average face time 17.5 min. per patient, an average 8 hours 8 min. of the office-based workday, an average 61% of the day spent in the exam room, and an average 23% devoted to medical activities outside the exam room. Moreover, "medically related out-of-examination room time averaged over the number of patients seen increased the average time spent per patient by 7 minutes." Using paired comparison *t* tests, they also compare real observation time to estimated service time by physician on various activities, resulting in significant overestimation on the part of physician service time as well as charting and dictation time. They also claim that since NAMCS uses self-report methods, service time is overestimated

Tai-Seale *et al*. (2007) employ videotaped data of elderly patients in three different types of practices, collecting from Cook (2002), to determine how specific tasks are allocated for time according to topic priority in the office visit. They collected 392 videotaped visits including 2,557 topics. On average, the number of topics was 6.5 and the visit length was 17.4 min. The median patient talk time was 5.3 min. and physician talk time was 5.2 min. They conclude that contents of visits have a wide range, and limited time is allocated on each topic.

Yawn (2003) conduct direct observations between October 1994 and August 1995 by trained research nurses. They use a logistic regression model with two types, acute and chronic type, as the dependent variable. They found that provider time-use differs between acute and chronic types. For example, chronic conditions are spent more time for history taking, compliance assessment, negotiating, and preventive services since patient behaviors have impact on such conditions. On the other hand, acute conditions include longer time on physical examination, procedure, feedback, and health education because this condition takes more time on diagnosis of symptoms. They conclude that an understanding their findings will assist primary care movements from acute care to chronic care.

Anderson *et al*. (2007) report "responses of a national cross-sectional, online survey of patient's satisfaction" with the goal of assessing the link between wait time and service time on patient satisfaction. In the web-based survey, patients are asked to recall the amount time they spent at the last visit - wait time (until seeing a provider) and service time with a provider. They found that there are 13.5% first visit and 28.4% routine exam or check-up; about 25% patients wait more than half an hour; and 38% patients see the provider less than 10 min. The service time with a provider affects patient satisfaction the most; the lowest level of patient satisfaction is combination of short service time and long wait time, thus both are important factors. To make their findings conclusive, they have another paper, Camacho *et al*. (2006), also investigating the relationship among wait time, service time and patient satisfaction in larger practice settings. They also used patient survey method employing a handheld computer right after a patient visit from two primary and sixteen specialty care clinics. They note that average wait time is about 21 min. (standard deviation - STD 15min.) and service time which is less than 5 min. is about 14%. They also found that satisfaction rating of provider and clinic is decreased by 0.1 as every 10 min. wait increases within less than 5 min. of service time, reduced by 0.3 when less than 4 min. service time, and decreased by 2% per min. in the probability of willingness to return. They conclude the same result as Anderson *et al*. (2007), short service time/long wait time negatively affect patient satisfaction.

Migongo *et al*. (2012) use data from the Kentucky Ambulatory Network (KAN), which "performed a modified replication of the 1997-1998 NAMCS involving 56 community-based primary care clinicians at 24 practice sites between May 2001 and June 2002." The service time with a physician is 14.5 min. on average (STD 8.0, min. 3.0, max. 65.0) and also 88.9% patients has been seen by physicians before. They use a regression tree and a linear mixed model with twenty two potential predictors related to service time with a physician. They found three highest impact factors on service time with a physician: number of diagnoses, non-illness care (such as child check-ups, supervision of pregnancy, and general medical examination has long service time), and previously seen by providers in practice. They also study factors which can influence between longest and shortest time: whether patient are seen by their primary care physicians, received non-illness care, and seen by physician assistant/nurse practitioner/nurse midwife along with the physician. They also provide rules of thumb for scheduling by increasing additional minutes depending on the main predictors.

Next, we study literature analyzing NAMCS data. Blumenthal *et al*. (1999) analyze NAMCS data from the 1991 and 1992. One of the major findings is mean service time of adult visits with primary care physicians (16.3 min), while in the family practice specialty average service time is 15.9 min. From multivariate analysis, the service time is affected by several factors such as elderly or new incoming patients, and the number of diagnostic tests ordered. Based on these factors, the researchers suggest that physicians' service time can be utilized more effectively by improving the patient scheduling.

Kimberly *et al*. (2009) examine NAMCS 2003 data focusing on visits to family medicine physicians. They compare service time of preventive, acute or chronic care from NAMCS 2004 to required service time of three types of care which meet current recommended guidelines. On average, service time per acute care takes the shortest among the three while physicians spend longest hours per day for this care since the physicians see urgent patients as the first consideration than long-term care. Based on number of hours complied with guideline recommendations, however, chronic, preventive, and acute care are in order. Thus, they claim that patients who have chronic disease need to be spent more time with physicians as well as preventive care so that patients can have better outcomes. It will take efforts to this transition due to physician mindset, a change of reimbursement and incentives, and information systems. They suggest a team-based care including physician assistants and nurse practitioners, and community engagements which may help increase time for chronic and prevent care.

Despite being aware of these important findings from previous research which consider factors only related to service time with providers, understanding the patient flow is crucial in primary care practices since problems can arise anywhere in patient flow: from check-in to check-out, and unorganized patient flow results in high wait times and low patient satisfaction (Potisek *et al*. 2007). Few relevant studies explicitly examine the subject of patient flow.

Potisek *et al*. (2007) analyze the patient flow process using a time-motion survey to assess the time usage of each patient during the visit in the anticoagulation and the chronic pain program at the UNC general internal medicine practice. Based on the patient flow analysis, they suggest that certain stages can be improved by relocating the patient room in the same area to make a simpler path for patients to check-in; by transferring certified nurse assistants to the staffs for simple work to improve nursing support in the anticoagulation program and by reviewing patients before the session to prioritize the patients who need to see the pharmacist practitioner in the chronic pain program. Then, they again measure service time after implementation of interventions: on average, total time patient spent at the practice is reduced by 25 min. and 22 min. in each program, respectively. In addition, wait time in certain stages is significantly decreased in both programs. They conclude that the patient flow analysis assists in detecting inefficient stages in the patient flow; and by suggesting brief interventions based on analysis, the patient flow can be significantly improved.

Another relevant study is Stahl *et al*. (2011) which also investigates the patient flow using a radio-frequency-identification (RFID)-based indoor positioning system (IPS) in two different practice settings: a primary care clinic (PC) that applies 15-min. and 20-min. appointment slots and an urgent care clinic (UC) that schedules patients first-in and first-out without fixed appointment slots. They mainly measure flow time which is the total time a patient spends at the practice, wait time from the time of initial tag registration until the time when both a patient and a provider are in the office/exam room, and service time which is the total time a patient and a provider spend together. They found that on average, service time with a provider is about 20 min. longer whereas wait time is around 15 min. shorter in the PC than in the UC. The flow time between two practices does not significantly differ. They discuss that the technology is able to collect patient flow time measure which can provide insight and identify bottlenecks.

2.2 Appointment Scheduling

Research on outpatient appointment scheduling is well established and growing. A comprehensive review of the topic is provided in Cayirli and Veral (2003) and Gupta and Denton (2008). Cayirli and Veral (2003) classifies analysis methodologies into queuing theory, mathematical programming methods, and simulation studies. Among these methodologies, we use mathematical programming methods, driven by empirical data, since they have benefits distinct from other methodologies: unlike queuing theory, no particular assumptions are necessary such as the distributions of inter-arrival times, distributions of service times, and queue capacity; and unlike simulation studies, we can find the exact optimal solution rather than using only heuristics (Berg, 2012). To maintain consistency with the literature, we use the broader term outpatient practice instead of primary care practice in this section. In addition, it is worth to clarify definitions among a scheduling rule, a sequencing rule and an appointment rule; the scheduling rule is composed of the sequencing rule that determines the sequence in which patients will be seen and the appointment rule which assigns specific appointment times to these patients.

Our goal is to provide easy-to-implement scheduling guidelines for primary care practices using a stochastic integer programming approach. We therefore review literature relevant to two main issues: scheduling guidelines irrespective of the methodology; and mathematical programming approaches to appointment scheduling. In addition, we review few papers relevant to a use of simulation by Excel since we develop an Excel simulation scheduling tool which can dynamically show performances of the schedule as patient requests happen. We consider any application setting in outpatient healthcare delivery, including surgery.

2.2.1 Scheduling guidelines irrespective of the Methodology

The most well-known outpatient appointment rule is the *Bailey-Welch* rule (Bailey 1952, Welch 1964) which assigns two appointments for the very first slot and one appointment in the rest of the slots. This rule was shown using queuing models and simulation studies with mean service times. Ho and Lau (1992) using simulation also prove that the Bailey-Welch rule is robust. Soriano (1966) compares one appointment per slot using mean service times to the *Two-at-a-time* rule (two double booked appointments, followed by an empty slot) using queuing theory. He finds that Two-at-a-time is successfully applied to an outpatient department in significantly reducing wait time.

Kaandorp and Koole (2007) use a heuristic local search algorithm to optimize wait time, idle time, and overtime with homogeneous patients with equal slot lengths. They consider three parameters: probability of no-shows, average service time, and total number of patients. They conclude that dome-shaped inter-appointment times are robust; dome-shaped indicates that inter-appointment intervals first increase and then decrease. The optimal appointment rule is very similar to the Bailey-Welch rule with particular parameter values (weights). Using a simulation-based optimization, Klassen and Yoogalingam (2009) find that the modification of dome-shaped inter-appointment times, *plateau-dome* pattern (slot lengths in the dome part are equal), is robust in considering various environment factors, such as number of appointment slots, probability of no-shows, and session lengths.

The literature cited above assumes fixed inter-appointment times. Chew (2011) relaxes this assumption and focuses on determining inter-appointment times given a known number of slots from historical data using a simulation-based heuristic algorithm to minimize expected wait time, idle time and overtime. He finds that as the unit cost for wait time is higher, the inter-appointment times are increased; as the unit cost for idle time is higher, the inter-appointment times are decreased; and if the unit cost for overtime is increased, the last slot is long enough to prevent overtime.

Hassin and Mendel (2008) study the two types of appointment systems, non-fixed inter-appointment times and fixed inter-appointment times, by using queuing systems with a single server considering different show rates for each patient. They find that with no-show rates, their optimal schedule with non-fixed inter-appointment times seems dome-shaped since the appointment interval increases for the first few appointments, then stays almost the same, and then decreases for the last few appointments. With fixed- inter-appointments, the slot length decreases as no-show rates increases.

The papers discussed above assume patients to be homogenous. However, the outpatient practice generally consists of various patient types, each of whose service times involves significantly high variability. Klassen and Rohleder (1996) evaluate different scheduling rules with different types of patient and equal slot lengths by conducting simulation. They conclude the best sequencing rule is to allocate all low variance patients at the beginning of the session and high variance patients toward the end to strike a balance between wait time and idle time. Although this sequencing rule is practical, it is often difficult to have knowledge of variance of each patient type. Based on our empirical study, we find that patients differ in their *mean* durations, but we are unable to classify patients by variance since all patient types vary significantly in their appointment durations (see Chapter 3). Hence, mean service durations could be more tractable to use in patient classification. Cayirli *et al.* (2006) employed mean service times to classify two different patient types (new and return patients, which correspond to long and short mean service times, respectively). They use discrete event simulation to evaluate various types of scheduling rules using empirical data with the goal of reducing wait time and idle/overtime. It is interesting to note that although service time variability is statistically different, it has less significant impact on the performance in comparison to the clinic size, no-shows, walk-ins, and patient punctuality. Among appointment rules, the Bailey-Welch rule is close enough to the efficient frontiers and can be applied to all sequencing rules they tested. Cayirli *et al.* (2008) extends the study of Cayirli *et al.* (2006) by comparing schedules with equal inter-appointment times with schedules that set two different inter-appointment times equal to the mean of new and return patient service times, respectively. They show that when the cost of provider idle time is high relative to that of patient wait, a schedule using an SPT (shortest processing time) sequence and following the Bailey-Welch rule along with the two different inter-appointment lengths performs very well.

Most papers have considered only a single step, the provider step, in the patient flow process. However, Gul *et al.* (2011) do consider three independent steps, intake, procedure, and recovery steps, in outpatient procedure centers with the goal of minimizing the expected patient wait time and overtime. They first use discrete event simulation; then they develop a genetic algorithm (GA) (Holland, 1975) to analyze simple sequencing heuristics. Among heuristics, SPT performs the best. In addition, they use their GA to see the impact of rescheduling procedures within a given time-horizon of *n*-days. They conclude that the rescheduling procedures significantly help reduce wait and overtime since a procedure can be assigned to a lower utilization day.

2.2.2 Mathematical Programming Approaches

We next turn to papers that use stochastic optimization programs. Outpatient surgical scheduling is relevant to our work since procedure durations – like service times in primary care – are highly variable. The most relevant papers are Robinson and Chen (2003), Denton and Gupta (2003), Denton *et al*. (2007), Mancilla and Storer (2012), Berg (2012) and Saremi *et al*. (2013).

Robinson and Chen (2003) formulate a stochastic linear program with empirically determined distributions of surgery service times in order to determine inter-appointment times given the known patient sequences. The objective is to minimize the expected weighted sum of patient wait time and provider idle time. They solve it by using Monte Carlo integration (see Hammersley and Handscomb, 1964; Halton, 1970; and Fishman, 1996). They propose a scheduling rule using two different inter-appointment durations, one is applied to the first appointment and the other is assigned to the remaining appointments, and show that it is close to the optimum. Denton and Gupta (2003) also optimize inter-appointment times by formulating a two-stage stochastic linear program, considering different coefficients of wait, idle and overtime. They exploit the L-shaped algorithm (Van Slyke and Wets, 1969) with sequential bounding. They show that inter-appointment times display a dome shape when the ratio of idle to wait cost is high, while look more uniform when the cost ratio is low.

Since the surgeries were all of the same type, the above papers focus only on optimal appointment times. Denton *et al*. (2007) and Mancilla and Storer (2012) consider appointment times as well as the sequencing decisions for different surgery types. Denton *et al*. (2007) optimize the sequences and appointment times of surgeries in operating rooms using a two-stage stochastic programming model. The surgery duration and the schedule are derived from historical data. They find that it is hard to review all possible combination of sequences (*n!*) in their stochastic programming formulation. Thus, they compare actual schedules used in the practice with three different heuristics. Their results confirm that low variance surgeries sequenced earlier in the schedule provides robust performance. In addition, Mancilla and Storer (2012) expand the work of Denton *et al*. (2007). They develop new algorithms using Bender's decomposition to determine the optimal appointment times in settings with fixed slot lengths. In Denton *et al*. (2007), appointment time decisions are not restricted by fixed slot lengths. Mancilla and Storer (2012) compare the cases with equal vs. unequal costs for the different surgeries. In the case of equal costs, the sequencing rule by Denton *et al*. (2007), the assignment of shorter variance cases first, performs quite well. In the case of unequal costs, however, the algorithms based on Bender's decomposition outperform the shorter variance first assignment.

Some papers not only use a mathematical model to find the optimal scheduling rules but also implement them in simulation studies to measure performance. Berg (2012) determine optimal scheduling rules and booking number of procedures using a two-stage stochastic mixed integer program with a single server and five different types of procedures in outpatient centers.

They consider no-show rates; an attendance binary random variable is defined by the no-show probability. They employ two decomposition methods based on the classic L-shaped method and a progressive hedging heuristic (Rockafellar and Wets, 1991). Each method improves solution times and optimality gaps. Their findings are the following: patients who have high variance procedure durations or high no-show probability need to be scheduled towards the end of the session; a double booking occurs as no-show probability increases; the Bailey-Welch rule is followed in the optimal schedule; and the optimal number of patients to schedule is quite robust with regard to estimates of the fixed cost of running the suite. In addition, they use discrete event simulation to compare the actual sequences and schedules from the practice with solutions derived by their single server stochastic model. The patient flow structure in the simulation is similar to Gul *et al.* (2011) and models the registration step and three types of procedure rooms. The stochastic program solutions yield up to 63% higher expected profits than the actual one followed by the practice.

Muthuraman and Lawley (2008), Chakraborty *et al.* (2010), Lin *et al.* (2011), Turkcan *et al*. (2011), and Chakraborty *et al.* (2012) focus on scheduling decisions as patient call-ins arrive sequentially in an outpatient practice. Patients are identical as far as service times are concerned, but differ based on their probability of no-show. These papers establish the importance of considering heterogeneous no-show probabilities of patients in appointment scheduling; they also consider the interaction between heterogeneous no-shows and aspects such as impact of pre-defined slot structures and fairness in performance across patients.

Papers cited above assume a single-provider practice. To further research, we have reviewed papers that have studied multiple resources and multi-steps using an mathematical model. Therefore, we focus on relevant papers that use a mathematical model to deal with complicated practical issues, such as dynamic scheduling, random service time, multiple appointment types, multiple human resources, and multi-steps in the patient flow.

Erdogan and Denton (2013) dynamically schedule appointments as a patient calls in and use random service time of appointments assuming in a single step in the patient flow process. They develop two stochastic linear programming models to optimize appointment times with the objective of minimizing total expected wait time and overtime. The first model is a two-stage stochastic linear program incorporating patient no-shows. They found that without no-shows, optimal inter-arrival times follow a dome-shape; and as no-show rates increase, inter-arrival times decrease and number of double booking increases. The second model is a multistage stochastic linear program which dynamically schedule appointments as patients' call-in based on the first come first serve rule. In order to solve the computationally expensive problem, they use nested decomposition integrated with valid inequalities, a set of two-variable linear programs, and multicut outer linearization. They conduct various computational studies with different cost ratios (overtime/wait), distribution of service time, and appointment request probabilities. They found that without add-on patients who may request appointments on short notice, the dome-shape is optimal, and as number of add-on patients increases, the inter-arrival times for early appointments increase while those for later appointments decrease.

Tang et al. (2014) also consider a solo provider outpatient practice and include multiple appointment types with random service time. More specifically, they develop two models with two types of patients: routine patients with no-show probability and urgent patients arriving on time for an appointment, with an objective of minimizing patient wait time, provider idle and overtime. They formulate a model to optimize the appointments with deterministic service time and prove that all urgent patients need to be booked at the beginning of the session and at least one routine patient is double-booked to moderate the effects of no-shows. In addition, they develop a heuristic algorithm for exponentially distributed service time, which can provide a local optimal solution, since the objective function does not satisfy multimodularity, and thus does not assure global optimality. They study the optimal schedules and compare the performance from policies studied in Cayirli et al. (2008). Also, they investigate various sensitivity analysis of different cost ratio between wait time and idle time, no-show probability, number of patients of each type, appointment slot length, and service time. They conclude that the circumstances of "lower no-show probability, smaller interval lengths, shorter service time, and more urgent patients" can improve the service efficiency of the clinic.

Next, Qu et al. (2013) consider a multiple physician specialty clinic (a women's' clinic) with stochastic service times for multiple appointment types in order to design a weekly scheduling template. The patients schedule appointments with any of the available physicians. They develop a two-phase approach: in Phase I, they formulate a mixed integer linear program to assign one of the appointment categories in a session and determine optimal number of appointments for a specific service type, with the aim of balancing of provider workload among sessions; and in Phase II, they model a two stage stochastic mixed-integer program to allocate the appointments into the equal-length time slots for a session given the optimal results from the Phase I, with the objective of minimization of patient wait time, provider idle time, and overtime. They also consider no-show rates for each appointment type in the model. In order to solve the Phase II problem, they propose a genetic algorithm incorporating a Monte Carlo sampling approach which can provide suboptimal solutions. They conclude that their proposed two-phase approach can obtain the effective scheduling templates, which can significantly reduce patient wait time and provider idle time. Although this paper accommodates multiple physicians, it accounts for a single step and also patients can schedule to any available physicians unlike a general primary care practice, where patients schedule appointments with their personal physician.

Saremi et al. (2013) incorporate random service time of multiple appointment types as well as multiple resources in the multi-stage scheduling of operating rooms for outpatient surgeries. They propose three methods integrates with a tabu search: a discrete-event simulation model using stochastic service time, deterministic integer programming model, and binary programming model using mean service time, so as to minimize patient wait time, completion time, and cancellations. They find that a tabu search method enhanced by the optimization models significantly improves the performance of wait time and completion time than simulation-based tabu search method. They also study several scheduling rules and find that the dome-shaped rule improves wait time; and rules that sequence in increasing order of variance and coefficient of variability of the service times decrease session completion time.

Perez et al. (2013) approach the scheduling problem in nuclear medicine clinics where multiple human resources and multi-steps are involved. They develop three models: offline, online, and stochastic online. The offline scheduling model uses an integer program with the assumption that requests are known in advance, to maximize the number of patients on a given day. For the online scheduling model, on the other hand, they assume that a patient is scheduled upon request. So, they use the same decision variables and constraints from the offline integer programming model but with an objective of minimizing the patient wait time since the online scheduling model intends to provide the best schedule for each new request. Then, they extend the online model, accounting for possible future patient requests. They develop a two stage stochastic integer programming model: the first stage determines schedules of the current patients and of resources; and the second stage solves the problem based on scenarios of possible future requests arrival. In this study, the wait time refers to the time duration between the request and the actual appointment. They conduct various experiments based on performance of average number of patients served, utilization of resources by each station and human resource, wait time, and patient preferences by different patient demand scenarios.

Next, we review papers related to the flexible flow shop problem since our scheduling problem has the similar structure to a two stage flexible flow shop: two sequential stages - nurse and provider and two machines (human resources) at each stage. While FFS determines the start time and sequence of multiple job types with deterministic processing time, our problem optimizes appointment times with homogeneous patients with stochastic service time. Wang (2005), Ruiz and Vazquez-Rodriguez (2010), and Ribas et al. (2010) reviewed literature on FFS scheduling problem classified by solution techniques: exact (optimization), heuristics, metaheurisitic, and hybrid approaches. Among solution approaches, Kis and Pesch (2005) review exact solution approach of non-preemptive FFS problems with two objectives: minimization of makespan and mean flow time. Santos et al (1995) develop the global lower bound for hybrid flow shop with the objective of the minimization of makespan. Sawik (2000) formulates mixed integer program for FFS scheduling problem with finite capacity buffers or without buffer. This model is expanded to adopt various configurations to schedule surface mount technology lines (Sawik 2001 and Sawik et al 2002). Sawik (2005) develops integer programming model incorporating make-to-order manufacturing environment with various due dates. The literature cited above discusses various techniques to improve lower bounds and branch-bound method, which have proven so far to assist the effective progress of the exact methods.

A few papers in health care consider scheduling problems as a job shop. Hsu et al. (2003) formulate a patient scheduling problem of an ambulatory surgical center as a no-wait, two-stage process shop scheduling problem, with an objective of minimization of number of PACU (postanesthesia care unit) nurses. They consider two sequential steps – several operating rooms and one PACU, and multiple resources in each step. They develop a tabu search-based heuristic algorithm to solve the problem and find near optimal schedules. Chien et al. (2008) structure a patient scheduling problem as a hybrid shop scheduling problem with partial precedence constraints. They develop a proposed genetic algorithm for rehabilitation treatment operations in order to reduce patient wait time and improve utilization of medical resources. In this case, each patient undergoes different physical therapies and uses the multiple medical resources. They also formulate a mixed integer programming model for small cases as a benchmark with deterministic parameters and multiple replications to validate the solutions of the proposed algorithm. In addition, they develop a decision support system incorporating the proposed algorithm and found that their proposed system increases service quality and improves the utilization based on experiments using empirical data. Pham and Klinkert (2008) also formulate a mixed integer linear programming model as an extension of job shop scheduling problem, for a surgical case scheduling of both inpatients and outpatients. They employ multiple resources in three surgical steps with known, deterministic service time. They conclude that small to medium size cases can be solved within feasible solutions. Shoshana et al. (2012) propose a dynamic scheduling template for a chemotherapy center. They formulate the scheduling problem as a constraint programming model of a three stage flexible flow shop problem with the aim of minimizing the makespan. They first create a proactive template based on known requests generated by a deterministic optimization model; then apply the template to schedule appointments as requests arrive. When a new request does not fit any appointments in the template, the template is dynamically updated using the model. They find that dynamic template scheduling significantly improves the makespan than current scheduling practice.

In order to improve scheduling in healthcare, many papers use simulation since it can accommodate complex queuing systems and environmental factors (Cayirli and Veral 2003). We review a few relevant papers which consider patient flow using simulation in advanced software tools. Hashimoto and Bell (1996) first conduct a time-motion study of the patient flow in an internal medicine academic practice. They use simulation, coded in Turbo Pascal, to observe the impact of increases in human resources and task variables such as appointment intervals, no-shows, and provider service time. Based on the time study and the simulation results, clinic managers made operational changes. Gul et al. (2011) consider multiple patient flow steps in outpatient procedure centers. They use discrete event simulation and develop a genetic algorithm with the goal of minimizing the expected patient wait time and practice overtime. They found that the shortest processing time rule performs the best among heuristics. Harper and Gamlin (2003) develop a simulation model in the Simul8 package which interfaces with Excel for an ENT (Ear, Nose, and Throat) clinic in a hospital in the United Kingdom. They collect arrival and service time data and also different human resources depending on the clinic session. They test a number of different schedules while considering three performance measures: 1) average wait time in the lobby, 2) percentage of patients who wait more than half an hour in the lobby, and 3) average total time patient spent at the practice. They find that the most significant factor is whether the clinic is able to start its day on time. If the start time is delayed, spreading out the appointments in the session, instead of scheduling patients at the beginning of the session, is a more effective policy.

Unfortunately, these simulation tools are not easy to access for small primary care practices. According to National Ambulatory Medical Care Survey 2010, about 32% visits are to single physician practices. For small practices, it may be effective to use something as widely available as Excel. We review a few papers that have used Excel. Rojas et al. (2011) use LpSolve in Excel to allocate medical staff to consulting rooms for the public hospital in Bogota, Colombia. They formulate a mixed integer linear program and solve it in two stages: the first stage to minimize the number of consulting rooms and the second stage to allocate related specialty physicians. Bagust et al. (1999) use an Excel spreadsheet simulation in order to examine the relationship between the stochastic patient admission demand and available inpatient bed capacity.

2.3 Contributions

In summary, outpatient appointment scheduling is a well studied area. We contribute to the literature in the following ways. Our empirical study identify inefficient components or bottlenecks in the patient flow and provides estimates on service time durations for common patient conditions typically seen in primary care. We then use this data to propose a practical, new patient classification scheme, and use the classification to develop scheduling guidelines. Previous mathematical programming approaches mostly consider a single stochastic service step; if multiple steps are modeled, service times in each step are all assumed to be deterministic.

First, we model a stochastic integer program for a single-provider primary care practice. We explicitly consider both nurse and provider steps in the patient flow process, with stochastic service times in both steps that depend on patient type. Furthermore, in our computational results, we consider a variety of heuristic schedules that accommodate patient time-of-day preferences.

We demonstrate that these schedules have a specific structure that makes them easy to implement in practice, while providing a good balance of patient wait and provider idle times.

Next, we propose a novel stochastic integer programming formulation which can be a representative of many practices where multiple nurses and providers are involved in the stochastic patient flow process. We also incorporate the practical issues at primary care practices: nurses can flexibly see patients; providers have their own dedicated panel appointments; and other intricacies such changes in the sequence of patients that a provider sees. To the best of our knowledge, previous papers have not studied team practices from a mathematical programming perspective. There are few papers that incorporate multiple resources in multiple steps in the patient flow. However, these papers have not studied team based practice and have used deterministic service time in the optimization model which is integrated with metaheuristic search method for random service time.

Then, we develop a user-friendly Excel simulation tool for schedulers to manage appointment schedules which accommodate multiple steps in the patient flow process. Additionally, we use three well-differentiated patient types and random service time. This tool can be easily modified to include more human resources, patient types, and performance measures. The previous studies commonly use an advance simulation tool which is not a readily accessible tool for small practices. In our case study, we use the Excel tool to compare the performance of a single-provider primary care practice versus team primary care practice using schedules from our previous work.

CHAPTER 3

DATA ANALYSIS

3.1 Introduction

Primary care practices involve a higher variety of cases: the same team cares for patients of all ages, from birth to end of life, who suffer from various types of ailments related to both their physical and mental health. There are multiple dimensions to variability in primary care: nature of patient complaint (acute versus chronic); mix of appointments (pre-scheduled versus same-day); and time spent with providers and non-provider staff (nurses/medical assistants). This variability may in turn influence patient wait time and the utilization of providers.

Therefore, understanding the variability is significant in order to increase efficiency of the practice. To comprehend variability and the key predictive factors, a whole patient flow at the practice needs to be analyzed since the problems can be any part of the patient flow: from check-in to check-out. These problems typically result in increase of waiting times and decrease of patient satisfaction from unorganized patient flow process; thus, identifying all factors causing variability is crucial in the patient flow (Potisek *et al*. 2007).

However, most literature relies on a single resource, service time with provider, since providers are the most expensive resource. Although service time with provider is a major factor at the practice, the primary care practices contain considerably more complex factors which results in high variability.

Thus, we analyze empirical data collected at a three provider family medicine practice in Massachusetts. The data was collected using a time-motion study conducted on nine work days in summer and fall of 2011. While the results of our time study may seem restricted to the practice we work with, they are in fact fairly general. This is because the majority of the primary care practices in the United States are small and include similar patient ailments. In fact, 32% of the practices in the U.S. are solo practices, and 78% of the practices consist of 5 providers or less

(Bodenheimer and Pham, 2010). In addition, the types of patient conditions seen at this practice – chronic conditions such as diabetes, depression, fatigue, routine physicals for adults and children; and acute conditions such as sore throat and migraine – are representative of the patients seen in all primary care practices.

In this chapter, we present the empirical study that motivated this dissertation. We first describe the practice and how the data was collected in Section 3.2. Section 3.3 analyzes the data and the insights obtained regarding the patient flow and the variability in service time with nurse and provider for different patient types and ailments. In Section 3.4, we propose the new patient classification and verify this new patient classification with national data in Section 3.5. We summarize our conclusions in Section 3.6.

3.2 Time Study

3.2.1 Data Collection Methodology

We collected data at a family medicine practice in Massachusetts. There are three providers, two physicians and one nurse practitioner, and seven nurses. In general, two providers and two nurses are working but if the waiting room becomes compacted, one more flexible nurse starts to see a patient. Figure 3.1 illustrates the layout of the practice. There are five exam rooms and one pediatric room. The black rectangle indicates the location of the observer who conducted the time-study. We gathered data on nine work days: July 7, 18, 22, August 3, 8, and Oct. 5, 7, 8, 9 in 2011. We observed all patients seen by the providers on these days. At the beginning of the day, we examined the list of prescheduled appointments; at the end of the day, we reviewed the list of all appointments including same-day appointments, no-shows, cancellations, and reschedules. We were thus able to collect the data of all patients during a workday. In other words, our data is not merely a sample; it can construct *a complete chronology of patient flow* on the nine work days.

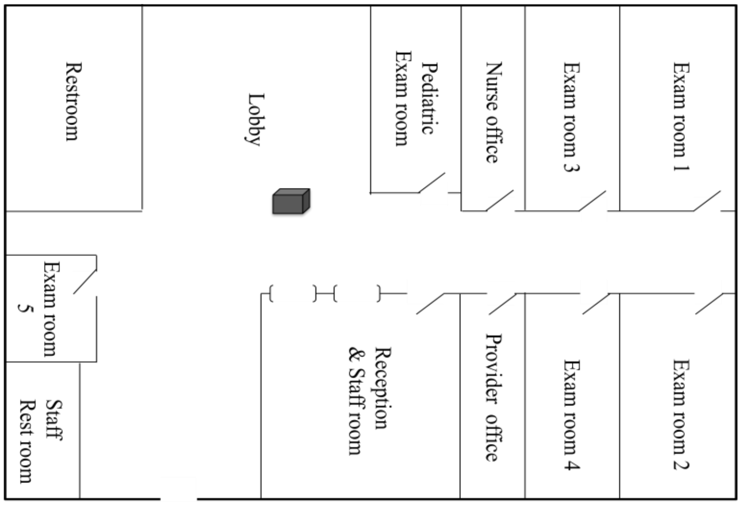


Figure 3.1: Layout and observer location at the studied family medicine practice

Once patients enter the practice, they proceed to the reception desk to notify their arrivals to the receptionist. They wait in the lobby until a nurse calls to examine the patient. After the exam, the nurse flips a flag indicating that the patient can now be seen by the provider. The patient waits in the exam room until a provider is available. Before seeing the patient, the provider flips another flag; and once the appointment has concluded, she/he flips down all flags. These flags are visible from the lobby, where the observer is present, and allow for the unobtrusive collection of the following time stamps: 1) wait time in the lobby; 2) service time with a nurse; 3) wait time in the exam room; 4) service time with a provider; and 5) total time of patient visits. In our wait and service time observations, we accounted for the fact that a nurse and/or a provider sometimes returned to visit the patient in the exam room even after the conclusion of the initial service time.

3.2.2 Summary of Appointment Mix

The data was collected using a time-motion study conducted on nine work days in summer and fall of 2011. A descriptive summary of the data is provided in Table 3.1.

Table 3.1: Appointment Mix

	Number	Percentage
Total number of scheduled appointments *	420	
Total number of observed appointments	364	
No-shows	13	3%
Cancellations and Reschedules	19	5%
Prescheduled appointments	317	75%
Same-day appointments	103	25%
30-min. appointments	121	33%
15-min. appointments	243	67%

* Total number of scheduled patients includes patients who were scheduled during the course of the study, including no-shows, cancellations and reschedules, and those who only received nurse care.

All told, 420 patients were scheduled for appointments during the course of the research study. We observed 364 patients from beginning to end. The total numbers of patients who were scheduled and patients who were actually observed are different because some patients saw only a nurse to receive a flu shot or simple treatment. The practice has fairly low percentage of no-shows, cancellations and reschedules. Literature study of Cayirli and Veral (2003) reports the range of no-show rates from 5 to 30 percent. Thus, 3 percent of no-show rates is significantly low. The pre-scheduled appointments are three-quarters of the total number of scheduled appointments while the same-day appointments are one-quarter of them.

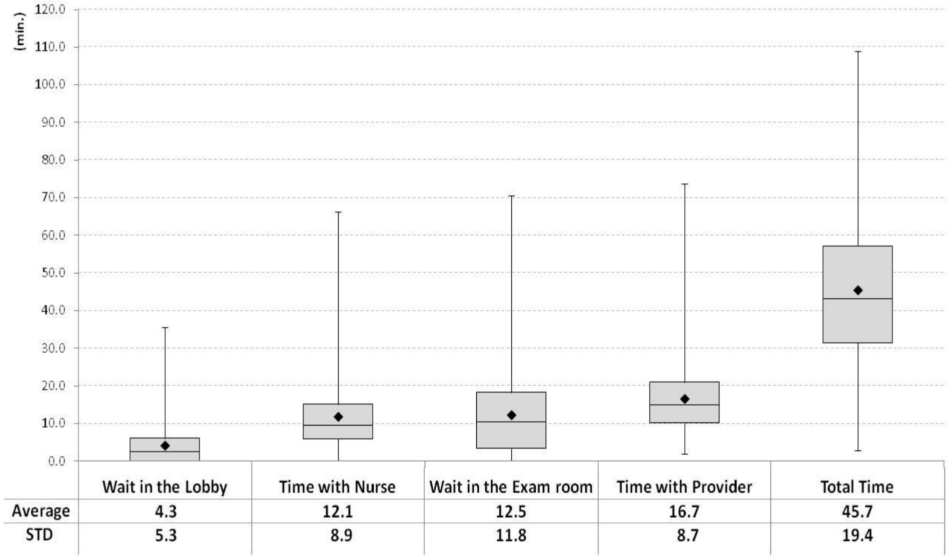
The practice schedules patients in 15-min. increments and reserves either a 15- or 30-min. appointment slot for each patient depending on their predicted complexity. 30-min. appointments are one third out of total number of observed patients whereas 15-min. appointments are two third. Same-day appointments are allocated a 15-min. slot, and occasionally double-booked.

3.3 Data Analysis

We collect data of the patient flow; wait time in the lobby, service time with nurse, wait time in the exam room, service time with provider, and total time patient spent in the practice.

This assists to understand where the high variability occurs and what causes this high variability in the patient flow.

3.3.1 Summary of Patient Flow Measures



	Wait in the Lobby	Time with Nurse	Wait in the Exam room	Time with Provider	Total Time
Average	4.3	12.1	12.5	16.7	45.7
STD	5.3	8.9	11.8	8.7	19.4

Figure 3.2: Box plot of Practice Performance (min.)

Figure 3.2 presents a box plot, the average and the standard deviation of each indicator of patient flow. On average, patients wait about 4 min. in the lobby, spend 12 min. with a nurse, wait 13 min. in the exam room, and finally spend 17 min. with a provider. In total, patients spend 46 min. at the practice. Although at first glance each of the performance indicators appears satisfactory, there is, in fact, significant variability among the time indicators. In particular, wait time in the exam room has a high standard deviation. Distributions of each recorded measure are shown in Figure 3.3.

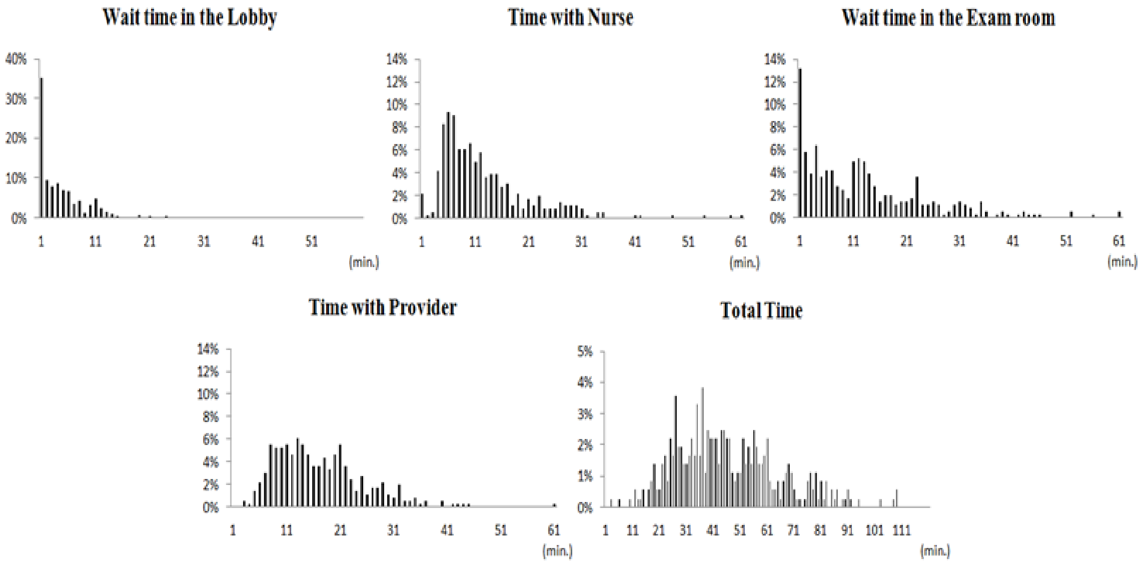


Figure 3.3: Distribution of Patient Flow

Patients must have the necessary amount of service time with nurses and providers. As shown in Figure 3.3, however, service time with a medical team (nurses and providers) is highly variable; furthermore, the service time distributions for both nurses and providers are skewed to the right. The variability is understandable as these distributions aggregate both 15-min. and 30-min. appointments and a great variety of patient needs. The data shows that 15-min. appointments often exceeded their anticipated durations; in fact, 42% of 15-min. appointments (whether prescheduled or same-day) took longer than 15 min. with providers, and 24% of them exceeded 20 min.

The histograms of both lobby and exam wait times resemble the Exponential distribution. In the lobby, 29% of patients had 0 wait time, and 68% of patients waited fewer than 5 min. In the exam room, we observed that 52% of patients waited more than 10 min. and 10% waited 30 min. or more. These relatively long wait times are of particular concern to the practice, as they erode patient satisfaction. Certainly, waiting in the exam room increases patient discomfort and anxiety and is not convenient. The distribution of total time that patients spend at the practice, which aggregates all the time measures, understandably looks less skewed.

3.3.2 Provider and Nurse Service time by Patient Condition

We found that service times vary significantly depending on the nature of the patient's ailment. Further, as shown in Figure 3.4, different patient conditions require different amounts of service time with nurses and providers. Notice that while service time with providers is typically higher, nurse times are non-trivial and in some cases higher. For instance, patients scheduled for well child check-ups or sore throat visits require longer time with nurses because specific medical tests need to be performed. Therefore, coordinating nurse and provider times for the various patient types in the schedule is essential if exam room waiting is to be reduced.

*TN: time with nurses TP: time with providers

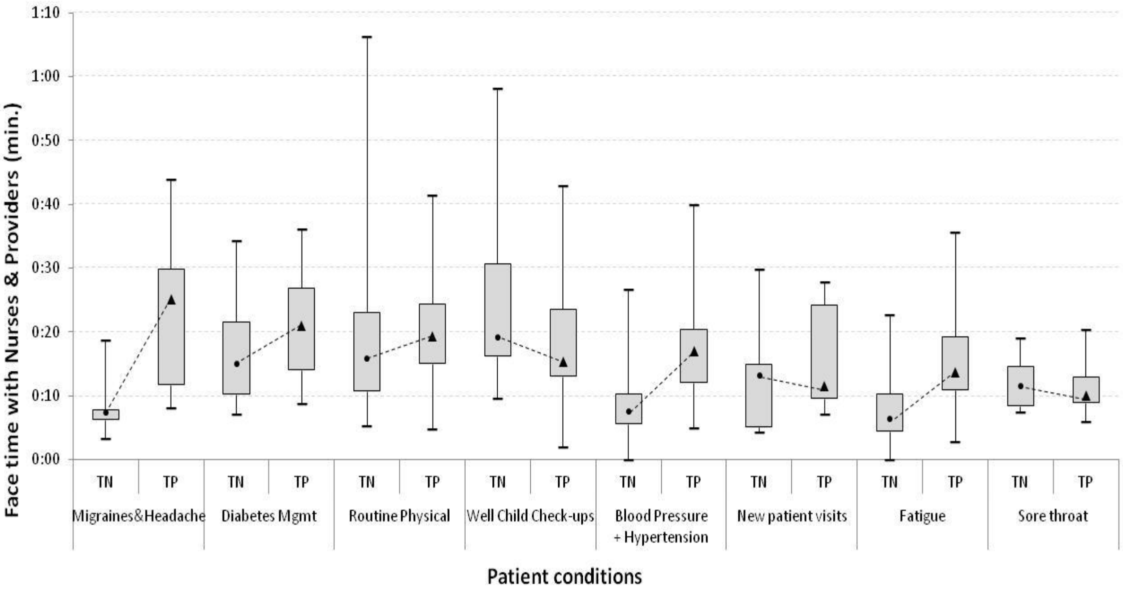


Figure 3.4: Box Plots of Service time with Nurses and Providers by Patient Conditions

3.3.3 Other Variability Factors

In this section, we study more factors which cause variability at the practice: service time by providers, wait times by hours, and practice utilization by days.

We have found that patient time spent with nurse is non-trivial. However, the service time with provider is typically higher and causes the bottle neck in the patient flow. Therefore, we closely look at service time of each provider, shown in Figure 3.5.

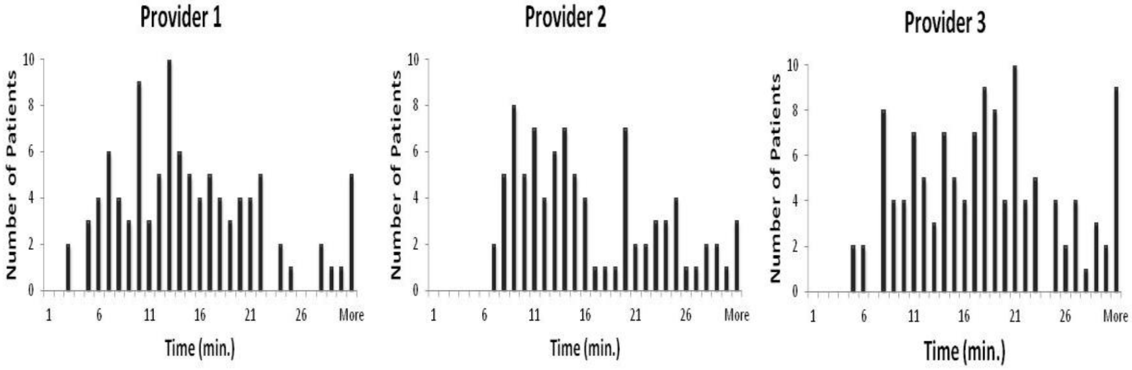


Figure 3.5: Variability of Service Time by Providers

The service time distribution changes significantly from provider to provider. Both the average and standard deviation of service time differ by providers: average service time is 14.7 min. (standard deviation: 7.5 min.) with provider 1; 16.0 min. (standard deviation: 7.3 min.) with provider 2; and 17.8 min. (standard deviation: 7.7 min.) with provider 3, respectively. The provider 1 and 2 take less average time with patients compared to provider 3 since they have known many of their patients more than 10-years. We also find that the panels of the provider 3 have not only the longest service time but also the longest wait time in the exam room. Indeed, her reaction is modified after she reviewed a report: her average service time is similar before-and-after, but standard deviation is reduced by two minutes. The significant adjustment is six minutes decrease in 90th percentile. Furthermore, average wait of her panels is decreased by seven minutes in attempting to do indirect service work in empty appointment slots due to no-shows or cancellations/reschedules, rather than using her idle time. For examples, she tried to take care of indirect service work such as email, phone calls, and paper work in the empty appointment slots because of no-shows or cancellations/rescheduled appointments. When the provider tends to use idle time for indirect work, she/her may overuse idle time causing delay in seeing the next patient, which significantly affects wait time accumulation over the day.

Next, we observe high variability of wait times by hours. This variability of wait times (both in the lobby and the exam room) is compounded by the fact that delays accumulate over the day.

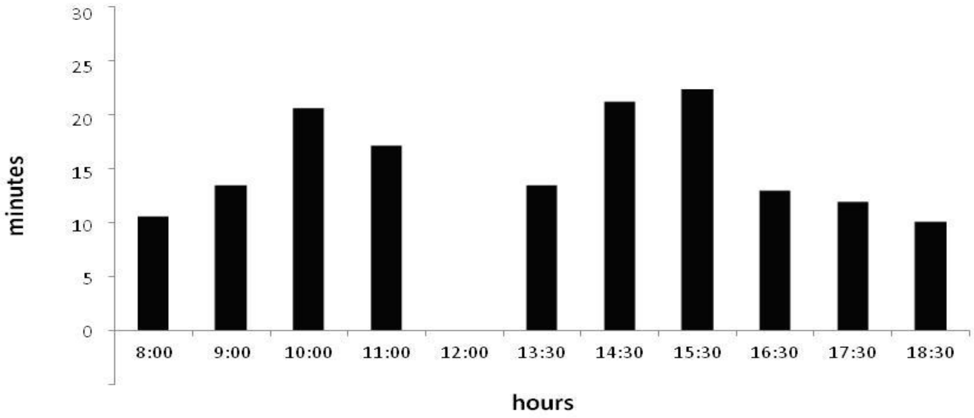


Figure 3.6: Wait Times by Hours

As displayed in Figure 3.6, the morning session decouples from the afternoon since the practice breaks for lunch. Wait times grow over time and then decrease at the end of the morning and afternoon sessions since same-day appointments, which require a shorter time from both nurse and provider, are more frequent scheduled toward to the end of the session. This seems to allow the providers to catch up with their delayed work. In fact, wait times are down towards to very end of the morning or very early of the afternoon. It clearly explains the relationship of wait times and same-day appointments.

The last variability we have analyzed is the practice utilization which varies day-to-day. Figure 3.7 illustrates the daily practice utilization, calculated as the proportion of the available 15 min. slots that were actually used up by both pre-scheduled and same-day appointments. It also displays the prescheduled utilization of the daily slots at the beginning of the day (gray line), that is, the proportion of total appointment slots that had been assigned to prescheduled patients before same-day cancellations and no-shows occur.

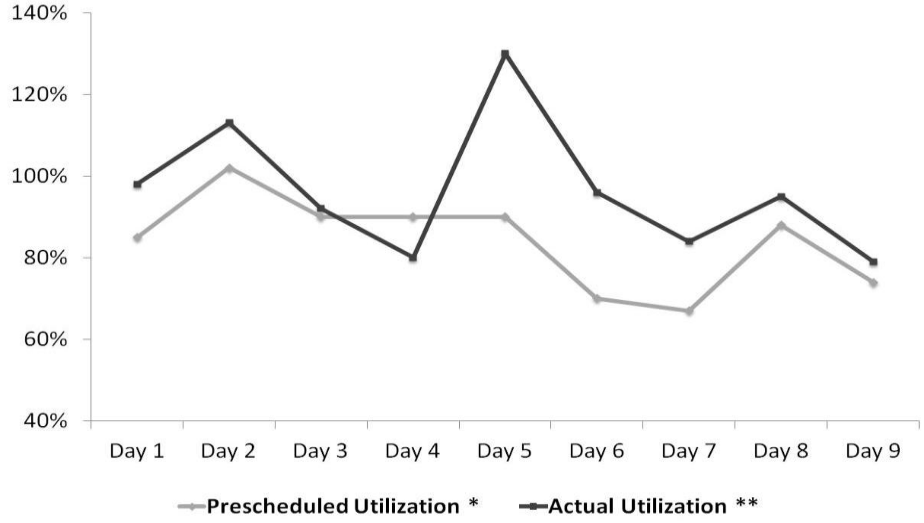


Figure 3.7: Daily Practice Utilization

* Pre-scheduled Utilization: number of prescheduled appointment slots/number of available slots
** Actual Utilization: (number of pre-scheduled plus same-day minus (no-shows plus cancellations/rescheduled appointment slots)) divided by number of available slots

On average of nine days, 84% of the available slots are filled by pre-scheduled appointments. The black line is actual utilization of the practice on that day including same-day appointments and excluding no-shows and cancellations/rescheduled appointments. Once you account for the same-day appointments, utilization has been increased by 96% on average. In day 2, utilization appears to be over 100% for pre-scheduled appointments. This is because of family groups which were expected to take less time than the corresponding slots for each of the family members. In day 4, the actual utilization is below pre-scheduled utilization since no-shows outpace the requests for same-day appointment slots. Therefore, the practice utilization experiences significant day-to-day variability.

3.4 Improved Appointment Classification

The practice currently schedules two types of appointments, 15-min. and 30-min., in 15-min. slots. In the provider's schedule, a 30-min. appointment takes up two consecutive 15-min. slots. The 30-min. appointments consist of routine physical exams; well child check-ups; diabetes and chronic condition management; new patient visits; procedures; and migraines and headaches. All other appointments – including same-day requests – are scheduled as 15-min. appointments.

Table 3.2: Service time with Nurse and Provider by Patient Type under Current Patient Classification (min.)

Current Classification	Medical Staff	Mean	Standard deviation	*T-test p-value*
30-min.	Nurse	18.5	10.7	*0.000*
	Provider	19.1	7.9	*0.000*
15-min.	Nurse	9.0	5.7	*Ref.*
	Provider	15.6	8.8	*Ref.*

As Table 3.2 shows, the mean and standard deviation of service times of the patients we observed in our time-study are indeed statistically different for these two types of appointments considered by the practice.

Our empirical study suggests that we can further refine the classification of appointments. Based on time requirements, we propose classifying patients into three easy-to-identify groups: prescheduled 30-min. appointments of high complexity (*HC*), which consist of the six conditions mentioned above; prescheduled 15-min. appointments, which include conditions of relatively low complexity (*LC*); and appointments scheduled on short notice, which consist of urgent, same-day appointments (*SD*).

Table 3.3: Service time with Nurse and Provider by Patient Type under New Patient Classification (min.)

New Classification	Medical Staff	Mean	Standard deviations	*T-test p-value*
HC (High Complexity)	Nurse	17.8	10.7	*0.000*
	Provider	19.5	8.2	*0.005*
LC (Low Complexity)	Nurse	8.5	5.1	*Ref.*
	Provider	16.6	9.0	*Ref.*
SD (Same-day)	Nurse	9.5	6.1	*0.239*
	Provider	12.7	7.0	*0.000*

Table 3.3 shows that the differences among the three groups we propose are indeed statistically significant. The practice currently lumps all LC appointments and SD appointments into the same 15-min. appointment category. On average, however, the LC patient spends three additional minutes compared to the SD patient while still remaining clearly distinct from the HC patient. Note that LC and SD patients will still be scheduled in 15-min. slots. However, if a number of LC appointments are scheduled in succession without an open slot (slack), wait times are more likely to accumulate than when the same number of SD appointments is scheduled in succession. This subtle point has implications for the scheduling questions we study in the next chapter.

The new classification makes also intuitive sense from the point of view of the practice since the SD appointments become known only as the work day progresses, whereas all prescheduled patients, whether in the HC or the LC categories, are known at the beginning of the work day. In addition, SD patients' calls have to be fulfilled at a short notice. The short notice here refers to a few hours or half a day (patients who need immediate care do not fall in this category and are typically directed to an emergency room). Thus, it is important to have slots available towards the end of a session. This also helps reduce the risk of unfilled slots (or double-booking) by allowing the practice to provide a patient who calls in at, say, 8 am with a late morning or early afternoon slot. Indeed, the practice we work with has followed this policy based on our recommendation. See Balasubramanian *et al*. (2013) for a detailed analysis. Figure 3.8 shows the average number of prescheduled and same-day appointments by time of day for nine work days with two providers working in parallel. SD appointments are mostly scheduled late in the morning. In the afternoon session, however, SD appointments can be more evenly distributed; yet, for the same reasons discussed above, some SD appointments are made available later in the afternoon.

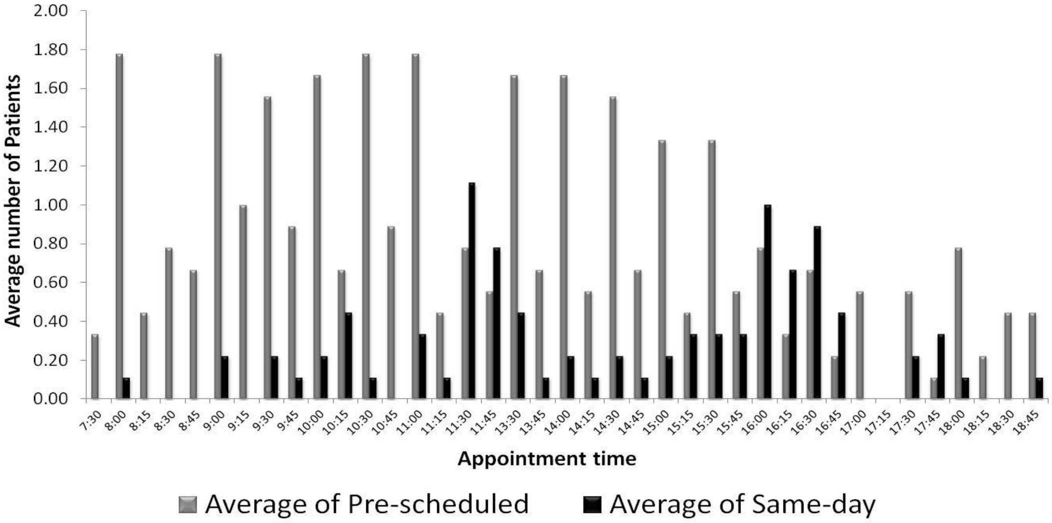


Figure 3.8: Pre-scheduled vs. Same-day by Time of Day

3.5 Appointment Classification with National Data

The nine work days chosen for the time-study may not be entirely representative of the volume and mix of patients served. Our main goal, though, was to capture the distribution and variability of nurse and provider service times for different patient types. Comprehensive self-reported data on patient conditions and provider service times does exist in the National Ambulatory Medical Care Surveys (NAMCS) conducted each year. We analyze NAMCS collected in 2010 to check whether the insights of our new patient classification apply to such a nationally representative data set. In particular, patient conditions in HC category which are distinguished from other appointments in LC and SD. Although there is no service time of same day appointments, 94% physicians accepts same day appointment and 81% physicians said their practice set aside time for same day appointments, out of 11,673 appointments which explains below.

Out of 31,229 appointments in NAMCS 2010, we mainly focus on the following aspects: primary care practice physicians including general/family practice, internal medicine, and pediatrics, out of sixteen specialty categories; and the category of the reason to visit the practice rather than the diagnosis which can be also a good measure to classify patient; however, when the scheduler books an appointment, it is based on the patient reasons. In all, there are 11,673 appointments.

First, under the category of new/established patients, we extract only the new patient which is the separate appointment type regardless of patient conditions in HC category. Excluding new patient appointments, then, we found 511 different reasons for visits. Table 3.4 shows the reasons for visits which include over 100 appointments, sorted by average service time with provider.

Table 3.4: Reasons for visits over 100 appointments (sorted by average service time)

Reasons for visits	Number of appointments	Average service time (min.)	STD (min.)
General medical examination (CPE)	1464	22.3	11.3
New Patient (First visit)	1221	22.0	11.5
Diabetes mellitus (Diabetes)	230	21.3	12.0
Headache, pain in head (Headache)	109	20.2	9.7
Hypertension	202	20.1	8.5
Well baby examination (WCC)	385	20.0	7.9
Back pain, ache, soreness, discomfort	162	19.7	10.2
Abdominal pain, cramps, spasms, NOS	100	19.3	8.9
Medication, other and unspecified kinds	381	19.3	8.8
Progress visit, NOS	783	19.1	9.5
For other and unspecified test results	278	18.4	8.6
Fever	227	17.4	12.1
Head cold, upper respiratory infectio...	161	17.0	8.0
Skin rash	169	16.9	6.3
Nasal congestion	124	16.7	6.6
Cough	554	16.6	7.1
Earache, pain	167	16.6	6.8
Throat soreness	208	16.4	8.5

* NOS: Not Otherwise Specified

In Table 3.4, bolds are the patient conditions/appointment types in our HC category. In particular, we suggested that headache condition, which was originally in 15-min. appointment bucket, needs to be considered in HC to the practice we work with; this has been already implemented. NAMCS 2010 data analysis also proves that headache condition requires long service time and also that all conditions we have included in HC need high service time with providers. Hypertension condition was considered in our LC category since all patients who have this condition were follow-up appointments.

3.6 Conclusion

In summary, our empirical study sheds light on the scheduling challenges facing family care practices. Primary care practices include multiple factors causing high variability: patient conditions, appointment types, random service times, multiple resources, and multiple service steps in the patient flow. This variability leads to patient long wait time and less satisfaction; as a result, significantly reduces the utilization of the practice. Thus, we analyze the data collected from the family medicine practice in Massachusetts for nine work days.

We mainly focus on factors leading to high variability at the practice. We first observe that service time in each stage of the patient flow is significantly variable. In particular, we find that wait time in the exam room is three times longer than wait time in the lobby and has the highest standard deviations among stages in the patient flow. It is due to lack of coordination of service steps since the wait time in the exam room is between the nurse and provider steps. Next, we analyze service time distributions with nurse and provider for specific conditions/appointment types. We find that although service time with provider has the significant impact on the practice utilization, service time with nurse is non-trivial. Some of patient conditions require more nurse service times than provider service time. In addition, we look at three more factors causing high variability: service time considerably varies from provider to provider; wait time accumulates over time; and the practice utilization notably changes day by day. All these factors significantly influence cumulative delays which highly affect long patient wait time. In order to reduce wait time, particularly cutting-off the probability of high waits, nurse and provider steps need to be coordinated effectively.

Based on data and statistical analysis, we suggest the new easy-to-identify groups: prescheduled high complexity (HC) appointments, prescheduled low complexity (LC); and same-day appointments (SD). This new patient classification scheme is meaningful and broadly applicable in the practices since patient conditions we analyze are common in many primary care practices. In addition, the analysis of data set from NAMCS 2010 proves patient conditions in HC type on the national level.

In the next two chapters, we study the scheduling problems with a stochastic model using two service steps – nurse and provider and new patient classification to provide robust scheduling guidelines. The objective of the models is to minimize wait times while keeping the bottleneck resource busy. In Chapter 4, we study a single-provider primary care practice where one nurse and one provider see patients. In Chapter 5, multiple human resources are included in the model.

CHAPTER 4

SINGLE-PROVIDER PRIMARY CARE PRACTICE

4.1 Introduction

In the previous chapter, we study different factors which cause high variability in the primary care practices: service time of each component in the patient flow, service time by patient conditions/appointment types, each provider service time, wait time by hours, and daily practice utilization. This variability significantly affects on accumulated patient wait time over the day and unnecessary provider idle time; in turn, negatively influence the practice utilization. To balance between patient wait and provider idle time, nurse and provider steps needs to be coordinated efficiently.

In this chapter, we model a two stage stochastic integer program that schedules and sequences patient appointments for a single-provider primary care practice where a single nurse and provider see patients. The objective of the model is to minimize a weighted measure of provider idle time and patient wait time. Key features of the model include: patient classification to accommodate different chronic and acute conditions seen in the practice; adequate coordination of patient time with a nurse and a provider; and strategies for introducing open slot (slack) in the schedule to counter the effects of variability in service time with providers and nurses. While outpatient scheduling is a well studied topic [see Cayirli and Veral (2003) and Gupta and Denton (2008) for a review], the current paper brings together the disparate elements mentioned above – which have typically been studied only in isolation – into a single, tractable optimization framework. For example, many stochastic optimization approaches schedule only the provider [Robinson and Chen (2003) and Denton and Gupta (2003)], but they do not coordinate patient service time with the nurse or consider the diversity of patient conditions. We use the model to create broader guidelines that can help practices carry out more effective scheduling while staying sensitive to current protocols and operational constraints. We also compare the proposed schedules to actual schedules used in practice.

To best fit our collaborating practice's needs, we fix certain parameters. For instance, we use a particular objective function coefficient since it provides the right balance between the provider idle time and patient wait time for the practice. Also, we consider zero no-show rates since the practice we work with has only 3% no-show rates. Thus, we also study an impact of different coefficients of the objective function and the effect of various no-show rates.

The rest of the paper is structured as follows. In Section 4.2, we describe a two stage stochastic integer program, a scheduling and sequencing model. In Section 4.3, we use this mathematical model to address five focused questions relevant for the practice. In Section 4.4, we study the sensitivity analysis of objective function coefficients and no-show rates. We summarize our conclusions and implications for practice in Section 4.5.

4.2 Integer Programming Formulation

4.2.1 Model Description

Based on findings from the data analysis in Chapter 3, we now study a two-stage stochastic integer program (SIP) for assigning multiple patient types to appointment slots in a *session*. A session refers to a block of time (typically a few hours) either in the morning or afternoon. The morning and afternoon sessions can be decoupled, and their schedules studied independently, since there typically is a break for lunch.

The objective of the SIP is to minimize a weighted measure of provider idle time and patient wait time in the session. Wait time in our model has two components: wait time in the lobby (until the nurse calls), and wait time in the exam room after the nurse exam (until the provider is ready). However, we simply consider total patient wait times as the measure of performance in the computational results. We assume that a provider's calendar for a morning or afternoon session consists of contiguous appointment slots, each having a fixed, predetermined length (15-min. in our case study). Each patient spends an uncertain amount of time with first the nurse and then the provider. The distribution of these service times depends on the type of patient being scheduled. The number of patients of each type to be scheduled is known beforehand. While this is not the case in reality, we demonstrate in our computational results that the guidelines we develop using our model are robust to changes in the mix of patients scheduled. We also assume that the patients arrive punctually and are not called by the nurse before their scheduled appointment times.

The first-stage decisions of the SIP involve both the *sequence* in which the patient types are scheduled, and the *appointment times* of each patient. Because the slots in our case-study are 15 minutes long, the appointment times are always in 15-min. increments. For any feasible first-stage decisions (which determine the schedule for the session), nurse and provider service times are realized in the second-stage, resulting in idle time for the provider and wait time for the patients.

We create 1000 scenarios or realizations by sampling randomly from the empirical service time distributions obtained from the field study. We use the sample average approximation method (see Kleywegt *et al*., 2002).

The two-stage stochastic integer program is described formally below.

Sets

I	Set of patients to be scheduled in the session, indexed by *i* = 1,…, *I*
S	Set of scenarios, indexed by *s* = 1,…, *S*

Parameters

α	Weight for idle time
β	Weight for wait time
N_{HC}	Number of patients of type HC to be scheduled
N_{LC}	Number of patients of type LC to be scheduled
N_{SD}	Number of patients of type SD to be scheduled
$\tau_{i,s}^{n,HC}$	Service time with a nurse for patient *i*, if of type *HC*, under scenario *s*
$\tau_{i,s}^{n,LC}$	Service time with a nurse for patient *i* if of type *LC*, under scenario *s*
$\tau_{i,s}^{n,SD}$	Service time with a nurse for patient *i* if of type *SD*, under scenario *s*
$\tau_{i,s}^{p,HC}$	Service time with a provider for patient *i* if of type *HC*, under scenario *s*
$\tau_{i,s}^{p,LC}$	Service time with a provider for patient *i* if of type *LC*, under scenario *s*
$\tau_{i,s}^{p,SD}$	Service time with a provider for patient *i* if of type *SD*, under scenario *s*

Variables

$\tau_{i,s}^{n}$	Service time of patient *i* with a nurse under scenario *s*
$\tau_{i,s}^{p}$	Service time of patient *i* with a provider under scenario *s*
$y_{i,s}^{start}$	Start time of patient *i* with a nurse under scenario *s*
$y_{i,s}^{finish}$	Finish time of patient *i* with a nurse under scenario *s*
$z_{i,s}^{start}$	Start time of patient *i* with a provider under scenario *s*
$z_{i,s}^{finish}$	Finish time of patient *i* with a provider under scenario *s*
$A_i \in \{0,1\}$	1 if patient *i* is HC, 0 otherwise
$B_i \in \{0,1\}$	1 if patient *i* is LC, 0 otherwise
$C_i \in \{0,1\}$	1 if patient *i* is SD, 0 otherwise
X_i	Appointment slot for patient *i*, X_i *in 0,1,2,…,15 for a 4-hour session*

The problem is modeled as the following integer program.

$$Min. \qquad \frac{1}{S}\left(\alpha\left[\sum_{s}\sum_{i=1}^{n}\left(z_{i,s}^{start}-z_{i-1,s}^{finish}\right)\right]\right.$$
$$\left.+\,\beta\left[\sum_{s}\sum_{i=1}^{n}\left(y_{i,s}^{start}-15X_i\right)+\left(z_{i,s}^{start}-y_{i,s}^{finish}\right)\right]\right) \tag{1}$$

Subject to.

$$y_{0,s}^{finish}=0, \qquad \forall s\in S \tag{2}$$

$$z_{0,s}^{finish}=0, \qquad \forall s\in S \tag{3}$$

$$X_1=0 \tag{4}$$

$$\tau_{i,s}^{n}=\tau_{i,s}^{n,HC}\times A_i+\tau_{i,s}^{n,LC}\times B_i+\tau_{i,s}^{n,SD}\times C_i, \qquad \forall i\in I, s\in S \tag{5}$$

$$\tau_{i,s}^{p}=\tau_{i,s}^{p,HC}\times A_i+\tau_{i,s}^{p,LC}\times B_i+\tau_{i,s}^{p,SD}\times C_i, \qquad \forall i\in I, s\in S \tag{6}$$

$$y_{i,s}^{start}\geq 15X_i, \qquad \forall i\in I, s\in S \tag{7}$$

$$y_{i,s}^{start}\geq y_{i-1,s}^{finish}, \qquad \forall i\in I, s\in S \tag{8}$$

$$y_{i,s}^{finish}=y_{i,s}^{start}+\tau_{i,s}^{nurse}, \qquad \forall i\in I, s\in S \tag{9}$$

$$z_{i,s}^{start}\geq y_{i,s}^{finish}, \qquad \forall i\in I, s\in S \tag{10}$$

$$z_{i,s}^{start}\geq z_{i-1,s}^{finish}, \qquad \forall i\in I, s\in S \tag{11}$$

$$z_{i,s}^{finish}=z_{i,s}^{start}+\tau_{i,s}^{PCP} \qquad \forall i\in I, s\in S \tag{12}$$

$$\textstyle\sum_{i=1}^{n}A_i=\ N_{HC} \tag{13}$$

$$\textstyle\sum_{i=1}^{n}B_i=\ N_{LC} \tag{14}$$

$$\textstyle\sum_{i=1}^{n}C_i=N_{SD} \tag{15}$$

$$A_i+B_i+C_i=1 \tag{16}$$

$$A,B,C\in\{0,1\}; \quad y^{start},y^{finish},z^{start},z^{start}\geq 0; \quad X\ int.$$

The objective function (1) minimizes the weighted sum of idle time and wait time over all scenarios. Note that computation of the provider's idle time is based on the difference between the start time of patient *i* and finish time of patient *i*-1. For the patients' wait time in the lobby, we look at the difference between the start time of patient *i* with a nurse and the appointment time. For the wait time in the exam room, we take the difference from the start time of patient *i* with a provider minus the finish time of patient *i* with a nurse. Constraints (2-4) initialize the finish time of the 0th patient with a nurse and a provider to 0, and the first patient start time with nurse to the beginning of the session, in every scenario. Constraints (5-6) ensure that proper service times are used given the patient type. Constraints (7-9) keep track of start time and finish time of patient *i* with a nurse, as well as set the appointment time given to patient *i*. Constraints (10-12) track start time and finish time of patient *i* with a provider. Constraints (13-15) ensure that the desired number of patients of each type is scheduled in the session. Constraint (16) enforces that only one patient type can be scheduled on the particular slot.

As a benchmark, we also consider a *deterministic* integer program (DIP) by assuming that nurse and provider service times take on their respective average values and have no variability. We use the CPLEX Solver Version 12.4 to solve the SIP and the DIP.

Notice that, for a predetermined patient sequence, the SIP can also be used to optimally determine the appointment times of each patient. The spacing between the scheduled arrivals of two patients determines *slack* in the schedule. Slack prevents the accumulation of patient waiting. Given that sequences can vary from day to day based on patient requests and time-of-day preferences, it is important to derive robust guidelines on where slack should be strategically positioned in the schedule.

4.2.2 Calibrating Weights in the Objective Function

How much should a unit of provider idle time be valued against a unit of patient waiting? This is a recurring issue in all appointment scheduling research (see Robinson and Chen, 2011 for a detailed discussion). We looked at five afternoon schedules from our time-study. The mix of patients varied from one afternoon to another. We compared the schedule used in practice with the schedules generated by the SIP and the DIP. While we tested a wide range of weights, we narrowed down our search to cases where a provider's idle time is equal to or higher than that of patient waiting. This makes intuitive sense since idle time is experienced by a single person while wait time accumulates across all scheduled patients. In addition, high idle time is unacceptable in the primary care practice as it would make it financially unviable.

The results are shown in Figure 4.1. As an example, DIP 0.8:0.2 implies that the weight on provider idle time is 0.8 and on patient waiting is 0.2 in the DIP.

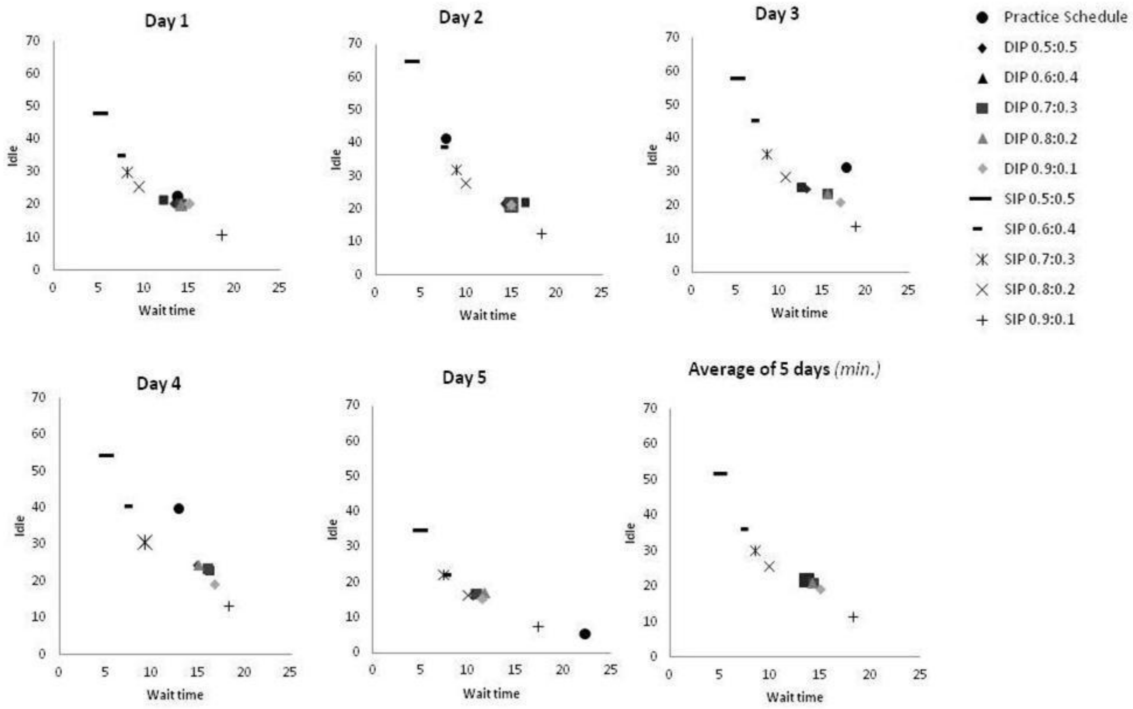


Figure 4.1: Expected performance of the schedules using different weight combinations (min.)

We observe that the DIP is mostly insensitive to the weights. This is understandable since the DIP does not capture variability and therefore grossly underestimates how wait times accumulate as the day progresses. The SIP, which considers variability, exhibits greater sensitivity toward changes in weights. We notice that the SIP 0.5:0.5 results in much higher idle time than other weight combinations; on average, idle time of the SIP 0.5:0.5 is more than 50 min. in a session with only 10 patients, which would be unacceptable in a primary practice. We also find that the SIP 0.7:0.3 schedules provide low wait times but more idle time than acceptable for the practice, while the SIP 0.9:0.1 schedules provide very little slack and thereby increase patient waiting beyond the desired levels. The practice needs to strike a careful balance between inducing high levels of provider idle time by adding too much slack in the schedule, and observing lengthy patient waits when not adding enough. Fortunately, the SIP 0.8:0.2 schedules tested provide the right balance between these two cases.

These observations are further illustrated in the schedules generated by the SIP when all patients are of the same type (the homogeneous patient case). Notice in Figure 4.2 (a) and (b) that in the 0.8:0.2 schedule, the number of empty slots (slack) is exactly one fewer and one more than the 0.7:0.3 and 0.9:0.1 schedules, thus striking a balance. Consequently, we will use the 0.8:0.2 weight combination in the remainder of our computational experiments.

# patient number

(a) 15-min. appointment type

0.7:0.3

#	Time	15-min.
1	0:00	▨
2	0:15	▨
3	0:30	▨
4	0:45	▨
5	1:00	▨
	1:15	
6	1:30	▨
7	1:45	▨
8	2:00	▨
9	2:15	▨
	2:30	
10	2:45	▨
11	3:00	▨
12	3:15	▨
13	3:30	▨
	3:45	
14	4:00	▨
15	4:15	▨
16	4:30	▨

0.8:0.2

#	Time	15-min.
1	0:00	▨
2	0:15	▨
3	0:30	▨
4	0:45	▨
5	1:00	▨
6	1:15	▨
	1:30	
7	1:45	▨
8	2:00	▨
9	2:15	▨
10	2:30	▨
11	2:45	▨
	3:00	
12	3:15	▨
13	3:30	▨
14	3:45	▨
15	4:00	▨
16	4:15	▨

0.9:0.1

#	Time	15-min.
1|2	**0:00**	2
3	0:15	▨
4	0:30	▨
5	0:45	▨
6	1:00	▨
	1:15	
7	1:30	▨
8	1:45	▨
9	2:00	▨
10	2:15	▨
11	2:30	▨
12	2:45	▨
13	3:00	▨
14	3:15	▨
15	3:30	▨
16	3:45	▨

(b) 30-min. appointment type

0.7:0.3

#	Time	30-min.
1	0:00	■
2	0:15	■
	0:30	
3	0:45	■
4	1:00	■
	1:15	
5	1:30	■
	1:45	
6	2:00	■
7	2:15	■
	2:30	
8	2:45	■

0.8:0.2

#	Time	30-min.
1	0:00	■
2	0:15	■
3	0:30	■
	0:45	
4	1:00	■
5	1:15	■
	1:30	
6	1:45	■
7	2:00	■
	2:15	
8	2:30	■

0.9:0.1

#	Time	30-min.
1	0:00	■
2	0:15	■
3	0:30	■
4	0:45	■
	1:00	
5	1:15	■
6	1:30	■
	1:45	
7	2:00	■
8	2:15	■

Figure 4.2: Optimal SIP schedules for homogeneous patients under different weight combinations

4.3 Computational Results

Our computational results consist of five distinct parts. To develop intuition, we first consider the structure of optimal schedules when all patients are of the same type. Second, we look at optimal sequences and appointment times under the DIP and SIP, when a mix of patient types has to be scheduled in a practice session. In practice, however, sequences need to be flexible so as to accommodate patient preferences and keep the practice financially viable. Hence, in the third part, we look at a variety of heuristic sequences that a practice might prefer, and how slack should be optimally introduced into these sequences to prevent the accumulation of wait time. In the fourth part, we conduct sensitivity on the length of appointment slots and its impact on a practice's performance. Finally, we compare schedules based solely on uncertain provider service time durations – a common practice in the appointment scheduling literature – to our model where both provider and nurse steps, with uncertain service times at both steps, are modeled.

4.3.1 Spacing Appointment Times for Homogeneous Patients

We start with the simplest case: How should appointment times be spaced throughout the session if all patients are homogeneous? We consider optimal appointment spacing for three homogeneous patient scenarios for a practice session: 1) 8 high-complexity (HC) patients; 2) 16 low-complexity (LC) patients; and 3) 16 same-day (SD) patients. The optimal schedules for these three scenarios are shown below:

8 HC patients

#	Time	HC
1	0:00	■
2	0:15	■
3	0:30	■
	0:45	
4	1:00	■
5	1:15	■
	1:30	
6	1:45	■
7	2:00	■
	2:15	
8	2:30	■

16 LC patients

#	Time	LC
1	0:00	▨
2	0:15	▨
3	0:30	▨
4	0:45	▨
5	1:00	▨
	1:15	
6	1:30	▨
7	1:45	▨
8	2:00	▨
9	2:15	▨
	2:30	
10	2:45	▨
11	3:00	▨
12	3:15	▨
	3:30	
13	3:45	▨
14	4:00	▨
15	4:15	▨
16	4:30	▨

16 SD patients

#	Time	SD
1	0:00	░
2	0:15	░
3	0:30	░
4	0:45	░
5	1:00	░
6	1:15	░
7	1:30	░
8	1:45	░
9	2:00	░
10	2:15	░
11	2:30	░
12	2:45	░
13	3:00	░
14	3:15	░
15	3:30	░
16	3:45	░

Figure 4.3: Spacing appointment times for homogeneous patients

Figure 4.3 shows that slack is necessary in schedules with HC and LC appointments, but not necessary when all are SD appointments. In the HC case, slack appears after two successive appointments, except at the beginning of the day, where it appears after three successive appointments. Figure 4.3 also illustrates that LC and SD are indeed different patient categories as we hypothesized: the former needs slack at regular intervals while the latter can do without slack. This is because SD appointments involve less variability in service times with provider (the bottleneck resource) than LC and HC appointments, as shown in Table 3.3 in Chapter 3. As a result, scheduling consecutive SD appointments without slack does not lead to any significant accumulation of patient waiting.

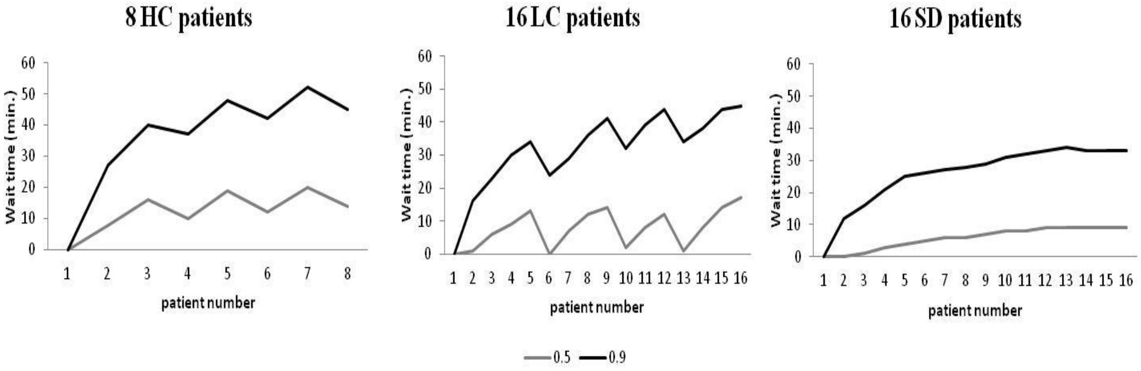


Figure 4.4: 50th and 90th percentiles of the patient wait times (min.)

Figure 4.4 shows the 50th and 90th percentiles of wait time by the patient number in the sequence. As the day progresses, wait time accumulates, but the introduction of slack brings it back down; hence the serrated shapes in the graphs for HC and LC cases. In the HC case, the wait time drops after the third, fifth, and seventh patients, due to slack. In the SD case, there is no slack, so we only have a gradual accumulation of wait time. Note that this accumulation is not as significant compared to the other two cases.

4.3.2 DIP vs. SIP

Consider Figure 4.5 (a), which shows the practice schedule for one of the five afternoon sessions observed. In total, 10 patients were scheduled for a provider: three HC patients; three LC patients; and four SD patients. Notice that there is slack after every HC patient. This is in fact the current scheduling policy: the practice uses two 15-min. slots in the calendar for every HC patient. The HC patient scheduled at 2 pm has until 2:30 pm; the HC scheduled at 2:45 pm has until 3:15 and so on.

The afternoon schedules created by the DIP and the SIP models (Figure 4.5 (b) and (c)) show that there is no need to book slack after every single HC appointment. We see that slack is typically scheduled after two successive HC appointments, consistent with what we found in the homogeneous HC patient case (see previous subsection). In the DIP, HC and LC appointments are double-booked at 0:45. The empty slot immediately after, at 1:00, provides the necessary time for the provider to see the second patient.

(a) Practice Schedule

#	Time	HC	LC	SD
1	0:00		▨	
2	0:15		▨	
3	0:30	■		
	0:45			
4	1:00		▨	
5	1:15	■		
	1:30			
6	1:45			░
7	2:00	■		
	2:15			
8	2:30			░
9	2:45			░
10	3:00			░

(b) DIP

#	Time	HC	LC	SD
1	0:00		▨	
2	0:15			░
3	0:30		▨	
4|5	0:45	5	4	
	1:00			
6	1:15	■		
7	1:30	■		
	1:45			
8	2:00			░
9	2:15			░
10	2:30			░

(c) SIP

#	Time	HC	LC	SD
1	0:00		▨	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00			░
6	1:15		▨	
7	1:30		▨	
8	1:45	■		
9	2:00	■		
	2:15			
10	2:30	■		

Figure 4.5: Schedules associated with one afternoon

The schedules we observe for the DIP and the SIP models consistently follow the features we see in the example shown in Figure 4.5. The DIP seems to be *dome-shaped* since it always schedules slack in the middle of the session which implies that the appointment interval lengths increase toward the middle and then decrease to the end of section. This slack in the middle session helps relieve the congestion that naturally accumulates over time. The sequence of the DIP locates HC appointments (with the longest average service time) towards the middle, LC towards the beginning, and most SDs towards the end. The SIP, meanwhile, follows, for the most part, the well known *SPT* (shortest processing time) rule. SPT translates to scheduling shortest mean appointments earlier in the schedule. The longer mean appointments, HC appointments, are scheduled towards the end, and the LC appointments are mostly clustered in the middle of the session.

In addition, these SPT sequences are fairly consistent with the sequences generated by the SIP under other different weight combinations that we discuss more details in Section 4.4.1.1. As the idle time weight increases, we observe less slack and more double booking. Also, none of the optimal SIP schedules under different weights start with HC appointments.

Table 4.1: Percent increase in the weighted sum of provider idle time and patient wait time (objective)

Average of 5 days	Practice schedule vs. DIP	Practice Schedule vs. SIP	VSS
Objective	16%	24%	10%

Table 4.1 shows percentage increase in the weighted sum of idle time and wait time. When averaged for five afternoon sessions, the practice's schedule is 24% worse in the objective value compared to the SIP and 16% worse than the DIP. The Value of the Stochastic Solution (VSS), the difference of performance between the SIP and the DIP, is 10%. In terms of total wait times (lobby + exam room), the SIP is 25% better than the practice schedule when averaged over the five afternoon sessions. Furthermore, the 90th percentile of waiting time in the exam room is 20% less in the SIP compared to the practice schedule. To further illustrate this point, Figure 4.6 displays the 90th percentiles of wait time by the patient number in the different schedules (Practice, DIP, and SIP) for one of the five afternoon sessions. While the wait time observed by the different patients in the sequence following the practice schedule is highly variable, wait times in the DIP and the SIP increase relatively smoothly. Wait time of the SIP is significantly below that of both the DIP and the practice schedule.

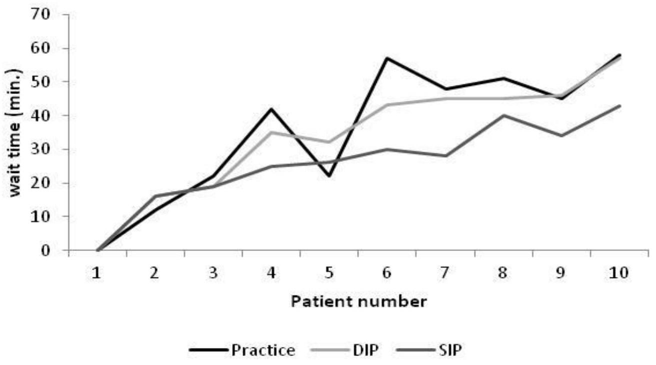


Figure 4.6: 90th percentiles of the patient wait times under three different schedules: Practice, DIP, and SIP

4.3.3 Heuristic Sequences

In two previous subsections, we have identified the structure of optimal schedules for homogeneous sets of patients, as well as a mix of patient appointment types. However, rigid adherence to sequences shown in Figure 4.5 (b) and (c) – based on the DIP and SIP – are not practical in reality. A dome-shape or SPT sequence is likely to be near optimal, but patients have time of day preferences; it is unrealistic to expect that all patients will be amenable to accepting slots only at a certain time of the day.

To be truly patient centered, therefore, we need to test schedules that provide sufficient flexibility for patients to have time-of-day options. For example, rather than having all HC appointments at the end or the middle of the day, the practice may like to make one HC appointment or LC appointment available each hour in a session.

On the other hand, the practice also has to stay financially viable. To do so, each provider in the practice we worked with needed to see at least three patients per hour. Hence, to satisfy the practical needs of patients and providers, we now explore a number of sequences that satisfy the 3-Appointments per Hour (3AH) criterion and provide a flexible schedule that allows an option for each type of patient class during every hour of the session. These sequences are shown in Figure 4.7.

(a) SD/LC/HC

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▓	
6	1:15	■		
	1:30			
7	1:45			░
8	2:00		▓	
9	2:15	■		
	2:30			
10	2:45			░

(b) LC/ SD/HC

#	Time	HC	LC	SD
1|2	0:00		1	2
3	0:15	■		
	0:30			
4	0:45		▓	
5	1:00			░
6	1:15	■		
	1:30			
7	1:45		▓	
8	2:00			░
9	2:15	■		
	2:30			
10	2:45			░

(c) SD/SD/HC followed by LC/LC/HC

#	Time	HC	LC	SD
1|2	0:00			1|2
3	0:15	■		
	0:30			
4	0:45			░
5	1:00			░
6	1:15	■		
	1:30			
7	1:45		▓	
8	2:00		▓	
9	2:15	■		
	2:30			
10	2:45		▓	

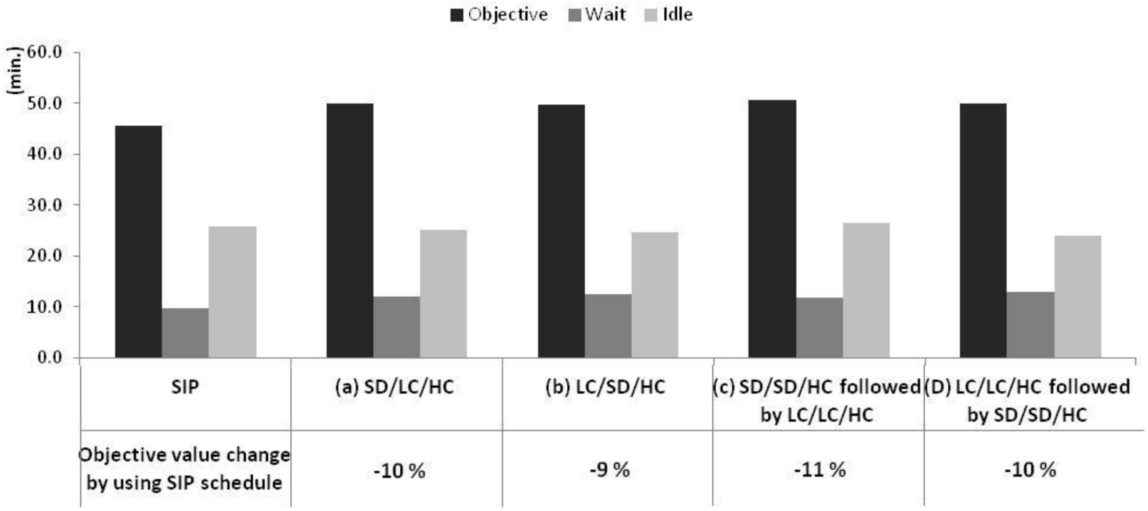
(d) LC/LC/HC followed by SD/SD/HC

#	Time	HC	LC	SD
1|2	0:00		1|2	
3	0:15	■		
	0:30			
4	0:45		▓	
5	1:00			░
6	1:15	■		
	1:30			
7	1:45			░
8	2:00			░
9	2:15	■		
	2:30			
10	2:45			░

Figure 4.7: 3-Appointments-per-Hour (3AH) schedules given optimal appointment times

In all four sequences, we have three appointments per hour; we call a block of three appointments, a *triad*. The triads are described by the sequence of patient types scheduled. For instance, an SD/LC/HC triad schedules a same-day patient, followed by a low-complexity prescheduled patient and then a high-complexity prescheduled patient. The last appointment of the triad in all four sequences is always a HC appointment. We also examined triads where an HC appointment comes first, but the performance is almost 50% worse than that of the triads where the HC comes last. Hence, we focus on sequences in which a HC appointment is always the last appointment in each triad.

The optimal appointment times for these sequences, which determine the positioning of the slack in the schedule, are obtained using the SIP. In all four sequences, the very first triad in the session involves a double booked slot. This follows Bailey-Welch rule (Bailey, 1952; Welch, 1964). We also see that the SIP consistently suggests the introduction of slack – an empty 15-min. slot – at the end of each triad, in each of the four sequences. The consistency of this pattern is a key finding: if the practice chooses any of the above triad structures for a session, then it is clear where slack should be located.



	SIP	(a) SD/LC/HC	(b) LC/SD/HC	(c) SD/SD/HC followed by LC/LC/HC	(D) LC/LC/HC followed by SD/SD/HC
Objective value change by using SIP schedule		-10 %	-9 %	-11 %	-10 %

Figure 4.8: Performance comparison of the SIP schedule vs. the four 3AH schedules

If we compare the performance of these four schedules (Figure 4.8) with the SIP in which both the sequence and appointment times are simultaneously optimized (see previous subsection), we find that they are between 9-11% worse in the objective value, when averaged over five afternoon sessions. This may be interpreted as the price of allowing greater flexibility in sequences to accommodate patient preferences. We note, however, that the four heuristic sequences are still 17% better on average than the ad-hoc schedules that were used in the practice.

In addition, we compare the performance of the 3AH schedules shown above with that of optimal schedules under different weights (More detailed discussion is in Section 4.4.1). We find that the average performance of the 3AH schedules over five sessions is not dominated by the SIP optimal schedules for weight combinations 0.5:0.5, 0.6:0.4 in terms of the two criteria, expecting waiting time and idle time. Indeed, idle time of the 3AH schedules is on average 27 minutes lower than that of the SIP 0.5:0.5 schedules while wait time is only 7 minutes higher.

4.3.4 Granularity of Appointment Slots

Thus far, we have assumed that our appointment slots are 15-min. long. Patient appointments will always be at the four quarters of the hour, and therefore easier to remember. But what if the practice tried appointment slots that were 5-min. long? Patients could be given appointments in 5-min. intervals and allocated a number of consecutive 5-min. slots depending on their needs. The results of such a change would be no worse, since the current 15-min. slot schedules are a feasible solution when the day is broken into 5-min. slots; in fact the schedule might use session time more efficiently. The only inconvenience would be that patients may find appointment times at, say, 9:35 am or 10:55 am, harder to recall and keep track of. We found making slot length more granular does improve the objective value, but only around 4%. The returns do not appear to be significant to justify a change.

What if the minimum slot length was 20-min. instead of 15-min.? This means that we are implicitly incorporating greater slack within each appointment. The performance of such a schedule is 6% worse compared to using 15-min. slots. As shown in Figure 4.9, we compared the different appointment slot lengths on the five afternoon sessions observed in practice. As appointment slot lengths become more granular, the objective values generated by the SIP are slightly reduced.

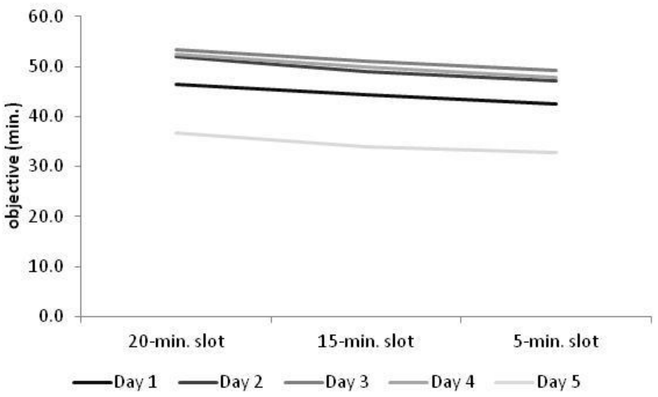


Figure 4.9: Weighted idle time and wait time (objective) by the SIP under different appointment slot lengths

Next, we study non-fixed inter-appointment times. We compare schedules and performance of the 15-min. fixed inter-appointment times and non-fixed inter-appointment times. Figure 4.10 displays optimal schedules of fixed versus non-fixed inter-appointment times from a particular day.

a) 15-min. fixed inter-appointment times

#	Time	HC	LC	SD
1	0:00		▓	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00			░
6	1:15		▓	
7	1:30		▓	
8	1:45	■		
9	2:00	■		
	2:15			
10	2:30	■		

b) non-fixed inter-appointment times

#	Time	HC	LC	SD
1	0:00		▓	
2	0:10			░
3	0:24			░
4	0:39			░
5	0:54			░
6	1:10		▓	░
7	1:25		▓	
8	1:35	■		
9	1:57	■		
10	2:21	■		

Figure 4.10: Optimal schedules of 15-minute fixed versus non-fixed inter-appointment times for a particular day

As Figure 4.10 shows, both optimal schedules follow the SPT-like sequences: SD appointments (which involve the shortest mean service time) toward to the beginning, LC in the middle, and HC appointments toward to the end. Next, Figure 4.11 illustrates the inter-appointment times between 15-min. fixed and non-fixed inter-appointment times on five afternoon sessions.

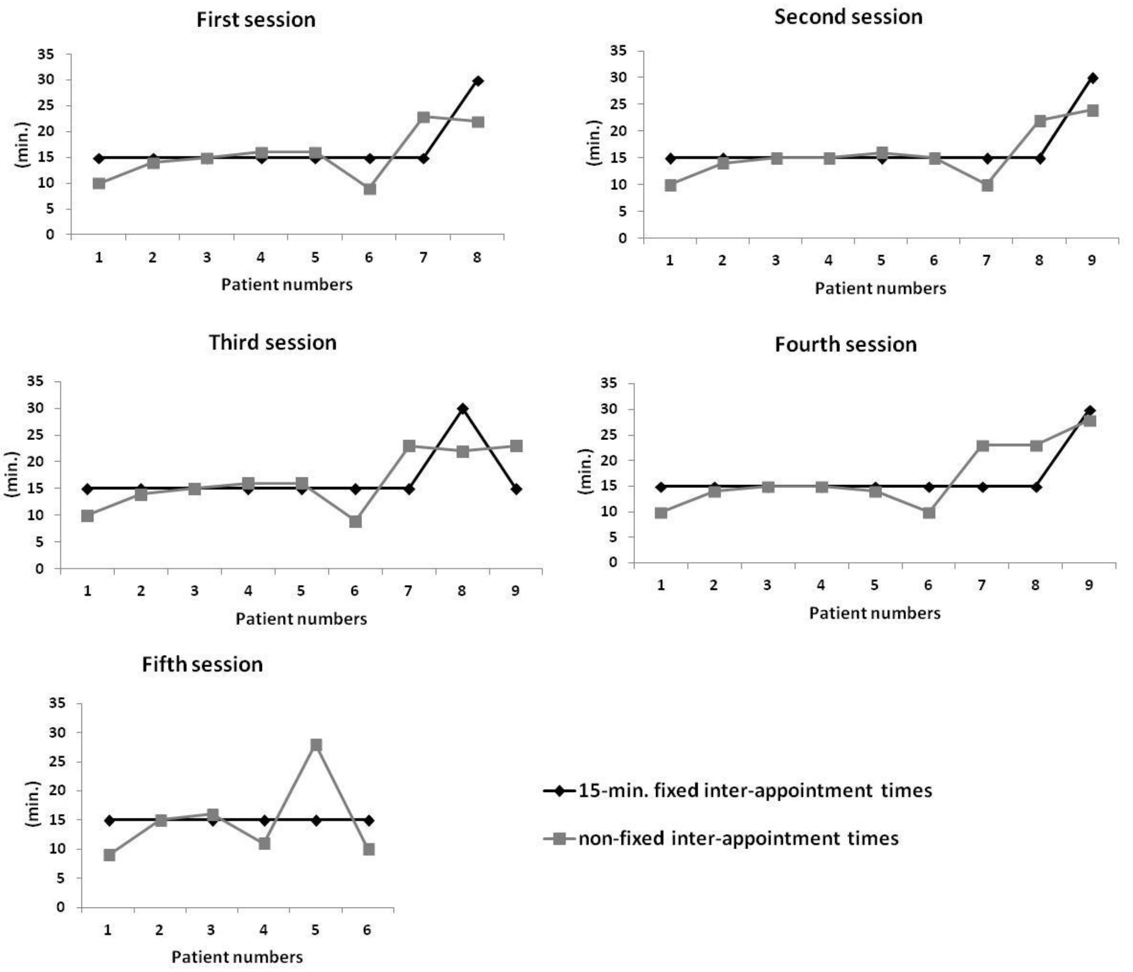


Figure 4.11: Inter-appointment times between 15-min. fixed and non-fixed on five afternoon sessions

As displayed in Figure 4.11, non-fixed inter-appointment times are fairly close to 15-min. fixed inter-appointment times over five afternoons. In 15-min. fixed inter-appointment times, 30-

min. inter-appointment times indicates to schedule an appointment in 15-min. appointment slot followed by slack. This always occurs after two HC appointments. In the non-fixed inter-appointment times, the inter-appointment times increases until the middle of the session with SD and LC patients and drops down, then increases again with HC patients. The reason the inter-appointment times falls in the middle is that an increase in the inter-appointment times provide slack, which make the inter-appointment times in the middle do not require as much as other inter-appointment times. Table 4.2 presents average of fixed and non-fixed inter-appointment times of all patients on five afternoon sessions. Overall, the difference of the objects between fixed and non-fixed appointment times is 4% on average of five afternoons.

Table 4.2: Average of inter-appointment times of all patients

Average Inter-appointment times	First session	Second session	Third session	Fourth session	Fifth session
15-min. fixed	16.9	16.7	16.7	16.7	15.0
Non-fixed	15.6	15.7	16.4	16.9	14.8

Next, we investigate the 15-min. fixed and non-fixed inter-appointment times given a heuristic triad sequences. As a reminder, we call a block of three appointments as a *triad*. Figure 4.12 and 4.13 displays the optimal appointment times between fixed and non-fixed inter-appointment times given a heuristic SD/LC/HC sequences.

a) 15-min. fixed
inter-appointment times

#	Time	HC	LC	SD
1|2	0:00		**2**	**1**
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▨	
6	1:15	■		
	1:30			
7	1:45			░
8	2:00		▨	
9	2:15	■		
	2:30			
10	2:45			░

b) non-fixed
inter-appointment times

#	Time	HC	LC	SD
1	0:00			░
2	0:08		▨	
3	0:18	■		
4	0:48			░
5	1:03		▨	
6	1:15	■		
7	1:45			░
8	2:00		▨	
9	2:11	■		
10	2:39			░

Figure 4.12: Optimal appointment times of 15-minute fixed versus non-fixed inter-appointment times given a heuristic SD/LC/HC sequence

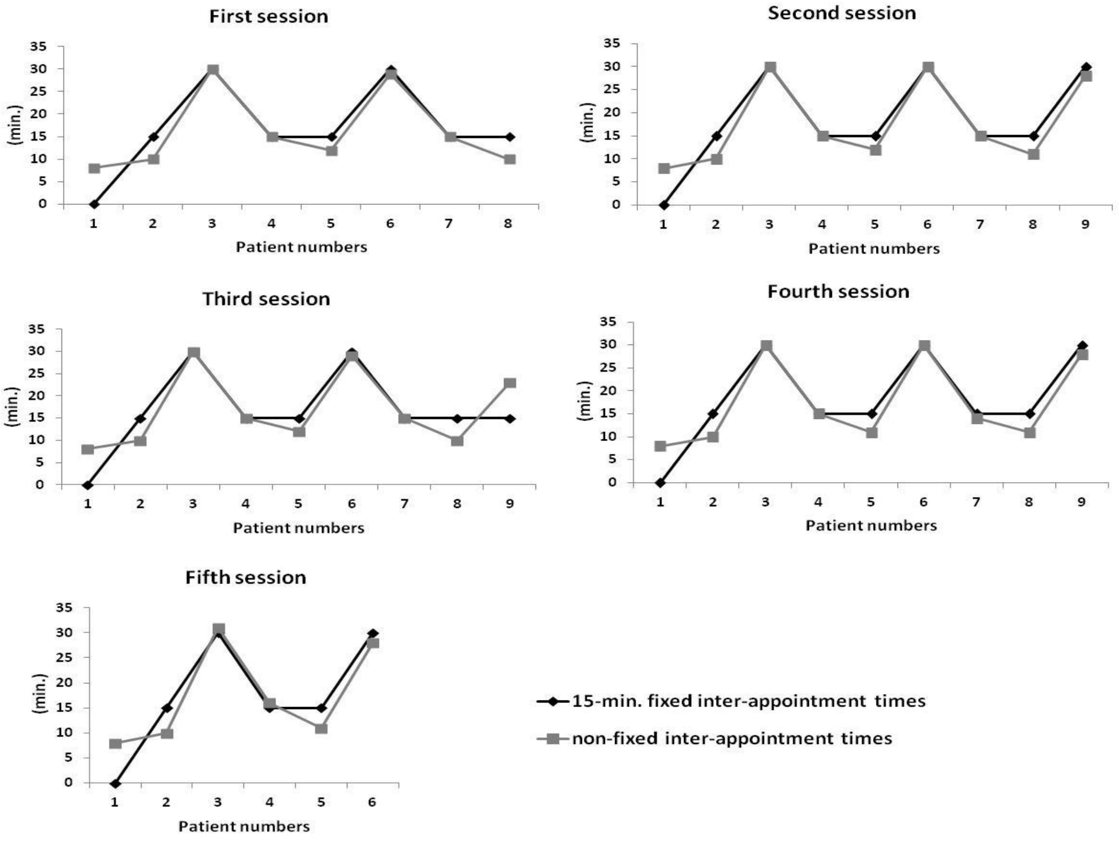


Figure 4.13: Inter-appointment times between 15-min. fixed and non-fixed on five afternoon sessions

The inter-appointment times between 15-min. fixed and non-fixed are approximate on five afternoons (Figure 4.13). In the 15-min fixed inter-appointment times, a double booking for the very first two patients and slack after a triad are optimal. In the non-fixed inter-appointment times, the first patients can be seen about 8 min. and the second patients around 10 min. in all five afternoons. Thus, it makes sense to schedule a double booking with 15-min. increment slot. Table 4.3 presents average of inter-appointment times of all patients. Overall, the difference of the objective between fixed and non-fixed inter-appointment times is 4% on average of five afternoons.

Table 4.3: Average of inter-appointment times of all patients

Average Inter-appointment times	First session	Second session	Third session	Fourth session	Fifth session
15-min. fixed	16.9	18.3	16.7	18.3	17.5
Non-fixed	16.1	17.7	16.9	17.4	17.3

4.3.5 Comparison with Provider-Only Models

We compare the performance of two models: 1) the nurse and provider model and 2) the provider only model. In the integer programs (DIP and SIP), thus, we use both steps, the nurse and provider steps, for 1), but we only account for the provider in 2). We compare the resulting schedules for each of the five afternoon sessions. The average results are summarized in Figure 4.14.

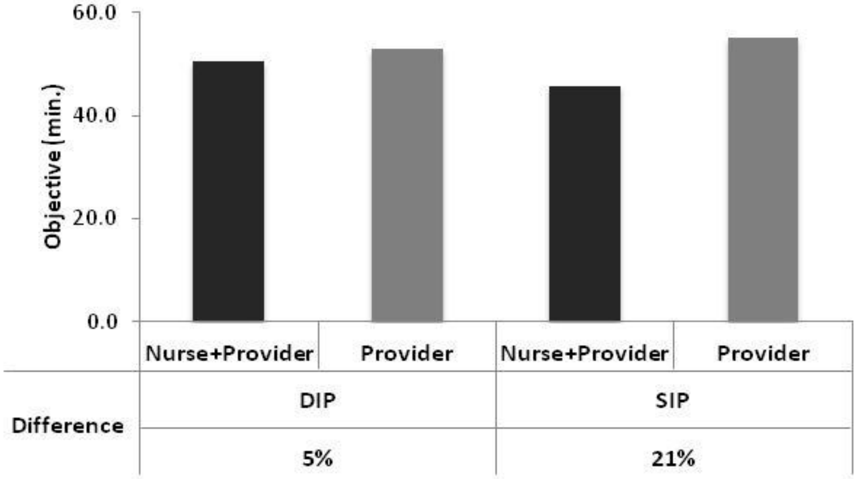


Figure 4.14: Performance improvement of nurse and provider steps vs. provider step using DIP and SIP (Nurse+Provider: the model using service time with both Nurse and Provider, and Provider: the model using service time with only Provider)

Figure 4.14 shows that considering the nurse step to generate the optimal schedule results on average in a 21% decrease in the weighted sum of provider idle time and patient wait time. In particular, wait times decrease by 64% which is fairly significant.

(a) SIP: Nurse+Provider

#	Time	HC	LC	SD
1	0:00		▓	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00			░
6	1:15		▓	
7	1:30		▓	
8	1:45	■		
9	2:00	■		
	2:15			
10	2:30	■		

(b) SIP: Provider

#	Time	HC	LC	SD
1	0:00	■		
2	0:15			░
3	0:30		▓	
4	0:45			░
5	1:00			░
6	1:15			░
7	1:30		▓	
8	1:45		▓	
9	2:00	■		
10	2:15	■		

Figure 4.15: Schedule for nurse and provider steps vs. provider step using SIP

Figure 4.15 shows the schedules generated by the SIP for 1) the nurse and provider model, versus 2) the provider model. The schedules are significantly different. The provider only schedule starts with a HC appointment at the beginning of the session and includes no slack, resulting in significantly increased patient wait times. Therefore, the nurse step is a critical factor in capturing patient wait times and needs to be considered in outpatient appointment scheduling.

4.4 Sensitivity Analysis: Objective function coefficient and no-shows

4.4.1 Coefficients of Objective Function (weight combinations)

4.4.1.1 Optimal Schedules

In the previous section, we found that 0.8:0.2 weight combination is the right balance between idle time and wait time for the practice we work with. The 0.8 weight is on provider idle time and the 0.2 weight is on patient wait time. However, other weight combinations could be more suitable depending on a practice is looking for. The practice may want to provide more weight on patient wait time than provider idle time, for instance, 0.3:0.7. However, we find that higher weight levels of wait time than idle time would render the practice financially unviable. For example, the idle time of the SIP 0.5:0.5 is already more than 50 minutes on average of five days considering only ten patients in a session. A provider cannot be idle this long and keep her practice financially viable. Therefore, we mainly focus on the same and higher weights of idle time than that of wait time: 0.5:0.5, 0.6:0.4, 0.7:0.3, 0.8:0.2, and 0.9:0.1.

Recall that deterministic integer program (DIP) uses average service time of nurse and provider and stochastic integer program (SIP) samples time of nurse and provider from the field study (more details in Section 4.2.1). Figure 4.16 and 4.17 displays the optimal DIP and SIP schedules under different weight combinations from a particular afternoon. All figures in this paper display the schedule from one particular afternoon out of five days, but the schedules we observe from other days consistently follow similar patterns.

0.5:0.5

#	Time	HC	LC	SD
1|2	0:00	2	1	
	0:15			
3	0:30	■		
4	0:45	■		
	1:00			
5	1:15			░
6	1:30			░
7	1:45		▨	
8	2:00			░
9	2:15		▨	
10	2:45			░

0.6:0.4

#	Time	HC	LC	SD
1	0:00		▨	
2	0:15			░
3	0:30		▨	
4	0:45			░
5|6	1:00	6	5	
	1:15			
7	1:30	■		
8	1:45	■		
	2:00			
9	2:15			░
10	2:45			░

0.7:0.3

#	Time	HC	LC	SD
1	0:00		▨	
2	0:15			░
3	0:30		▨	
4|5	0:45	5	4	
	1:00			
6	1:15	■		
7	1:30	■		
	1:45			
8	2:00			░
9	2:15			░
10	2:45			░

0.8:0.2

#	Time	HC	LC	SD
1	0:00		▨	
2	0:15			░
3	0:30		▨	
4|5	0:45	5	4	
	1:00			
6	1:15	■		
7	1:30	■		
	1:45			
8	2:00			░
9	2:15			░
10	2:45			░

0.9:0.1

#	Time	HC	LC	SD
1	0:00		▨	
2	0:15			░
3	0:30		▨	
4|5	0:45	5	4	
	1:00			
6	1:15	■		
7	1:30	■		
	1:45			
8	2:00			░
9	2:15			░
10	2:45			░

Figure 4.16: DIP optimal schedules under different weights from a particular afternoon

As shown in Figure 4.16, the DIP optimal schedules maintain the similar patterns as the performance of wait time and idle time generated by the DIP which is insensitive to the weights. Most of the schedules from 0.6:0.4 to 0.9:0.1 weight combinations have similar sequence patterns; LC appointments toward the beginning, HC appointments are in the middle with slack, and SD appointments toward to the end of session. Since slack is booked in the middle of the session with HC appointments which has the longest mean service time, the schedules follow the dome-shaped pattern and the slack helps reduce congestion. The schedule of the 0.5:0.5 weights is not robust since on two out of five days tested, HC appointments are booked toward to the beginning of the session, and the opposite is true on the remaining tested days. Regardless of any weight combinations in the DIP, however, slack typically occurs after two successive HC appointments, and also the double-booking between LC and HC appointments appears with slack in the middle of the session.

0.5:0.5

#	Time	HC	LC	SD
1	0:00			░
2	0:15			░
3	0:30		▓	
	0:45			
4	1:00			░
5	1:15		▓	
	1:30			
6	1:45			░
7	2:00		▓	
8	2:15	■		
	2:45			
9	3:00	■		
	3:15			
10	3:30	■		

0.6:0.4

#	Time	HC	LC	SD
1	0:00		▓	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00	■		
	1:15			
6	1:30			░
7	1:45		▓	
8	2:00	■		
	2:15			
9	2:45		▓	
10	3:00	■		

0.7:0.3

#	Time	HC	LC	SD
1	0:00		▓	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00			░
6	1:15		▓	
7	1:30	■		
	1:45			
8	2:00		▓	
9	2:15	■		
10	2:45	■		

0.8:0.2

#	Time	HC	LC	SD
1	0:00		▓	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00			░
6	1:15		▓	
7	1:30		▓	
8	1:45	■		
9	2:00	■		
	2:15			
10	2:45	■		

0.9:0.1

#	Time	HC	LC	SD
1|2	0:00		▓ 2	░ 1
3	0:15			░
4	0:30			░
5	0:45			░
6	1:00		▓	
7|8	1:15	■ 8	▓ 7	
	1:30			
9	1:45	■		
10	2:00	■		

Figure 4.17: SIP optimal schedules under different weights from a particular afternoon

On the other hand, the schedules generated by the SIP are sensitive to the weights. Figure 4.17 illustrates that when idle time weight increases, more double-booking and less slack are optimal in the SIP schedules. With 0.5:0.5 weight combination, the schedules follow the current scheduling policy: the practice uses two 15-minute appointment slots for every HC appointment, which causes high wait times. Under 0.6:0.4 weight, the schedule in the middle of the session partly resembles the heuristic 3-appointments-per-hour (3AH) schedule proposed in the previous section. We will discuss more about the heuristic schedule in the next section. From 0.7:0.3 to 0.9:0.1, slack is scheduled after two HC successive appointments. There is another slack right after double booked HC and LC appointments in the 0.9:0.1 weight combination, which is similar to the DIP optimal schedule. In general, the SIP produces an SPT-like sequence (shorter mean appointments first) under all weight combinations; HC appointments are typically scheduled toward to the end of session. Another interesting finding is that no HC appointments are scheduled at the beginning of the session in optimal SIP schedules.

As a result, the DIP optimal schedule maintains the similar patterns irrespective of weight combinations. The SIP optimal sequences also keep similar patterns under different weight combinations (SPT-like) while the SIP optimal appointment times are sensitive to the weight combinations. Both DIP and SIP optimal schedules, however, suggest one slack after two HC appointments.

4.4.1.2 Heuristic Schedules

In this section, we study our proposed heuristic 3AH schedules (3 appointments per hour – see Section 4.3.3) under different weight combinations . Recall that in the 3AH schedules, we first fix a block of three appointments (we call it a triad), which provides greater scheduling flexibility and options to patients, and then optimally solve for appointment times, assuming the weight combination 0.8:0.2. We obtained an empty slot (slack) after each triad, or there are three appointments in each hour; hence, the name is 3AH. There are four different triad sequences, a) SD/LC/HC, b) LC/SD/HC, c) SD/SD/HC followed by LC/LC/HC, and d) LC/LC/HC followed by SD/SD/HC.

First, we compare the performance of wait, idle, and session completion times of the 3AH schedules with that of optimal SIP schedules under different weight combinations. Figure 4.18 shows the performance of 3AH schedules in comparison to the SIP optimal schedules under different weight combinations. When 3AH schedules are fixed, all performance of wait, idle, and completion time are approximately under all weight combinations.

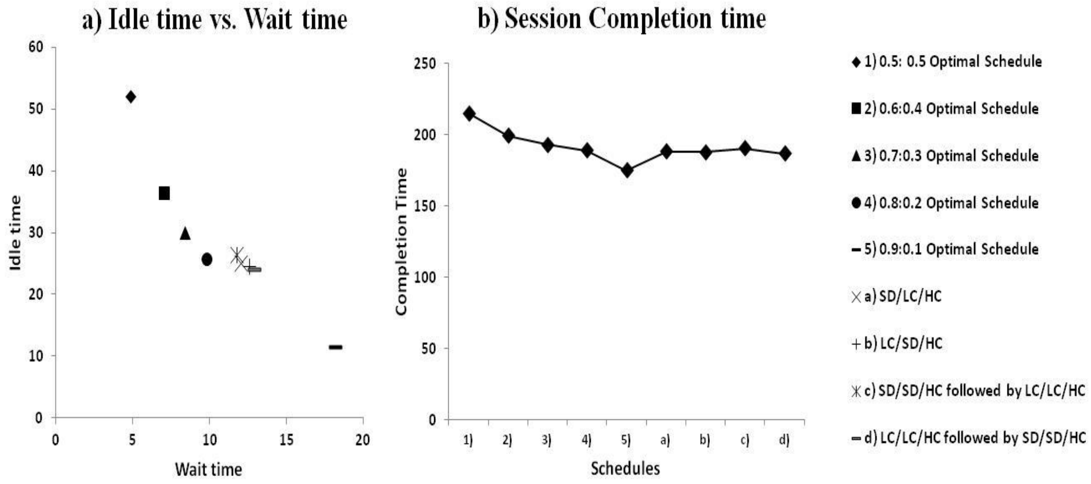


Figure 4.18: Performance comparison of optimal SIP and 3AH schedules

The average performance of the 3AH schedules over five afternoon sessions is not dominated by that of the optimal SIP schedules under any weights (Figure 4.18 a). The 3AH schedules have higher wait time but lower idle time than these optimal schedules with 0.5:0.5 to 0.8:0.2 while the opposite occurs with 0.9:0.1. Average session completion times of the 3AH policies are 12%, 6% and 2% better than those in the 0.5:0.5, 0.6:0.4, and 0.7:0.3 optimal schedules, respectively. Although the 3AH schedules do not significantly perform better than the optimal schedules with different weights, they are quite well in comparison to the optimal. They are also easy to implement and provide schedulers and patients with increased flexibility.

Next, we optimize the appointment times given a block of three appointments, a triad, under different weight combinations. This provides how the appointment times with the fixed triad sequence are affected by different weights. Figure 4.19 shows the optimal appointment times given one of four triad sequences, SD/LC/HC under multiple weights for a particular session.

0.5:0.5

#	Time	HC	LC	SD
1	0:00			░
2	0:15		▨	
3	0:30	■		
	0:45			
	1:00			
4	1:15			░
5	1:30		▨	
6	1:45	■		
	2:00			
	2:15			
7	2:30			░
8	2:45		▨	
9	3:00	■		
	3:15			
10	3:30			░

0.6:0.4

#	Time	HC	LC	SD
1	0:00			░
2	0:15		▨	
3	0:30	■		
	0:45			
4	1:00			░
5	1:15		▨	
6	1:30	■		
	1:45			
	2:00			
7	2:15			░
8	2:30		▨	
9	2:45	■		
	3:00			
10	3:15			░

0.7:0.3

#	Time	HC	LC	SD
1	0:00			░
2	0:15		▨	
3	0:30	■		
	0:45			
4	1:00			░
5	1:15		▨	
6	1:30	■		
	1:45			
7	2:00			░
8	2:15		▨	
9	2:30	■		
	2:45			
10	3:00			░

0.8:0.2

#	Time	HC	LC	SD
1|2	0:00		▨	░
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▨	
6	1:15	■		
	1:30			
7	1:45			░
8	2:00		▨	
9	2:15	■		
	2:30			
10	2:45			░

0.9:0.1

#	Time	HC	LC	SD
1|2	0:00		▨	░
3	0:15	■		
4	0:30			░
5	0:45		▨	
6	1:00	■		
	1:15			
7	1:30			░
8	1:45		▨	
9	2:00	■		
10	2:15			░

Figure 4.19: Optimal appointment times given SD/LC/HC sequence under different weights from a particular afternoon

As shown in Figure 4.19, number of slack decreases as the weight on idle time increases. 3AH rule (three appointments schedule in an hour; a triad followed by slack) is fairly robust under 0.6:0.4, 0.7:0.3, and 0.8:0.2 weights. We find that the objective of the SIP optimal schedules performs approximately 7 to 13% better than that of four triad sequences under the same weight combination. Hence, if the practice chooses the triad sequence, slack can be scheduled after a triad. However, the number of empty slots (slack) depends on the weight combination the practice wants to consider.

4.4.2 No-show Rates

We now consider no-show rates of the prescheduled appointments, and the chance that a same-day appointment may go idle if no patient ends up being scheduled in the slot. We model different no-show rates by allowing zero service times with provider and nurse in the data. In generating the scenarios, we randomly select zero-length durations in the sample average approximation method. The number of times zero-length duration appears is set equal to the no-show or idle probability. According to Cayirli and Veral (2003), no-show rates range from 5 to 30 percent. Thus, we have examined 5 to 30 percents of no-show rates in 5% increments. In addition, we investigate 3% which is observed from the practice we collected data. We examine three cases under different no-show rates: 1) optimal schedules, 2) optimal appointment times given heuristic triad sequences, and 3) performance comparison between a triad sequence with slack and one without slack, in order to analyze how the schedule and the performance adjust to different circumstances with no-show rates.

First, we study optimal schedules (appointment times and sequences) under different no-show rates to examine the performance and the provision or location of slack. Figure 4.20 and 4.21 display the schedules and the performance of average, median, and 90th percentile of the optimal schedules out of five days (Refer to Section 4.3.2 for the practice schedule and 0% no-show schedule which could not be included in this section for brevity.)

3%

#	Time	HC	LC	SD
1	0:00		▓	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00			░
6	1:15		▓	
7	1:30		▓	
8	1:45	■		
9	2:00	■		
	2:15			
10	2:30	■		

5%

#	Time	HC	LC	SD
1	0:00		▓	
2	0:15			░
3	0:30			░
4	0:45			░
5	1:00			░
6	1:15		▓	
7	1:30		▓	
8	1:45	■		
9	2:00	■		
	2:15			
10	2:30	■		

10%

#	Time	HC	LC	SD
1|2	0:00		▓ 1	░ 2
3	0:15			░
4	0:30			░
5	0:45			░
6	1:00		▓	
7	1:15		▓	
8	1:30	■		
9	1:45	■		
10	2:00	■		

15%

#	Time	HC	LC	SD
1|2	0:00		▓ 1	░ 2
3	0:15			░
4	0:30			░
5	0:45			░
6	1:00		▓	
7	1:15		▓	
8	1:30	■		
9	1:45	■		
10	2:00	■		

20%

#	Time	HC	LC	SD
1|2	0:00		▓ 1	░ 2
3	0:15			░
4	0:30			░
5	0:45			░
6	1:00		▓	
7|8	1:15	■ 8	▓ 7	
9	1:30	■		
10	1:45	■		

25%

#	Time	HC	LC	SD
1|2	0:00		▓ 1	░ 2
3	0:15		▓	
4	0:30			░
5	0:45			░
6|7	1:00		▓ 7	░ 6
8	1:15	■		
9	1:30	■		
10	1:45	■		

30%

#	Time	HC	LC	SD
1|2	0:00		▓ 1	░ 2
3	0:15			░
4	0:30			░
5	0:45			░
6|7	1:00		▓ 6|7	
8	1:15	■		
9	1:30	■		
10	1:45	■		

Figure 4.20: Optimal Schedules under different no-show rates from a particular afternoon

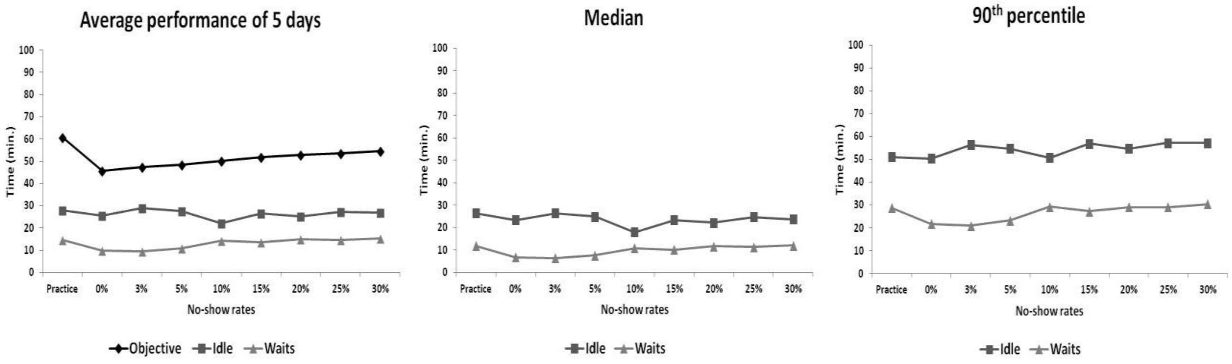


Figure 4.21: Performance of average, median and 90th percentile of optimal schedules under different no-show rates out of five days

As shown in Figure 4.20 and 4.21, the objectives of optimal schedules keep increasing when no-show rates rise since the schedule gets packed. With 10% no-show rates, we can see idle time dropping down and waits going up slightly because the first two appointments are double booked and there is no slack. With higher than 10% no-show rates, slack hardly exists and more double booking occurs in the schedule. The double booking for the first and second appointments is a promising scheduling strategy if the practice has more than 10% no-show rates. If the practice has fewer than 5% no-show rates, slack should be provided after two HC appointments type since the schedule hardly changes until 5% no-show rates, when compared to 0% no-shows. The optimal SPT-like sequence is fairly robust irrespective of no-show rates.

Second, we determine optimal appointment times of a heuristic triad sequence under different no-show rates so as to study optimal appointment times and performance. We select one of the triad sequences (SD, LC and HC). Since we want to study how the optimal appointment times change under different no-show probabilities, investigating only SD/LC/HC sequence provide sufficient schedule/performance information. Figure 4.22 shows optimal appointment times given SD/LC/HC sequence under different no-show rates from one of five days. The performance of average, median and 90th percentile of optimal appointment times with respect to SD/LC/HC out of 5 days is shown in Figure 4.23.

3%

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▨	
6	1:15	■		
	1:30			
7	1:45			░
8	2:00		▨	
9	2:15	■		
	2:30			
10	2:45			░

5%

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▨	
6	1:15	■		
	1:30			
7	1:45			░
8	2:00		▨	
9	2:15	■		
	2:30			
10	2:45			░

10%

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▨	
6	1:15	■		
7	1:30			░
8	1:45		▨	
9	2:00	■		
	2:15			
10	2:30			░

15%

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▨	
6	1:15	■		
7	1:30			░
8	1:45		▨	
9	2:00	■		
	2:15			
10	2:30			░

20%

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
	0:30			
4	0:45			░
5|6	1:00	6	5	
	1:15			
7	1:30			░
8	1:45		▨	
9	2:00	■		
10	2:15			░

25%

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
4	0:30			░
5	0:45		▨	
6	1:00	■		
7	1:15			░
8	1:30		▨	
9	1:45	■		
10	2:00			░

30%

#	Time	HC	LC	SD
1|2	0:00		2	1
3	0:15	■		
4	0:30			░
5	0:45		▨	
6	1:00	■		
7	1:15			░
8	1:30		▨	
9	1:45	■		
10	2:00			░

Figure 4.22: Optimal appointment times given SD/LC/HC sequence with different no-show rates from a particular day

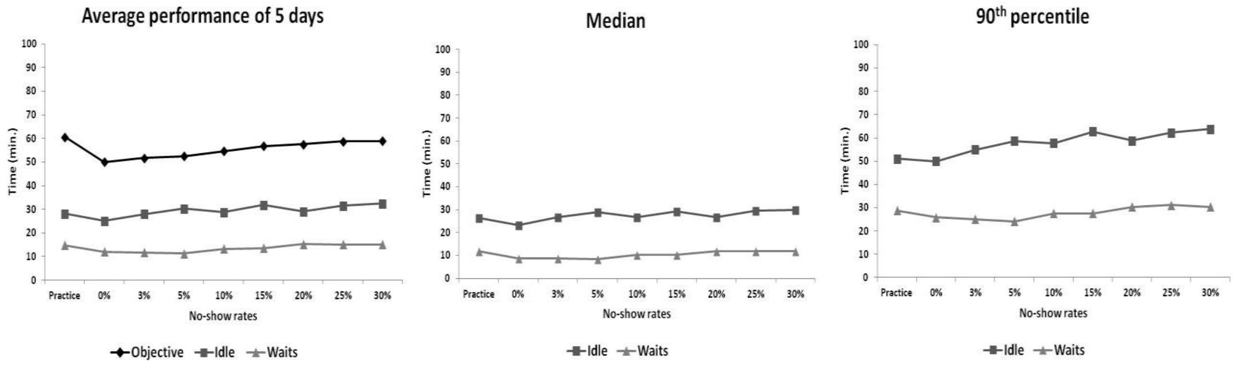


Figure 4.23: Performance of average, median and 90th percentile of optimal appointment times given SD/LC/HC sequence under different no-show probabilities out of five days

As shown in Figure 4.22 and 4.23, the objectives have slightly increased when no-show rates increase since the schedule gets packed – less slack and more double booking. Similar to optimal schedule under different no-show rates, the optimal appointment times given the heuristic schedule have almost no changes until 5% no-show rates, when compared with 0% no-show. When slack needs to be provided, it is scheduled after a triad. In addition, slack in the schedule scarcely occurs after 20% no-show rates. All schedules in any no-show rates start with a double booking of same-day and LC appointments.

Third, we use two schedules: a) slack triad schedule which represents the schedule involving slack after each triad (same as 3AH schedules), and b) no-slack triad schedule which indicates the schedule with no slack after each triad, shown in Figure 4.24. Again, we used the SD/LC/HC triad sequence. Given these two schedules, we want to observe that the 3AH is a secure schedule policy in any different no-show rates. In other words, we want to study whether slack is necessary with triad sequences under different no-show rates. We also keep the double-booking in the first slot since all schedules under different no-show rates involve the double-booking.

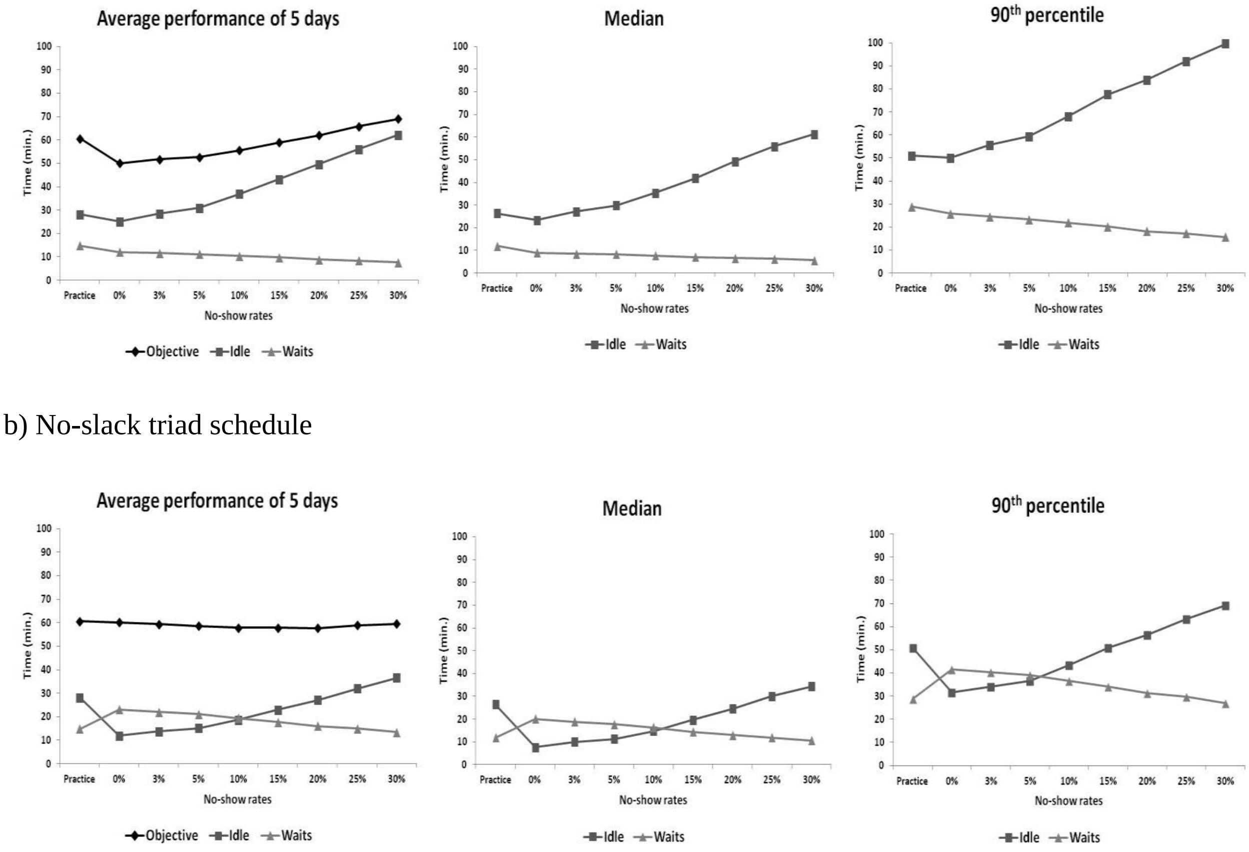
a) Slack triad schedule

#	Time	HC	LC	SD
1|2	0:00		**2**	**1**
3	0:15	■		
	0:30			
4	0:45			░
5	1:00		▓	
6	1:15	■		
	1:30			
7	1:45			░
8	2:00		▓	
9	2:15	■		
	2:30			
10	2:45			░

b) No-slack triad schedule

#	Time	HC	LC	SD
1|2	0:00		**2**	**1**
3	0:15	■		
4	0:30			░
5	0:45		▓	
6	1:00	■		
7	1:15			░
8	1:30		▓	
9	1:45	■		
10	2:00			░

Figure 4.24: Slack triad schedule vs. No-slack triad schedule from a particular afternoon


a) Slack triad schedule

b) No-slack triad schedule

Figure 4.25: Performance of average, median and 90th percentile of slack triad schedule and no-slack triad schedule under different no-show rates out of 5 days

In the slack triad schedule shown in Figure 4.24 and 4.25 a), the objective gradually increases as the no-show rates raise because while patients wait decrease, a provider idle time significantly increases. Since we fix the triad schedule with slack, the idle time dramatically increases as no-show rates increase. On the other hand, the objectives of the no-slack schedule under different no-show rates shown in Figure 4.24 and 4.25 b) stay within the 4% range; these objectives are almost similar object level of the current practice schedule. Since there is no slack until 10% no-show rates, wait time is higher than the idle time and vice versa after 10%, on average and median. We find that if the practice has more than 10% no-show rates, it may be a proper strategy not to have slack in the schedule.

4.5 Conclusion

We formulate a stochastic program to model the appointment sequencing and scheduling problem under the new classification and two sequential service steps (nurse and provider). The objective is to minimize a weighted combination of patient wait time and provider idle time. The model sequences patient types with different nurse and provider time requirements and staggers their appointment times appropriately while keeping the basic slot structure traditionally used by the schedulers at the practice.

The contributions of our research are as follows. First, from an operational point of view, we demonstrate that different amounts of slack are necessary in the schedule depending on the type of patient. It is known that patients with chronic conditions need longer appointments with their providers. Our model provides sufficient space in the schedule for such patients, yet ensures that provider idle time is not more than necessary.

Second, from a modeling point of view, we develop, unlike previous studies, a stochastic program that captures both the patient classification and the entire patient flow through the practice including initial wait, nurse check-up, wait in exam room and provider check-up. Third, we determine the optimal placement of slack (unscheduled slot times) to mitigate the effect of variability of service time with the nurse and the provider, under various patient sequences; this includes sequences that are attractive to the practice because they facilitate the accommodation of patient preferences and yet are financially viable. Our analysis of these sequences shows that optimal appointment times consistently follow a specific structure: an empty slot after every group of three scheduled appointments that includes a 30-min. patient. This results in easy to implement guidelines. Finally, we compare the proposed optimal and heuristic schedules with schedules actually used in practice.

Although our collaborating practice can be a representative of primary care practices, we also study the general settings of practices by investigating the wide range of objective function coefficients and no-show rates. We find that sequences of patient types maintain the same regardless different coefficients of the objective function and no-show rates; however, number of slack is sensitive to different coefficients and no-shows.

We have studied scheduling problems for a single nurse and provider. In the next chapter, we extend our model and the Excel simulation tool to practices with shared resources. For instance, we consider two nurses that can flexibly attend to the needs of the patients of two providers.

CHAPTER 5

TEAM PRIMARY CARE PRACTICE

5.1 Introduction

Effective scheduling in primary care practices plays an important role in smoothing patient flow. Many papers have studied the scheduling problem in the outpatient setting, but commonly assume a single step in the patient flow process: the provider step. However, many practices also involve a nurse step prior to the provider step. According to our empirical data analysis in Chapter 3, nurse service time durations are comparable for many appointments to provider service time durations. For example, for routine physicals and well child exams – two common appointment types in primary care – nurses spend as much time with the patients as providers. In addition, we found that there is a significant difference in the performance as well as the structure of the optimal schedule when the nurse step is explicitly considered in the scheduling formulation compared to when it is not.

Another common assumption is a single resource at each step: for example, a solo provider working at the practice. However, the majority of practices (68%) have more than two providers (Bodenheimer and Pham, 2010). Also, while collecting data, we have observed that nurses may work as a team in prepping patients for provider appointments. We call this a *team primary care practice*. In this case, nurses flexibly see patients scheduled on providers' calendars whenever they are available while providers stay dedicated to their appointment schedules.

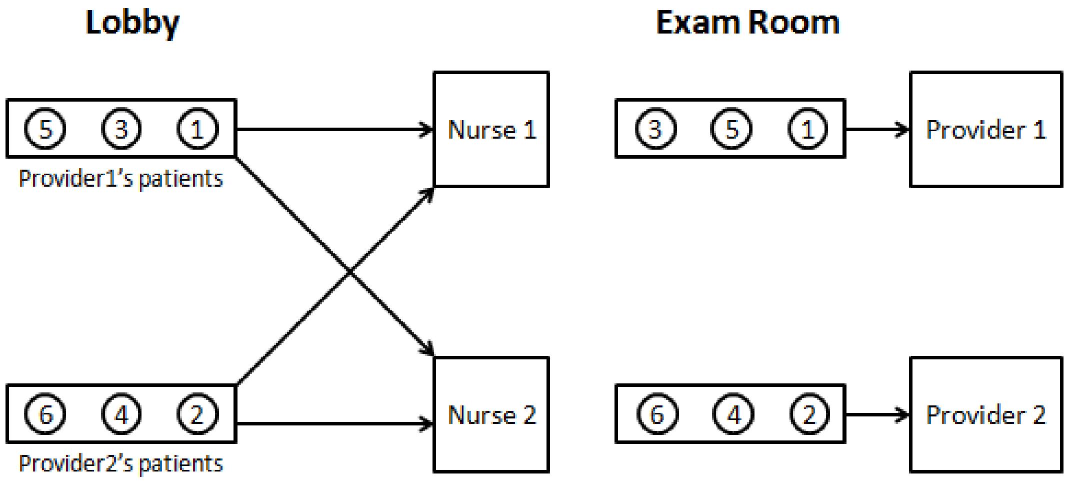


Figure 5.1: Patients schedule appointments with and are seen by their own personal provider, but they can be examined by any of the two nurses in the practice. Because of the flexible nurse step, the order in which appointments for a provider are scheduled may not be the order in which the provider ends up seeing them. For Provider 1 above, Patient 3 had an earlier appointment than Patient 5, but after the nurse step, Patient 5 arrived at the exam room earlier (i.e. crossover)

This multi-step patient flow process with multiple human resources at each step coupled with diverse patient conditions and uncertain service times make the problem extremely challenging from an optimal scheduling viewpoint. In our previous chapter, we have proposed three well differentiated appointment types based on time requirements: high complexity (HC), low complexity (LC) and same day (SD). With these three appointment types, we optimize the appointment times and sequences for a *single primary care practice* where one nurse and one provider see patients and conclude that the slack (open empty slot) position has a significant impact on reducing patient wait time and catching up the delayed work for providers. In addition, we optimize the appointment times with each appointment type and obtain the same guidelines of appointment times as we found from multiple appointment types.

We develop a mixed integer programming model incorporating multiple resources at two sequential steps with stochastic service times to minimize the weighted measure of patient wait time and provider idle time. In this paper, we limit ourselves to scheduling patients of type HC – that is, patients with complex conditions who need sufficient time with providers. Practices that schedule only HC appointments do exist in reality. For example, federally qualified health centers in the US provide primary care to complex patients with multiple conditions. Providers in such practices primarily book type HC appointments for their patients.

From the modeling perspective, the structure of the model is similar to flexible flow shop (FFS) – two stages: nurse and provider; and two human resources (machines) at each stage. In addition, each job is processed on one machine through each stage sequentially - each patient is seen by first a nurse and then a provider. While the FFS problem involves the deterministic processing time and decides the start time and sequencing with multiple types of job, our problem includes stochastic service time and determines optimal appointment times with one appointment type. The unique structure in our model is to take into account flexibility among machines in the first stage and flexibility among jobs to the dedicated machine in the second stage; in other words, nurses are flexibly seeing patients while providers are dedicated to their own panel and within the panel, a provider sees earliest available patients after the nurse step.

To extend the scheduling problem of a single appointment type, we develop an Excel scheduling simulation tool to accommodate multiple appointment types. The main is to provide a user-friendly Excel simulation tool for schedulers to manage appointment schedules which accommodate three well-differentiated patient classes and multiple steps in the patient flow process. This tool can be easily modified to include more human resources, patient types, and performance measures. In the case study, we compare the performance of a single-provider and team primary practices using schedules from the previous work.

In summary, we propose a novel stochastic integer programming formulation which can be a representative of many practices where multiple nurses and providers are involved in the stochastic patient flow process. To the best of our knowledge, previous papers have not studied team practices from a mathematical programming perspective. Since this problem is computationally challenging, we develop tightening constraints and lower bounds to improve running time. In the computational study, we study various experiments with empirical data that we collected from a family medicine practice. Since we limit to schedule for one appointment type, we develop the Excel scheduling tool to study the scheduling problem with multiple appointment types.

The rest of the article is structured as follows. In Section 5.2, we explain the team practice with visualized aid-Gantt chart, address the mathematical model and solution method, and discuss computational study for a single appointment type. In Section 5.3, we present features of Excel scheduling tool and study scheduling with multi-appointment types. In Section 5.4, we summarize our conclusions.

5.2 Single Appointment Type

5.2.1 Model Construction

5.2.1.1 Description of Team Primary Care Practice

We collected data over nine days by conducting an observational time study of the patient flow at a three-provider family medicine practice in Massachusetts (see Chapter 4 for details). In a previous study, we focused on the *single-provider* primary care practice composed of a single nurse and provider. While collecting the data, however, we observed that *two* nurses and *two* providers typically see patients in each session. This *multi-provider* or *team* primary care practice is the focus of our current model.

There are morning and afternoon sessions distinguished by a lunch break. The patient visit consists of the following steps: after check-in, a patient waits in the lobby until a nurse calls (wait time in the lobby); the first available nurse calls the patient into the exam room and examines the patient (nurse service time); after the nurse step, the patient waits in the exam room until her/his primary provider is available (wait time in the exam room); and once the provider finishes with the previous patient, she/he will examine the patient (service time with provider). In this team practice, the two nurses flexibly share patients (*flexible nurses*) while each provider oversees appointments only from his/her own panel. A provider takes care of the earliest available patient from his own panel after the nurse step. In other words, the provider sees patients in the order of their finish times at the nurse step, instead of the order of appointment times. *Patient crossover* will thus occur when a patient with an earlier appointment may have a long nurse service and end up seeing the provider after a patient with a later appointment. In the following, we describe how we incorporate flexible nurses at the nurse step and patient crossover at the provider step into the model.

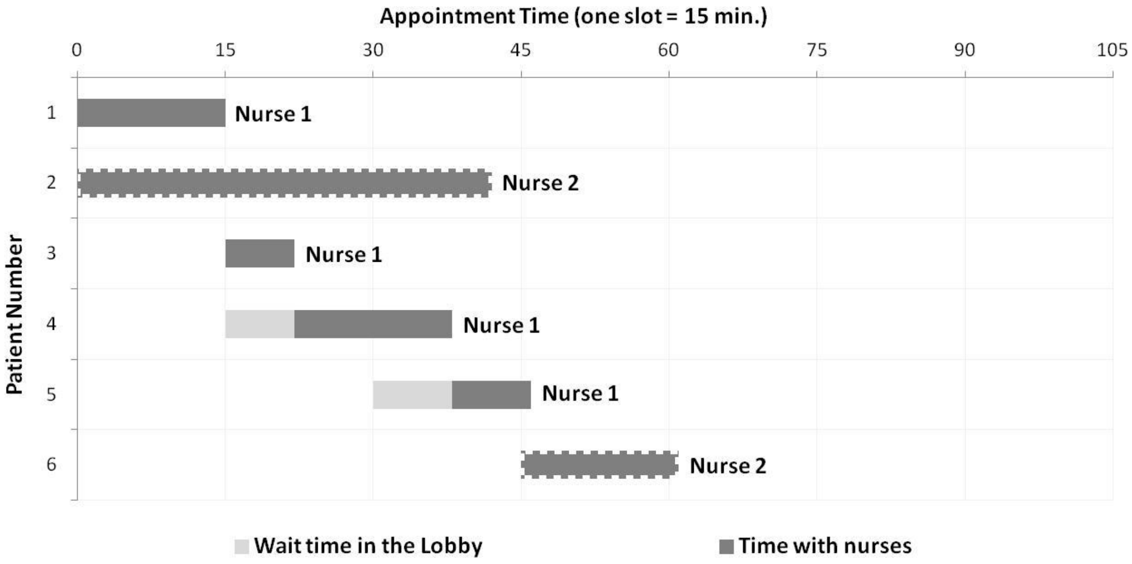


Figure 5.2: Example of patients' flow through the nurse step under nurse flexibility

The Gantt chart in Figure 5.2 illustrates the sharing of patients by flexible nurses. Each appointment slot is 15-min. long (common for American primary practices). For simplicity, the figure includes only three components of the patient flow: appointment time, wait time in the lobby, and service time with nurses. At the beginning of the session, nurse1 and 2 take care of patients 1 and 2, respectively. Since nurse2 sees patient 2 for 42 minutes, nurse1 proceeds to see patients 3, 4 and 5. Finally, nurse2 completes patient 2's visit while nurse 1 is busy with patient 5; as a result, nurse2 sees patient 6. In the flexible nurse environment, the earliest available nurse sees the next patient. Mathematically, we capture the flexible nurse schedule by dynamically comparing, for each subsequent patient, the finish times of the current patients with each nurse: the time at which a nurse will be available to see patient *i* can be recursively calculated as *the second largest value of the finish times of the nurse visits with patients 1 to i-1*. This second largest value, in turn, can be calculated as the minimum of the finish time of patient *i*-1 with a nurse and the recursively calculated maximum of the finish times of patients 1 through *i*-2.

The Gantt chart in Figure 5.3 illustrates an example of patient crossover in the provider step. In addition to the three components of patient flow discussed above, we now account for service time with provider and provider idle time.

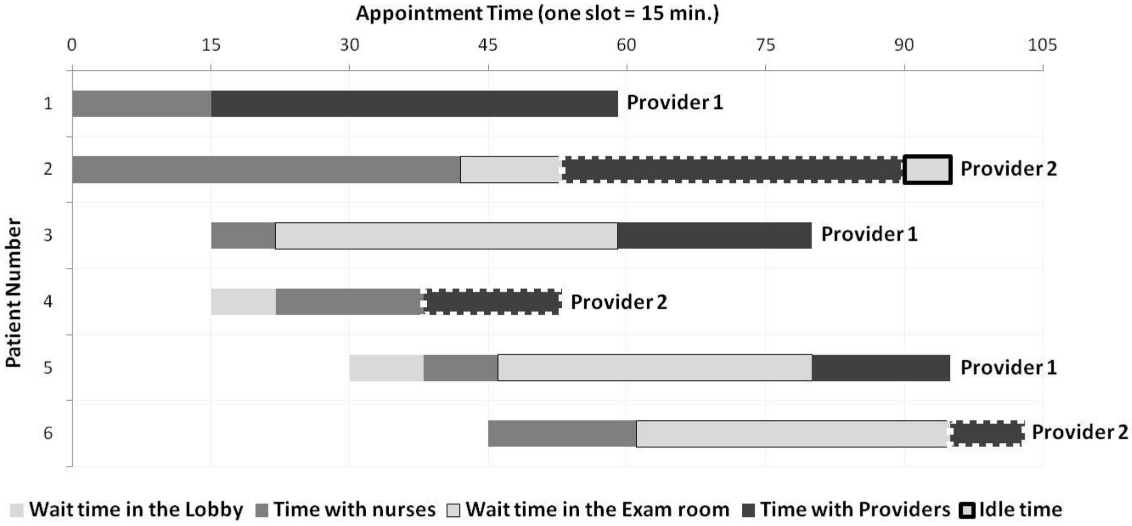


Figure 5.3: An example of patient crossover in the provider step

As shown in Figure 5.3, each provider sees patients exclusively in their own panel: provider 1's panel includes patients 1, 3, and 5 (black bars) and provider 2's panel consists of patients 2, 4, and 6 (serrated bars). Within their panels, the providers will see the earliest available patient after the nurse step. For example, provider1 sees patient1 first since pateint1 is the earliest available one among her/his panel after seeing the nurse. However, the first patient of provider2 is patient4, who is the earliest available after the nurse step within her/his panel. That is, a provider examines patients in the order of completion of the nurse step instead of the scheduled order of patient appointments. To mathematically capture this patient crossover in the providers' schedule, we dynamically compare the finish times with nurse within each provider's panel: the time when the next patient completes the nurse step and is ready to be seen by his primary care provider is calculated recursively applying the second largest logic again.

We conclude this section with a full list of our modeling assumptions: the practice's schedule is structured in 15-min slots, with each patient appointment to be allocated to a fixed 15-min slot; all patients arrive punctually for their appointments; patients can be seen by the earliest available nurse; each provider is exclusively dedicated to patients of her/his own panel; providers will see patients from their panels in the order in which they complete the nurse step; and service times with nurses and providers are independent and identically distributed.

5.2.1.2 Distribution of Service Time

In Figure 5.4, we present the distributions of service time with nurse and provider for high complexity patient visits, which we denote as type HC appointments. Type HC involves physicals and complex conditions, which require long service time with nurses and providers.

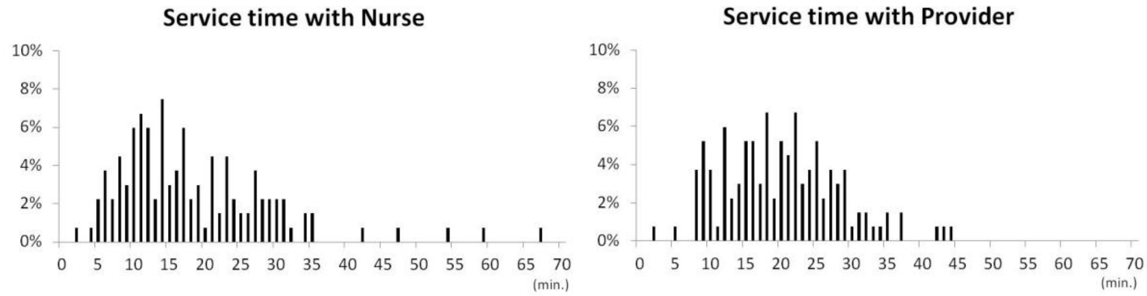


Figure 5.4: Distribution of service time with nurse and provider

As shown in Figure 5.4, service times with both nurse and provider are highly variable. Although provider service times tend to be longer, the service time distribution for the nurse step is skewed to the right, leading to nurse visits that are significantly longer than the provider visits. It is apparent that the nurse and provider steps should be effectively coordinated in order to avoid long patient waits or low provider utilization. On average, a type HC appointment takes 17.8 min (standard deviation: 10.7 min.) with nurse, and 19.5 min. (standard deviation: 8.2) with provider - See Oh et al. 2013 for service time information of other appointment types. The variability of service times makes the scheduling problem challenging.

In the next section, we formulate an integer program to find the optimal patient appointment times in a practice with flexible nurses, patient crossover, and stochastic service times in the nurse and provider steps.

5.2.1.2 Integer Programming Formulation

We formulate a mixed integer program to schedule patients into appointment slots. Key features of the model are to accommodate two sequential steps - nurse and provider, multiple human resources at each step, stochastic service times, flexible nurses, providers dedicated to own panels, and patient crossover. We use a fixed, predetermined appointment length of 15-minutes and consider homogeneous patients; we focus type HC patients in our computational work as described above. The objective of the model is to minimize a weighted measure of provider idle time and patient wait time across all scenarios. We assume that the patients punctually arrive at the appointed time since 89% of patients come early or on time based on our data analysis.

We will use the following notation to formulate the problem.

Sets:

I Set of patients to be scheduled in the session, indexed by i = 1,…, I

J_k Set of patients to be scheduled with provider k, indexed by j = 1,…, J_k

S Set of scenarios, indexed by s = 1,…, S

K Set of providers, indexed by k = 1,2

Parameters

α Weight for idle time

β Weight for wait time

$\tau_{i,s}^{N}$ Service time of patient i with nurse under scenario s

$\tau_{j,s}^{P_k}$ Service time of patient j with provider k under scenario s

$f[j,k]$ Patient index (in the overall set of patients in the practice) of the jth patient of provider k

Variables

$y_{i,s}^{start}$ Start time of patient i with nurse under scenario s

$y_{i,s}^{finish}$ Finish time of patient i with nurse under scenario s

$t_{j,s}^{k}$ Finish time with nurse of the j^{th} patient in provider k's panel under scenario s

$z_{j,s}^{k,start}$ Start time of the j^{th} patient to visit with provider k under scenario s

$z_{j,s}^{k,finish}$ Finish time of the j^{th} patient j to visit with provider k under scenario s

$N_{i,s}^{max}$ Maximum of the finish times of patients 1,…, i-1 with nurses under scenario s

$P_{j,s}^{k,max}$ Maximum of the finish times of patients 1,…, j of provider k's panel under scenario s

X_i Appointment slot assigned to patient i, an integer variable in {0,1,2,...}.

The problem is modeled as the following integer program.

$$Min. \qquad \frac{1}{S}\Bigg(\alpha\Bigg[\sum_{s}\Bigg(\Bigg(z_{J_1,s}^{1,finish}-\sum_{j_1=1}^{J_1}\tau_{j,s}^{P_1}\Bigg)+\Bigg(z_{J_2,s}^{2,finish}-\sum_{j_2=1}^{J_2}\tau_{j,s}^{P_2}\Bigg)\Bigg)\Bigg]$$

$$+\beta\Bigg[\sum_{s}\sum_{i=1}^{n}\left(y_{i,s}^{start}-15X_i\right)$$

$$+\sum_{s}\Bigg(\sum_{j=1}^{J_1}\left(z_{j,s}^{1,start}-t_{j,s}^{1}\right)+\sum_{j=1}^{J_2}\left(z_{j,s}^{2,start}-t_{j,s}^{2}\right)\Bigg)\Bigg]\Bigg) \qquad (1)$$

Subject to.

$$y_{1,s}^{start}=0 \qquad \forall s\in S \qquad (2)$$

$$y_{2,s}^{start}=0 \qquad \forall s\in S \qquad (3)$$

$$z_{0,s}^{k,finish}=0 \qquad \forall k\in K, s\in S \qquad (4)$$

$$X_1=0 \qquad (5)$$

$$X_2=0 \qquad (6)$$

$$y_{3,s}^{start}\geq \min\left(y_{1,s}^{finish}, y_{2,s}^{finish}\right) \qquad \forall s\in S \qquad (7)$$

$$N_{3,s}^{max}\geq \max\left(y_{1,s}^{finish}, y_{2,s}^{finish}\right) \qquad \forall s\in S \qquad (8)$$

$$N_{i,s}^{max}\geq \max\left(N_{i-1,s}^{max}, y_{i-1,s}^{finish}\right) \qquad \forall i\in 4..I, s\in S \qquad (9)$$

$$y_{i,s}^{start}\geq \min\left(N_{i-1,s}^{max}, y_{i-1,s}^{finish}\right) \qquad \forall i\in 4..I, s\in S \qquad (10)$$

$$y_{i,s}^{finish}=y_{i,s}^{start}+\tau_{i,s}^{N} \qquad \forall i\in I, s\in S \qquad (11)$$

$$y_{i,s}^{start}\geq 15X_i \qquad \forall i\in I, s\in S \qquad (12)$$

$$t_{j,s}^{k}=y_{f[j,k],s}^{finish} \qquad \forall k\in K, j\in J_k, s\in S \qquad (13)$$

$$z_{1,s}^{k,start}\geq \min\left(t_{1,s}^{k}, t_{2,s}^{k}\right) \qquad \forall k\in K, s\in S \qquad (14)$$

$$P_{2,s}^{k,max}\geq \max\left(t_{1,s}^{k}, t_{2,s}^{k}\right) \qquad \forall k\in K, s\in S \qquad (15)$$

$$P_{j,s}^{k,max}\geq \max\left(P_{j-1,s}^{k,max}, t_{j,s}^{k}\right) \qquad \forall k\in K, j\in\{3..J_k\}, s\in S \qquad (16)$$

$$z_{j,s}^{k,start}\geq \min\left(P_{j,s}^{k,max}, t_{j+1,s}^{k}\right) \qquad \forall k\in K, j\in\{2..J_k-1\}, s\in S \qquad (17)$$

$$z_{J,s}^{k,start}\geq P_{J_k,s}^{k,max} \qquad \forall k\in K, s\in S \qquad (18)$$

$$z_{j,s}^{k,finish}=z_{j,s}^{k,start}+\tau_{j,s}^{P_k} \qquad \forall k\in K, j\in J_k, s\in S \qquad (19)$$

$$z_{j,s}^{k,start}\geq z_{j-1,s}^{k,finish} \qquad \forall k\in K, j\in J_k, s\in S \qquad (20)$$

$$X\geq 0, INT;\; y^{start}, y^{finish}, z^{start}, z^{finish}\geq 0$$

The objective function (1) minimizes a weighted average measure of provider idle time and patient wait time across all scenarios. Note that provider idle time is calculated as the finish time of the last patient minus the sum of the service times of all patients with provider *k* under each scenario. The wait time in the lobby is the difference between the patient's start time with nurse and the appointment time. The wait time in the exam room is calculated as the sum of the differences of the patients' start times with provider and finish times at the nurse step. Constraints (2-6) initialize the start time with nurses for the first two patients, and set the 0th patient finish time with provider *k* to be zero in every scenario. Constraint (7) makes sure that patient 3 is seen by the earliest available nurse, by comparing the finish times of the first two patients with nurses. Constraint (8) calculates the maximum finish time of the first two patients with nurses. Similarly, Constraint (9) keeps track of maximum finish time with nurse for patients 1 to patient *i*-1. The max value for patients 1 through *i*-2 is used to compare with the finish time of patient *i*-1 with nurses in constraint (10). This makes sure that the earliest available nurse is scheduled to take care of the subsequent patient *i*. Constraint (11) calculates the finish time of patient *i* with nurse, as the start time plus the service time with nurse. Constraint (12) ensures that a nurse can only see a patient after the patients' appointment time (recall that patients arrive punctually; they are not available any earlier or later than their appointment time). Constraint (13) makes sure that the nurse finish time matches the finish time with nurse of the corresponding patient *j* in provider *k*'s panel under scenarios *s*. Constraints (15 and 16) track the maximum of the nurse finish times of the first *j*-1 patients scheduled from provider *k*'s panel, and this max value is recursively updated in constraint (17). Constraints (14 and 17) ensure that each provider *k* serves the patient *j* who finishes the nurse step earlier; this is done by comparing the nurse finish times of the first *j*+1 patients in provider *k*'s panel, to account for possible crossover. Constraint (18) ensures the start time of the last patient seen by provider *k* is no sooner than the finish time with nurse for all the patients in the panel. Constraint (19) calculates the finish time of patient *j* which is start time plus service time with provider *k*. Constraint (20) ensures that provider *k* starts to examine the *j*th patient after seeing the *j*-1th patient.

The current model with min and max constraints can be efficiently solved using the following reformulation. For min constraints (7, 9, 14, and 16), we apply a big M method and introduce two sets of integer variables.

$n_{i,s}$ 1 if the earliest nurse available to see patient *i* is the one that serves patient *i*-1, that is, there is some earlier patient that is still seeing the other nurse; 0, otherwise

$p_{j,s}^{k}$ 1 if crossover occurs, that is, the *j*th patient to see provider k is the $j+1$ patient in his appointment schedule; 0, otherwise

Each of the min and max constraints is reformulated into two constraints. Constraint (10) transforms to constraints (10-1 and 10-2). Constraint (7) follows the same structure.

$$y_{i,s}^{start} \geq N_{i-1,s}^{max} - M^1 n_{i,s} \qquad \forall i \in 4..I, s \in S \tag{10-1}$$

$$y_{i,s}^{start} \geq y_{i-1,s}^{finish} - M^1\left(1 - n_{i,s}\right) \qquad \forall i \in 4..I, s \in S \tag{10-2}$$

Constraint (17) converts to constraints (17-1 and 17-2). Constrain (14) follows the same structure.

$$z_{j,s}^{k,start} \geq P_{j-1,s}^{k,max} - M^2 p_{j,s}^{k} \qquad \forall k \in K, j \in \{2..J_k - 1\}, s \in S \tag{17-1}$$

$$z_{j,s}^{k,start} \geq t_{j+1,s}^{k} - M^2\left(1 - p_{j,s}^{k}\right) \qquad \forall k \in K, j \in \{2..J_k - 1\}, s \in S \tag{17-2}$$

The max constraints (8, 9, 15 and 16) can also be reformulated into two constraints, respectively; for example, constraint (9) is substituted by the two following constraints (9-1 and 9-2). Other max constraints (8, 15, and 16) follow the same structure.

$$N_{i,s}^{max} \geq N_{i-1,s}^{max} \qquad \forall i \in 4..I, s \in S \qquad (9\text{-}1)$$

$$N_{i,s}^{max} \geq y_{i-1,s}^{finish} \qquad \forall i \in 4..I, s \in S \qquad (9\text{-}2)$$

The model with reformulated constraints is provided in Appendix A. We use the resulting integer program throughout the computational study.

5.2.1.3 Tightening of the Formulation

The proposed integer programming model is computationally challenging. The number of scenarios needs to be sufficiently high to ensure robustness of the solution; we use 1000 scenarios in our experiments. For instances with more than 5 patients per provider, the general model fails to find a guaranteed optimal solution within 4-hours of computation time. We thus, seek strategies to tighten the formulation. Specifically, we derive tight lower bounds on the big M parameters and propose stage-based bounds and additional constraints to eliminate unnecessary processing and strengthen the formulation. As we shall see in the computational section, this significantly helps reduce the computational time.

First, we tighten the big M constraints, constraints (7 and 10) with M^1 and constraints (14 and 17) with M^2. M^1 is bounded by the difference of nurse finish times of patient i+1 and the maximum of patients 1 through *i*; and M^2 is derived from the difference of nurse finish times between provider k's patient *j* and *j*+1. The following theorems provide closed form expression for the resulting tight values of M^1 and M^2, respectively. The proofs are provided in Appendix B.

Theorem 1. The value of M^1 for each patient under each scenario can be given by

$$M_{i,s}^1 = Max\{\tau_{i-1,s}^N + Max\{0,30 - \tau_{i-2,s}^N\}, Max_{r=1,\dots,i-2}\{\tau_{r,s}^N - \sum_{u=r+1}^{i-1} \tau_{u,s}^N\}\}$$

Theorem 2. The value of M^2 for patient j of provider *k* under scenario s can be provided by

$M_{j,s}^{2,k} =$

$$Max\left\{\tau_{i+2,s}^{N} + \tau_{i+1,s}^{N} + Max\{\,0, -\tau_{i,s}^{N} + 30\},\ \underset{r=1,\dots,j, r\ in\ provider's\ k\ panel}{Max}\{\tau_{r,s}^{N} - \textstyle\sum_{u=r+1}^{i+1} \tau_{u,s}^{N}\}\right\}$$

where *i* = *f[j,k]*; that is, *i* is the patient number in the overall practice schedule corresponding to the *j*th patient in provider *k*'s schedule.

In addition, we propose stage-based lower bounds (see Santos et al. 1995) for both the nurse and provider stages. At the nurse stage we derive lower bounds for the start time and finish time with nurses for each patient under each scenario *s*. At the provider stage, the finish time of the last patient with each provider *k* which is essentially session completion time of each provider. Our lower bounds are derived using constraints (7-11) to calculate the start time and finish time with nurses without consideration of the appointment times introduced in constraint (12). In other words, the earliest time a patient visit starts can be calculated recursively as the second largest value of the finish times up to patient *i*-1 at the nurse step. This provides tight lower bounds for the nurse start and finish times of patient *i* in the nurse stage and the completion time with provider *k* in the provider stage under scenario *s*.

We also introduce additional constraints to further tighten the formulation and reduce unnecessary processing. First, the appointment times can be required to be in ascending order, w.l.o.g.; that is, the appointment time of patient *i*+1 must be greater than or equal to that of patient *i*.

$$X_i \le X_{i+1} \le \cdots \le X_I, \qquad \forall i \in I \qquad (21)$$

Second, we restrict the appointment schedule to have at most one open slot (slack) between consecutive patients, both within the overall set of patients in the practice [constraint (22)] and within the patients in a provider *k*'s panel [constraint (23), with *i* and *i*+2 as consecutive patients in provider *k*'s panel]. Observe that constraint (23) does not allow for double booking within provider k's panel. This is appropriate for the complex appointments, Type HC, under consideration, as they require long service times; double-booking would highly increase patient wait time. Note also that we are assuming that all patients show up at their appointment time. When no-shows are prevalent, this constraint will be relaxed to allow for double-booking.

$$X_{i+1} - X_i \leq 2, \quad \forall i \in I \quad (22)$$

$$1 \leq X_{i+2} - X_i \leq 2, \quad \forall i \in I \quad (23)$$

Intuitively, it makes sense to set a limit of at most one open slot between consecutive patients in a provider k's schedule. According to our data, 12% of the patients take over 30 minutes (2 slots) and only 3% over 45 min. with nurse. 10% of the patients spend time with provider over 30 min. and the maximum service time is 44 min. In general, it is more advantageous to have a single slot open after each of consecutive appointments, than having two slots open in a row.

Based on our previous experiments, we may safely assume that the schedule is staggered, which means that slack or idle slots are scheduled at different times for the two providers. In addition, by symmetry w.l.o.g. we can assume that the second provider's schedule gets the earlier empty slots. Under these assumptions, constraint (22) can be tightened to

$$X_{i+1} - X_i \leq 1, \quad \forall i \in I \quad (22')$$

5.2.2 Computational Study

In our experiments, we study small instances and large instances: 10 patients (5 per provider) and 16 patients (8 per provider), respectively. Sixteen patients are a reasonable workload for a 4-hour morning or afternoon session in primary care. In the objective function, we use coefficients of 0.8 for idle time and 0.2 for wait time since we find these weights align best with the desired performance of the practice (see Chapter 4 for sensitivity analysis of coefficients).

In the small instances, we jointly optimize the appointment times of the 10 patients seeing the providers. First, we compare the computational performance of models with and without tightening constraints. Then, we study schedules generated by the model and propose scheduling guidelines. In addition, we analyze the sensitivity of the schedules to no-show probabilities to indentify robust scheduling strategies. In the large instances, due to significantly high computational time, we consider a “divide and conquer” approach: we optimize the schedule of one provider while keeping the schedule of the other provider fixed. For example, we use the optimal schedule we have learned from small instances and also the optimal schedule provided by a single-provider practice model for provider 1. Finding optimal appointment times for provider 2’s patients is equivalent to determining the location of slack in provider 2’s schedule when provider 1’s schedule is fixed.

5.2.2.1 Scheduling Approach for Small Instances

5.2.2.1.1 Computational Performance

With consideration of five patients per provider, the optimal appointment times are determined for both providers. First, we evaluate computational performance of the model, with and without tightening constraints. In the model without tightening constraints, we still apply the big M method. However, we use a simple, large M which is the average completion time under scenarios generated by the model of a single-provider practice because the difference of nurse finish times between successive patients (left side of the big M constraints) will never be greater than the completion time.

In evaluating the computational performance of various approaches, we report the *optimality gap*, which can be defined as the relative gap between the objective of the best integer solution and the objective of the best node remaining generated by CPLEX. Our model is implemented with IBM ILOG Optimization Programming Language using CPLEX 12.6 and run on a Windows 8.1 pro and 64 bit with Intel(R) Core™ i7-4770 CPU @ 3.40 GHz, 3401 Mhz, and 32GB RAM. We generate two replications of 1000 scenarios by randomly sampling from the empirical service time distribution. We allow 14,400 CPU seconds. The model contains 118,002 constraints, 15,000 binary variables, and 10 integer variables. Table 5.1 and 5.2 present the optimality gap for the various models with and without tightening formulations after 1 hour and 4 hour run times, respectively.

Table 5.1: Computational performance for models with and without tightening constraints with allowance of 1 hour

Gap	Model with large M	Model with tight M	Model with large M & bounds	Model with tight M & bounds
1st replication	46.93%	14.02%	1.46%	1.05%
2nd replication	59.80%	15.59%	1.54%	1.34%

Table 5.2: Computational performance for models with and without tightening constraints with allowance of 4 hours

Gap	Model with large M	Model with tight M	Model with large M & bounds	Model with tight M & bounds
1st replication	28.52%	10.54%	1.27%	0.91%
2nd replication	32.16%	10.74%	1.32%	1.13%

As shown in Table 5.1 and 5.2, the gap significantly decreases when incorporating tight M values and bounds, with the bounds narrowing the optimality gap far more quickly. It is interesting to note that when running the model for 4 hours, all models produce the same objectives and schedules. However, we cannot confirm the quality of the solution produced by the formulation without any tightening bounds or tight M. Due to the time limit, the search process has not been completed to guarantee optimality; however, the best integer solution has not been improved after a certain time. The significant computational effort shows that "one of the incumbents found in the first minutes of the branch and bound process was indeed the best solution that was to be found (Topaloglu 2006)." The objectives and schedules obtained by the model satisfy the goal of the study from the practical viewpoint.

Next, we investigate the computational performance of the tightened formulation in Figure 5.5.

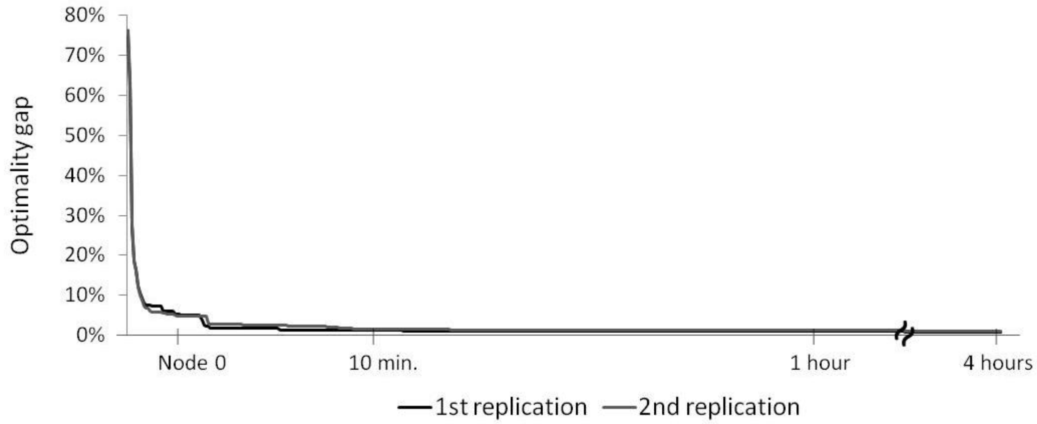


Figure 5.5: Computational performance of tightened formulation

As shown in Figure 5.5, at the end of node 0, the gap reaches close to 5.24% in 62 seconds in the 1st replication and 4.81% in 70 seconds in the 2nd replication. Within 10 mins, the gap is 1.2% in the 1st replication and 1.7% in the 2nd replication. The objective after 10 min. is only 0.03% and 0.2% lower than that after 4hours, respectively. Therefore, the formulation can guarantee a near optimal solution very quickly.

5.2.2.1.2 Schedule Comparison

In our effort to derive scheduling guidelines, we compare three schedules: practice policy schedule, identical schedule, and staggered schedule. The *practice policy schedule* follows the scheduling rules of the practice that inspired our study. Their policy is to book a HC appointment in two 15-min slots, as they regard HC appointments as 30-min appointments – in other words, a

15-min slack is placed after every HC appointment. The *identical schedule* is determined by the solution of our model with an additional constraint (24) which makes sure that both providers have identical schedules.

$$X_i = X_{i+1}, \qquad \forall i \in 1,3,5 \qquad (24)$$

Last, the *staggered schedule* optimizes appointment times for both providers with a constraint imposing that open slots are never placed at the same appointment slot for both providers. Figure 5.6 displays the schedules of practice, identical, and staggered policies.

	Practice Policy		**Identical**		**Staggered**	
Time	PCP 1	PCP 2	PCP 1	PCP 2	PCP 1	PCP 2
0:00	■	■	■	■	■	■
0:15			■	■	■	■
0:30	■	■	■	■	■	
0:45						■
1:00	■	■	■	■	■	■
1:15			■	■	■	
1:30	■	■				■
1:45						
2:00	■	■				

Figure 5.6: Schedules of practice, identical, and staggered policies for small instances

As shown in Figure 5.6, the identical schedule consists of three appointments followed by slack and two appointments. The first three appointments are consecutively scheduled since the wait time and idle time have not accumulated yet. In the staggered schedule, the schedule of provider 1 follows the identical schedule while the schedule of provider 2 assigns slack after two appointments; staggering in this fashion allows a steadier flow into the flexible nurse step. In general, our model suggests to schedule two HC appointments followed by slack. This schedule maintains the similar scheduling structure we have proposed in Chapter 4. In addition, the model gives the nurse priority to the patients of the busier provider. Since our model gives priority to the patients of provider 1 over provider 2 given the same appointment times, the schedule for provider 1 is packed.

Next, we discuss the performance of wait time, idle time and completion time for the first replication (i.e. 1000 scenarios) we run. The second replication leads to similar results. The identical schedule provides about 25% better objective value and 45% better idle time compared to the practice policy, on average over the 1000 scenarios. In the practice policy, however, the wait time performance is significantly better (277%) while the average idle time is more than one hour with only five patients per provider. The practice schedule introduces more than enough slack, which causes very low wait times but unsustainably high idle times.

Comparing the identical and the staggered schedules, the objective difference is 2% which does not seem significant. It is because although the staggered schedule improves 17% on wait time, the idle time decreases 5% compared to the identical schedule. But the decrease of idle time is essentially only 2 minutes summed across all patients.

Next, we display the performance of the practice as the session unfolds, for each of the ten patients in the sequence. Figures 5.7 and 5.8 show the wait time per patient and idle time between patients for all three schedules.

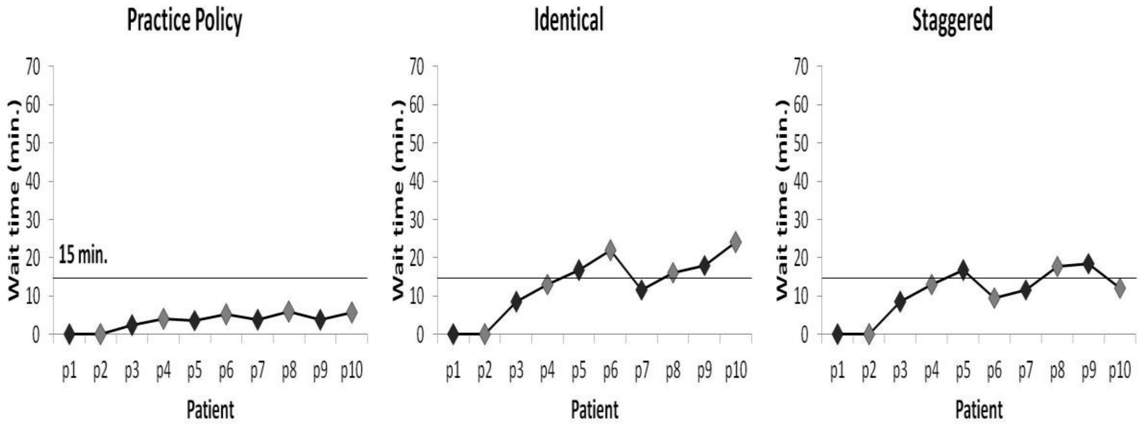


Figure 5.7: Average wait time per patient

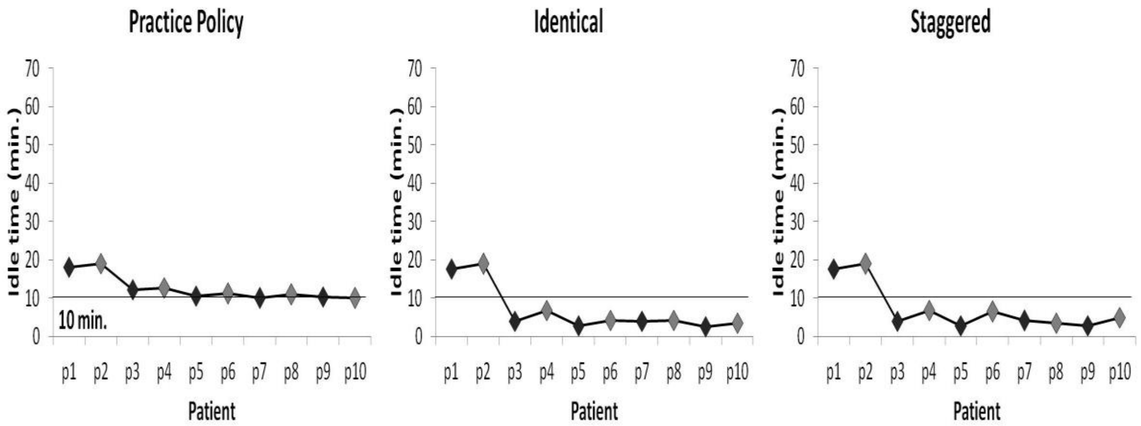


Figure 5.8: Average provider idle time between patients

Figure 5.7 shows that the wait time per patient followed by the practice policy is way below the 15-min. line but providers go idle more than 10 minutes per patient, on average. It is because unneeded slack is scheduled, which results in inefficient performance. In the identical and staggered schedules, the wait time accumulates and then drops down where slack has been added. The patient wait time of the staggered schedule stays consistently around the 15-min line. Thus, patients in the staggered schedule experience significantly less wait time than those in the identical schedule: three patients wait slightly more than 15 min. in the staggered while five patients wait more than 15 min in the identical.

The idle time of both the identical and the staggered schedules (Figure 5.8) is in a fairly similar range and much less than 10-min. per patient after the very first two patients. Note that the providers idle times after the first two patients are essentially nurse service times, since providers need to wait for patients to see nurses first in the first appointment slot. Next, we study the 90th percentile of wait time per patient and idle time between patients for the three schedules, to see how they fair in the "worst case".

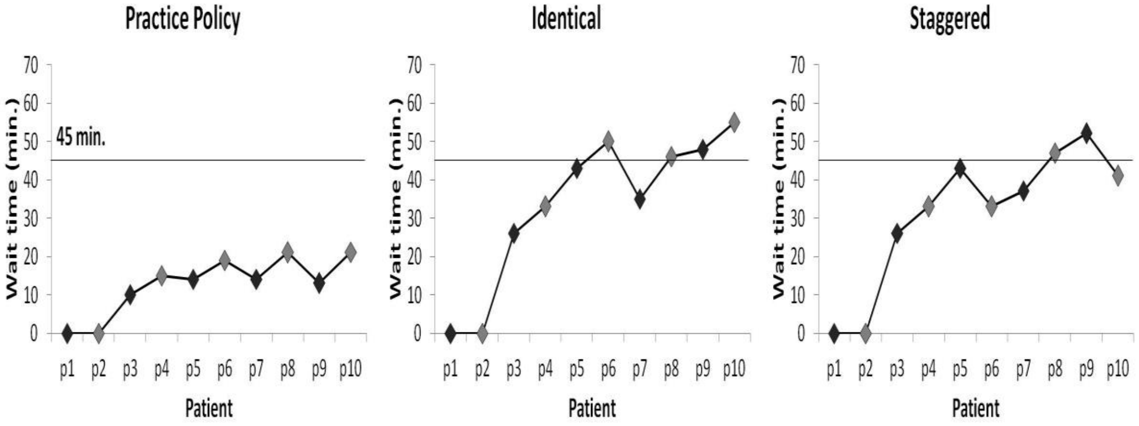


Figure 5.9: 90th percentile of wait time per patient

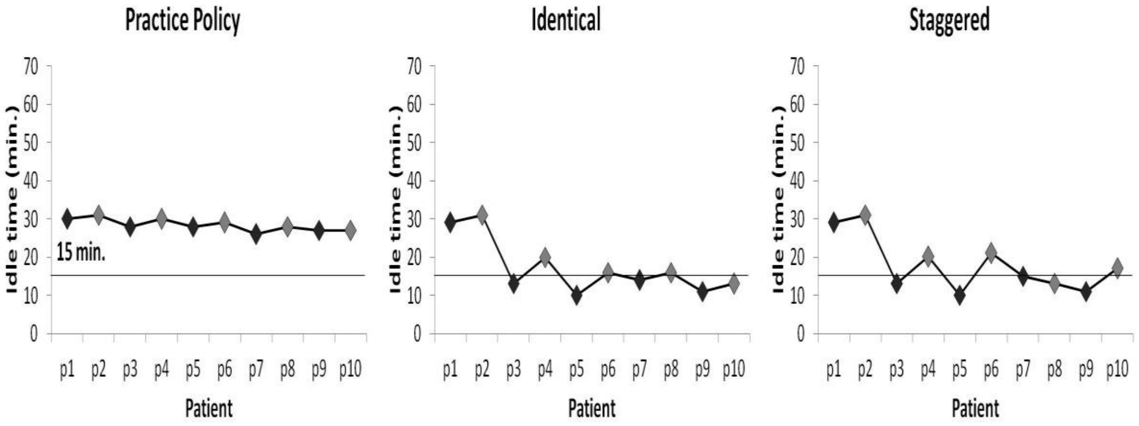


Figure 5.10: 90th percentile provider idle time between patients

Figure 5.9 and 5.10 show that each patient's wait time in the practice policy is considerably below the wait times associated with the identical and staggered policies. On the other hand, the provider idle time in the practice policy is almost twice higher than that of the identical and staggered. Comparing wait time between the identical and the staggered policies, only two patients in the staggered schedule wait more than 45 min. while four patients spend more than 45 min. to wait in the identical schedule. Therefore, the staggered schedule performs fairly well; in particular, in wait time per patient.

Next, we compare the joint performance in wait time and idle time between two single-provider practices and a 2-flexible-nurse, 2-provider team practice. Figure 5.11 displays the 90th percentile of wait time and idle time for the two practices under the various scheduling policies.

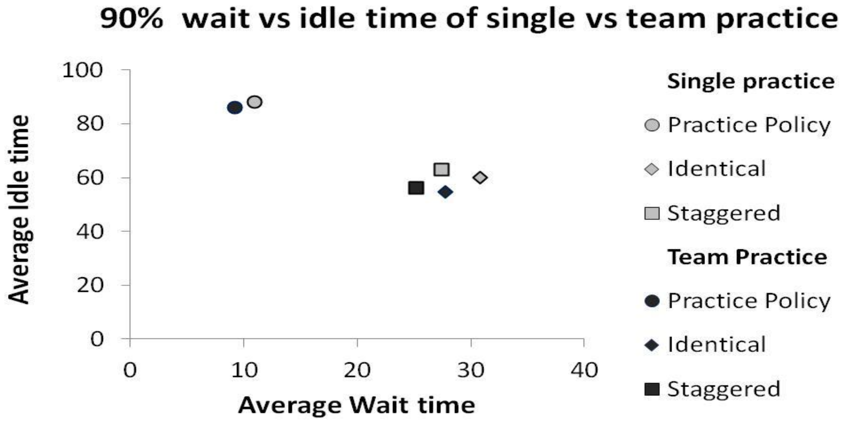


Figure 5.11: 90th percentile of wait time and idle time of single practice vs. team practice

Figure 5.11 illustrates that the wait time and idle time performance of schedules in team practices dominates that in single-provider practices. Thus, allowing flexible nurses and patient crossover has a significant impact on operational performance.

In summary, we derive the following guidelines: 1) team practices are better off staggering slack slots rather than locating them identically in both providers' schedules; 2) two

HC appointments should be followed by a slack slot, except perhaps in the first sequence of the session for only one of the providers; 3) no double-booking for a provider in the absence of no-shows, since HC appointments have long and highly variable service times; and 4) the patients of the busier provider should have priority in the nurse step given same appointment times.

5.2.2.1.3 Sensitivity to the number of scenarios

In this section, we assess the impact of reducing the number of scenarios considered in the stochastic program. This is of great interest since the scheduling problem is computationally very challenging; a smaller but reasonable number scenarios could result in a robust schedule at a much lower computational burden. First, we run 10 replications with 50 scenarios in each replication. Although the same general scheduling guidelines stated above still apply, there is variation in the optimal schedules from one replication to the next. However, when we use 100 scenarios and 10 replications, the integer program suggests only two different optimal schedules: staggered schedule 1 which occurs in four of the ten replications; and staggered schedule 2 which occurs in six out of the ten replications. Figure 5.12 presents these two schedules: staggered schedule 1 and staggered schedule 2.

	Staggered 1		**Staggered 2**	
Time	PCP 1	PCP 2	PCP 1	PCP 2
0:00	■	■	■	■
0:15	■	■	■	■
0:30	■		■	
0:45		■	■	■
1:00	■	■		■
1:15	■		■	
1:30		■		■
1:45				
2:00				

Figure 5.12: Staggered schedule1 and staggered schedule 2 are the only two optimal schedules generated in solving the integer model over 10 replications of 100 scenarios

The staggered 1 in Figure 5.12 looks exactly same as the staggered schedule using 1000 scenarios we saw in Figure 5.6. Despite the slight difference between the two schedules in Figure 5.12, notice that the scheduling guidelines we previously mentioned still hold. The objective between two schedules differs a mere 1.2% among replications.

To compare their performance, we simulate these two schedules over 1000 scenarios. The objective difference between staggered 1 and staggered 2 falls further to only 0.2%. The performance of the staggered 1 is 8% better with wait time and 3% worse with idle time than the staggered 2. Overall, the staggered schedule 1 performs slightly better than the staggered schedule 2. We conclude that the model with 100 scenarios results in fairly robust schedules.

5.2.2.1.4 Sensitivity to no-show rates

Until now, we assume the no-show rate is 0% since the practice that inspired our study has only 3% patient no-show rate. In this section, we study the performance of our models for various no-show rates. We consider no-show rates ranging from 5 to 30 percent in increments of 5 percent. According to Cayirli and Veral (2003), no-show rates of primary care practices commonly fall into that range. The method to model the different no-show rates within our stochastic programming formulation is to randomly place zero-length visit durations with nurse and provider in the data used to generate the scenarios. We again allow a maximum running time of 4 hours.

As previous sections, we optimize appointment times applying the model with the tightened formulation, except that we use a modified constraint (23') instead of (23). While double-booking was not beneficial (and thus not allowed by our formulation) under the assumption of all patients meeting their appointments, it becomes an attractive policy when no-shows are possible. We use the following constraint (23') to ensure that double-booking is feasible in each provider's schedule.

$$0 \leq X_{i+2} - X_i \leq 2, \quad \forall i \in I \qquad (23')$$

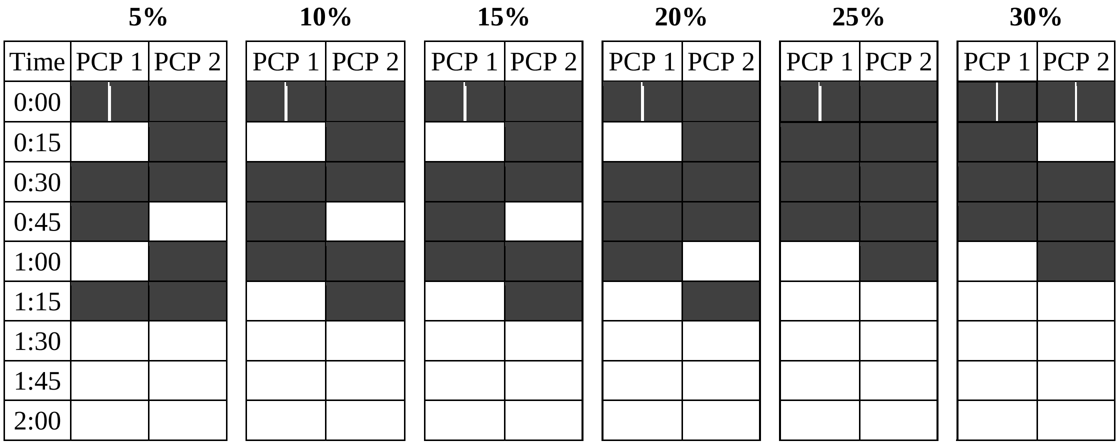
Due to the weaker constraint and highly variable service times experienced with no-shows in the patient mix, the model produces approximately 5% gap under all no-show rates. To improve computational performance, we further restrict the formulation and consider four special cases that cover all the potential optimal solutions to the original unrestricted model. Note that the unrestricted model is the tightened formulation with constraint (23') discussed above. Case 1 imposes a double–booking for the very first two appointments of provider1 and no double-booking for provider 2, by fixing $X_1 = X_2 = X_3 = 0$ and $X_4 = 1$. Case 2, considers double-bookings for the first two patients of both provider 1 and provider 2; we use the same formulation as Case1 except $X_4 = 0$. As in Case 2, Case 3 imposes double booking at the beginning of both provider's schedules, but unlike Case 2, it allows to book slack on the same appointment slots for both providers by applying constraint (22) instead of constraint (22'). Finally, Case 4 does not allow any double-booking. To solve the problem more effectively in Case 4, we add the following constraints (25 and 26):

$$1 \leq \mathrm{X_i} - \mathrm{X_{i-1}} \leq 2, \qquad \forall \mathrm{i} \in 3,5,7,9 \qquad (25)$$

$$\mathrm{X_i} - \mathrm{X_{i-1}} \leq 1, \qquad \forall \mathrm{i} \in 4,6,8,10 \qquad (26)$$

Double-booking the first two patients of just one of the providers (Case1) results in better schedule performance than double-booking the first two patients of both providers (Case 2 and 3), up until a no-show rate of 25%. For a 30% no-show rate, double-booking both providers results in slightly better (0.1%) performance. The optimality gap for all cases and no-show rates is 2% or less.

The objective differences between imposing double-booking for one provider (Case 1) and no double-booking (Case 4) are 0.4%, 1.5%, 2.5%, 2.6%, 3.5%, and 5.2% for no-show rates of 5%, 10%, 15%, 20%, 25%, and 30%, respectively. As expected, the difference increases as the no-show rate rises. Figure 5.13 displays the schedule with the best performance among cases, under different no-show rates.



	5%		10%		15%		20%		25%		30%	
Time	PCP 1	PCP 2	PCP 1	PCP 2	PCP 1	PCP 2	PCP 1	PCP 2	PCP 1	PCP 2	PCP 1	PCP 2
0:00	■|■	■	■|■	■	■|■	■	■|■	■	■|■	■	■|■	■|■
0:15		■		■		■		■	■	■	■	
0:30	■	■	■	■	■	■	■	■	■	■	■	■
0:45	■		■		■		■	■	■	■	■	■
1:00		■	■	■	■	■	■			■		■
1:15	■	■		■		■		■				
1:30												
1:45												
2:00												

Figure 5.13: Schedules under different no-show rates

As expected in Figure 5.13, the schedule gets packed when no-show rates increase. With a 25 percent no-show rate, slack is no longer needed in the schedule, even when double-booking the first two patients of one of the providers. The optimal schedule under 30% no-shows includes one open slot (slack) since both providers are double-booked at the beginning of the session. Double booking the first two patients of provider 1 is a robust scheduling guideline in the range of 5-30% no-shows. The double-booking is followed by an open slot for no-show rates in the range of 5-20% no-shows; no slack is necessary under 30% no-shows. It is also interesting to note that the slack position is pushed down, to later in the schedule, as the no-show rates increase. Because of no-shows, wait time and idle time accumulates at a slower pace. In addition, although our formulation allows double-booking any two consecutive patients, the optimal solutions generated only suggest double-booking the very first two appointments.

5.2.2.2 Scheduling Approach for Large Instances

In this section, we study moderately large instances considering 16 patients: eight patients per provider. Due to the high computation times, we use a heuristic solution method: we fix the schedule of provider 1 as the optimal schedule we learned from the single-provider practice and from our experiments with small instances; we then optimize the schedule of provider 2. As a reminder, the common scheduling guideline we obtained from the single-provider practice and small instances is to book slack after two successive HC appointments, except at the beginning of the session when three appointments are scheduled in a row. We apply this scheduling rule to create a fixed schedule for provider 1.

We run two cases: case 1 is to allow slack to be booked simultaneously, on the same time slot, for both providers; and case 2 is to impose a staggered schedule, which was shown to be optimal in the experiments with small instances. The optimality gap in case 1 is 0.71% with 4 hours running time; however, the gap in case 2 is 0% within 10 seconds running time. The schedule and objective function value of these two cases are the same.

As in the previous section, we compare the practice policy, identical, and staggered schedules. The identical schedule is fully predetermined by fixing provider 1's schedule using our scheduling guidelines. The staggered schedule is optimized using the Case 2 described above. Figure 5.14 shows the practice, identical, and staggered schedules.

	Practice Policy		**Identical**		**Staggered**	
Time	PCP 1	PCP 2	PCP 1	PCP 2	PCP 1	PCP 2
0:00	■	■	■	■	■	■
0:15			■	■	■	■
0:30	■	■	■	■	■	
0:45						■
1:00	■	■	■	■	■	■
1:15			■	■	■	
1:30	■	■				■
1:45			■	■	■	
2:00	■	■	■	■	■	■
2:15						■
2:30	■	■	■	■	■	
2:45						■
3:00	■	■				
3:15						
3:30	■	■				

Figure 5.14: Practice, benchmark, and staggered schedules for large instances

As shown in Figure 5.14, the scheduling guidelines derived for small instances hold for these moderately large instances as well. On average, the practice schedule performs much better in wait time, about 270%; but has a 26% worse objective value and a 52% worse idle time performance when compared to the identical schedule. In the practice policy, the wait time is around 4 minutes per patient yet the total idle time for each provider is 90 minutes when seeing 8 patients. The staggered schedule results in a 3% objective improvement relative to the identical schedule, with a 20% improvement in wait time and a 9% increase in idle time.

In sum, our computational experiments for large instances show that the scheduling guidelines derived for small instances are robust.

5.3 Multiple Appointment Types

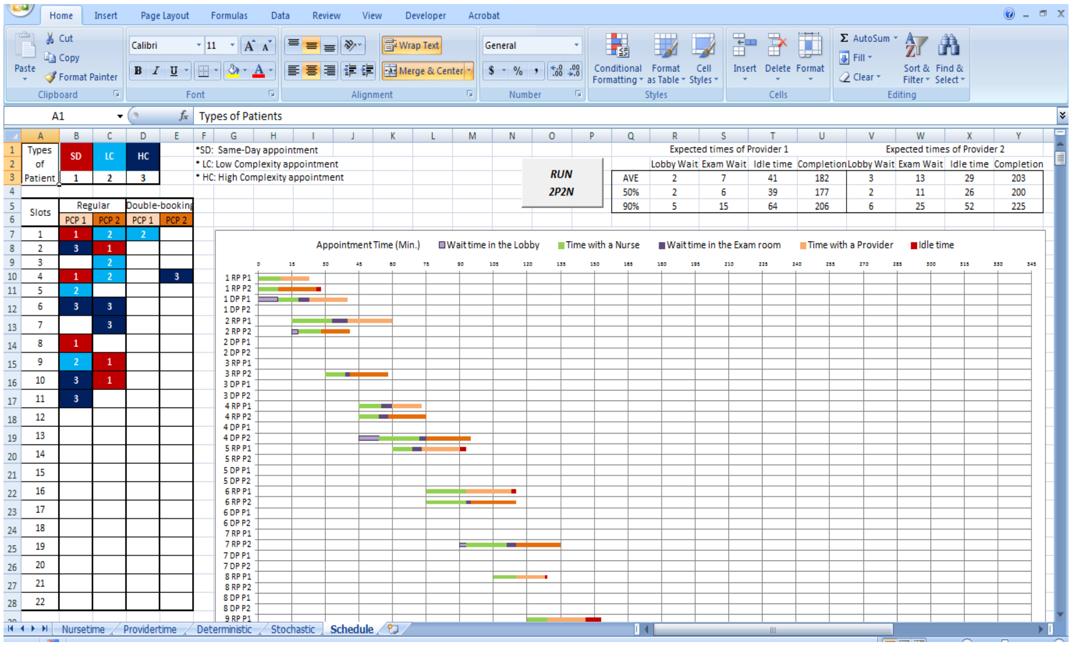
While we were principally interested in the structure of appointment times and sequences, the stochastic optimization program can also be used in a *dynamic* sense. This is important because schedules are not constructed all at once but as calls come in, one at a time. As a companion to the models presented in the previous chapters, we have developed a practical Excel simulation tool that allows the practice to explore the performance of different schedules in real time as patients call in. The scheduler can dynamically insert new patients into the schedule and obtain the expected performance based on 1000 scenarios randomly sampled from the empirical data. The measures provided are: wait time (total as well as by patient position in the sequence), idle time, and finish time. We provide the capability for measuring averages, 50th and 90th percentiles for the current partial schedule. A Gantt chart in the spreadsheet allows the scheduler to visualize how the appointments are staggered. Double booking of slots is allowed in the tool.

5.3.1 Excel Scheduling Tool: Motivation and Key Features

In Chapter 4, we focus on practices where one nurse and one provider work as a medical team. We refer to such practices as a single-provider primary care practice, or we call *dedicated nurse practices* in this section since a nurse exclusively takes care of patients in the panel of the particular provider. As we explained in the previous section, however, we have often observed two nurses flexibly sharing the patients of a two-provider team; we call these *flexible nurse practices*. Each provider still keeps her/his own panel of appointments and can choose to see her/his patients according to the original appointment schedule or, more commonly, in the order in which they complete the nurse step (that is, *allowing patient schedule crossover*). Thus, we refer to the practices allowing both flexible nurse and patient crossover as team primary care practices in previous section. In this section, therefore, we consider three cases: 1) dedicated nurses, 2) flexible nurses, and 3) flexible nurses & crossover. Therefore, we account for the following factors in an Excel simulation tool: 1) patient classification into three well-differentiated patient types – HC, LC, and SD; 2) stochastic service times for both nurse and provider; 3) potential patient sharing by nurses; and 4) no sharing of appointments between providers.

Our goal is to provide an Excel simulation tool that allows the scheduler in the practice to explore the performance of different schedules in real time. The stochastic performance of the schedule can be thus assessed dynamically as patients requests arise. As a case-study, we compare the performance of *dedicated* versus *flexible nurse practices* in a primary care setting with two providers and two nurses. We use actual schedules observed in practice, optimal schedules, and heuristic schedules from Chapter 4. A preliminary version of the Excel simulation tool is available at “UMass blog by Oh”

The Excel tool contains five different spreadsheets: nurse time, provider time, deterministic, stochastic, and schedule. We describe them in detail below. Figure 5.15 shows the snap shot of the Excel simulation tool.



Types of Patient	SD	LC	HC
	1	2	3

*SD: Same-Day appointment
* LC: Low Complexity appointment
* HC: High Complexity appointment

Slots	Regular		Double-booking	
	PCP 1	PCP 2	PCP 1	PCP 2
1	1	2	2	
2	3	1		
3		2		
4	1	2		3
5	2			
6	3	3		
7		3		
8	1			
9	2	1		
10	3	1		
11	3			
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				

	Expected times of Provider 1				Expected times of Provider 2			
	Lobby Wait	Exam Wait	Idle time	Completion	Lobby Wait	Exam Wait	Idle time	Completion
AVE	2	7	41	182	3	13	29	203
50%	2	6	39	177	2	11	26	200
90%	5	15	64	206	6	25	52	225

Figure 5.15: Snap shot of Excel simulation tool. On the left, colored slots indicate provider calendars. A Gantt chart on the right indicates how the schedule will play out in practice; this is intended as a visual aid to the scheduler

The *nurse time* spreadsheet includes 1000 scenarios of service time with a nurse for each patient type, randomly sampled from the empirical study. The *provider time* spreadsheet includes 1000 scenarios of service time with each provider for each patient type, randomly sampled from the empirical study.

The *deterministic* spreadsheet contains the average service times of nurse and provider steps for the different patient types. This information is linked to the Gantt chart in the *schedule* spreadsheet, which shows the scheduler how the schedule would fare under average service times. The *stochastic* spreadsheet uses service time data from the *nurse time* and *provider time* spreadsheets. All patient flow indicators are calculated by algorithms coded in visual basic for applications (VBA) in EXCEL 2007.

The algorithms for the dedicated nurse practices in VBA are based on the stochastic integer programming model in Chapter 4. We use the appointment time, start time with nurse/provider, and finish time with nurse/provider for each patient to calculate wait time in the lobby (start time with nurse minus appointment time), wait time in the exam room (start time with provider minus finish time with nurse), and idle time (session completion time minus service time of all patients with a provider). The appointment times are given in 15 min. slots, as is the case in the family care practice that inspired this study. This is a trivial calculation in the case of dedicated nurses. In the case of flexible nurse practices, patients see the nurse that first becomes available. The algorithms for this case in VBA are based on the stochastic integer programming model from the previous section. The time when a nurse becomes available for patient *i* can be calculated recursively as the second largest value of the finish times of earlier patients, 1 to *i*-1, with the nurses. Similarly, providers will see the patient from their panel that finishes the nurse step earlier. The time at which the provider's j^{th} patient is ready can again be calculated recursively, using the second largest logic (now applied to the finish times with nurses of the earlier patients).

In addition, the *stochastic* spreadsheet links to the performance table in the *schedule* spreadsheet.

The scheduler needs to use the *schedule* spreadsheet to input the desired schedule and click the run box button to get the associated performance estimates. A *schedule* spreadsheet consists of three parts:

1) Schedule management: consisting of two columns for each provider to input both regular booking and double booking appointments with the three easy-to-identify appointment types proposed in Chapter 4. The appointment types are denoted by different numbers and colors: same-day (*SD*) appointment – 1 and red; Low Complexity (*LC*) appointment – 2 and light blue; and High Complexity (*HC*) appointment – 3 and dark blue. If an additional patient needs to be

assigned to an appointment slot already filled, double-booking occurs and the patient types can be scheduled in the double booking columns.

2) Gantt chart: allowing the scheduler to visualize how the appointments are staggered. The Gantt chart provides six indicators of patient flow with different color codes (assuming 15 min. slot length): wait time in the lobby - light purple; time with nurse - green; wait time in the exam room - dark purple; time with provider - orange; and idle time - red.

3) Performance: including the average, 50th and 90th percentiles of lobby wait, exam wait, idle time and completion time derived using the results of a simulation of 1000 scenarios from the *stochastic* spreadsheet. Note that when the scheduler plugs number indicators of appointment types in the schedule management columns, the *stochastic* spreadsheet populates the appropriate data from the *nurse time* and *provider time* data spreadsheets.

5.3.2 Case Study

5.3.2.1 Schedule Comparison

To illustrate the use of the Excel tool, we use schedules studied in Chapter 4: schedules observed in the practice, optimal schedules generated by the Deterministic Integer Program (DIP) and the Stochastic Integer Program (SIP), and heuristic schedules. While we use these as examples, note that a practice can choose to evaluate any schedule it likes. We run all possible schedule combinations and patient mix which we observed from the practice. Figure 5.16 displays, for instance, one provider uses one of the heuristic schedules, 1) SD/LC/HC schedule and another provider employs the DIP optimal schedule: 1 + DIP. Also, Figure 5.16 shows the appointment mix from one of the afternoons: ten patients (3 SD patients, 3 LC patients, and 4 HC patients) for provider 1 and nine patients (just one less HC patients than provider 1) for provider 2. Since our study is inspired by the three-provide family medicine practice, we use six instances, or combinations of three providers' data, for example, data of provider1 and provider2; provider1 and provider3; and so on.

1) SD/LC/HC Optimal schedule by DIP

Time	Provider 1		Provider 2	
	Regular booking	Double booking	Regular booking	Double booking
0:00	**1**	**2**	**2**	
0:15	**3**		**1**	
0:30			**2**	
0:45	**1**		**2**	**3**
1:00	**2**			
1:15	**3**		**3**	
1:30			**3**	
1:45	**1**			
2:00	**2**		**1**	
2:15	**3**		**1**	
2:30	**3**			

Figure 5.16: An example of schedule combination: 1) SD/LC/HC and optimal schedule by DIP and appointment mix from a particular afternoon

Figure 5.17 compares the performance of a variety of schedule combinations under dedicated and flexible nurse settings with six instances. The performance is measured by the weighed combination of provider idle time and patient wait time; the weight on idle time is 0.8 and wait time is 0.2. This weight combination was shown to be appropriate in our previous work. Again, a practice is free to use weights that are better suited to its operations.

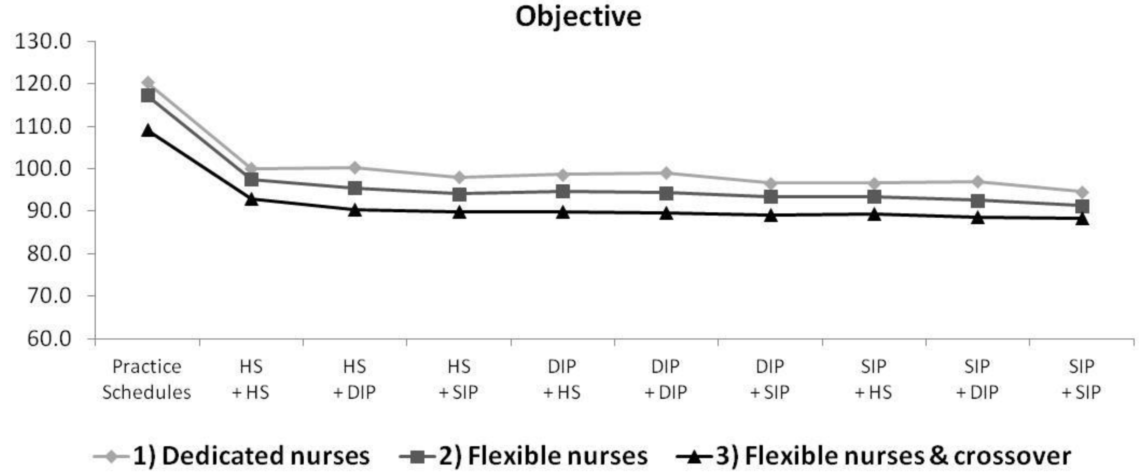


Figure 5.17: Objective performances among practice schedule and combinations of optimal and heuristic schedules among 1) dedicated nurses, 2) flexible nurses, and 3) flexible nurses & crossover on average of six cases

As shown in Figure 5.17, each flexible practice yields a 4% and 8% improvement in performance relative to the dedicated practice on average of six instances, respectively. Observe that schedules using SIP for at least one provider tend to slightly outperform the others in 1) dedicated nurses, 2) flexible nurses and 3) flexible nurses and patient crossover. The performance of the combinations of optimal and heuristic schedules we generated with a single provider is 19%, 20%, and 18% better on average than that of the practice schedule in dedicated and flexible practice settings, respectively. To provide more details, Figure 5.18 illustrates average performance of each component of patient flow: wait time in the lobby per patient, wait time in the exam room per patient, and idle time per provider (average idle time of providers), on average of six instances.

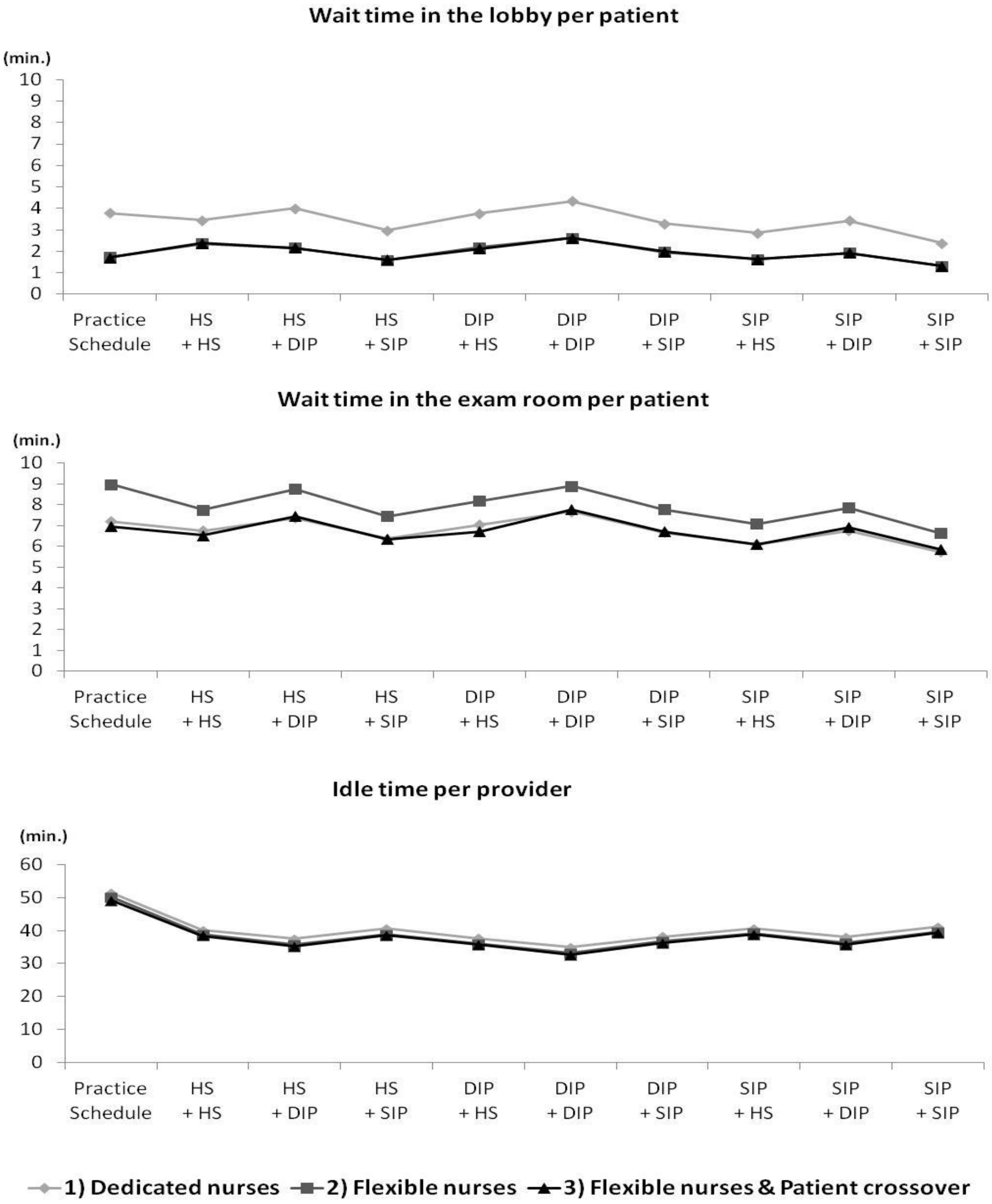


Figure 5.18: Performance of wait time in the lobby and exam room per patient, and average idle time of providers among schedules between 1) dedicated nurses, 2) flexible nurses, and 3) flexible nurses & crossover, on average of six instances

As expected (shown in Figure 5.18), having flexible nurses significantly improves the wait time in the lobby, by 42%, except when the practice schedules are used. With 2) flexible nurses, unfortunately, these time savings are partially washed away by an increase exam room wait by 16% on average of the schedules. As a result, the total patient wait time of flexible nurse practices is 3% better than that of dedicated nurse practices. In addition to the flexible nurses, however, when the provider decides to see patients whoever finishes earlier with nurses, patient wait time in the exam room stays the same as the dedicated nurse practices. In this case, the total wait time has 14% improvement than that of the dedicated nurse practices.

In addition, provider idle time is reduced by 4% by adding flexibility in the nursing step. Along with flexible nurses, the crossover also grants 5% improvements in the idle time. Therefore, having flexible nurses and allowing the patient crossover significantly improves the practice utilization: patient wait time and provider idle time.

Next, we study the worst case, 90th percentile of idle time versus wait time in Figure 5.19. For brevity, we show only some combinations of schedules.

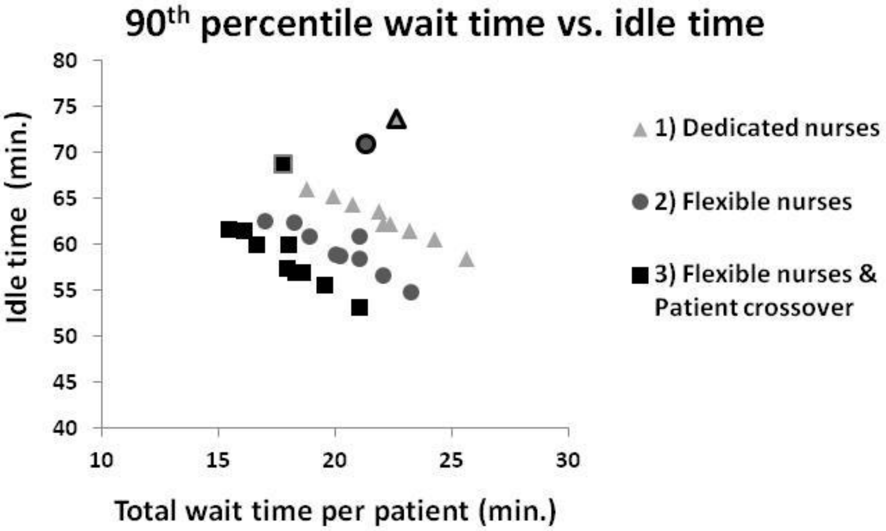


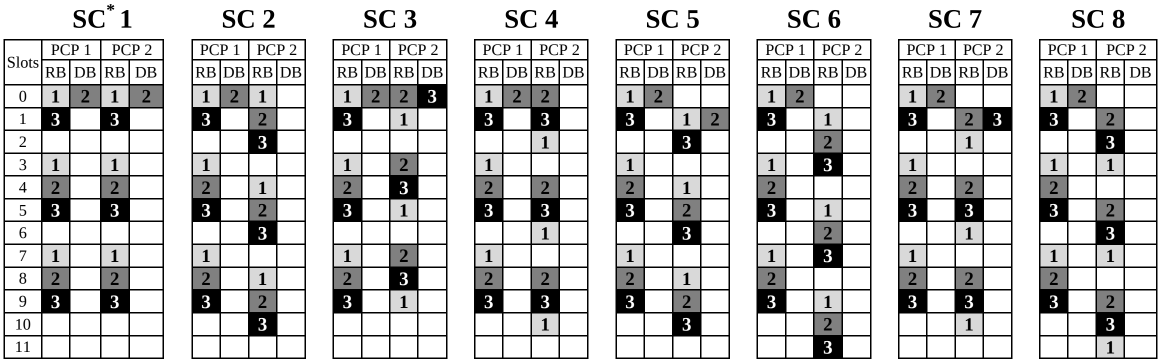
Figure 5.19: 90th percentile of idle time versus wait time among practice schedules (lined shape) and combinations of optimal and heuristic schedules (non-lined shape) between 1) dedicated nurses, 2) flexible nurses, and 3) flexible nurses & crossover, on average of

When we look at the 90th percentile of idle time versus wait time (Figure 5.19), we see that combinations of optimal and heuristic schedules have consistently better performance in both idle and wait time compared to practice schedules (lined shape). Also, optimal and heuristic schedules of 3) flexible nurses & crossover dominate those of 1) dedicated nurses and 2) flexible nurses. Most of the schedules in the efficient frontier are optimal schedules by DIP or SIP. We also study 90th percentile of session completion time; combinations of optimal and heuristic schedules improve approximately 6% when compared to the practice schedule. Therefore, allowing for flexible nurses and crossover has significant impact on the schedule performance.

5.3.2.2 Heuristic Schedule Comparison

In this section, we study the different mixes of heuristic schedules to find the robust combinations. The heuristic schedule provides time-of-day preferences for patients and also can be readily implementable in primary care practices. We use six instances, combinations of three providers' data and consider only the case of flexible nurses and patient crossover since we have verified that allowance of both flexible nurses and patient crossover significantly improves the performance of wait time and idle time in the previous section. Figure 5.20 shows the combinations of heuristic schedules.

*SC: Schedule Combination, PCP: primary care provider, RB: Regular Booking, DB: Double Booking



SC* 1

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2	1	2
1	3		3	
2				
3	1		1	
4	2		2	
5	3		3	
6				
7	1		1	
8	2		2	
9	3		3	
10				
11				

SC 2

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2	1	
1	3		2	
2			3	
3	1			
4	2		1	
5	3		2	
6			3	
7	1			
8	2		1	
9	3		2	
10			3	
11				

SC 3

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2	2	3
1	3		1	
2				
3	1		2	
4	2		3	
5	3		1	
6				
7	1		2	
8	2		3	
9	3		1	
10				
11				

SC 4

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2	2	
1	3		3	
2			1	
3	1			
4	2		2	
5	3		3	
6			1	
7	1			
8	2		2	
9	3		3	
10			1	
11				

SC 5

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2		
1	3		1	2
2			3	
3	1			
4	2		1	
5	3		2	
6			3	
7	1			
8	2		1	
9	3		2	
10			3	
11				

SC 6

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2		
1	3		1	
2			2	
3	1		3	
4	2			
5	3		1	
6			2	
7	1		3	
8	2			
9	3		1	
10			2	
11			3	

SC 7

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2		
1	3		2	3
2			1	
3	1			
4	2		2	
5	3		3	
6			1	
7	1			
8	2		2	
9	3		3	
10			1	
11				

SC 8

Slots	PCP 1		PCP 2	
	RB	DB	RB	DB
0	1	2		
1	3		2	
2			3	
3	1		1	
4	2			
5	3		2	
6			3	
7	1		1	
8	2			
9	3		2	
10			3	
11			1	

Figure 5.20: Combinations of heuristic schedules

As shown in Figure 5.20, there are eight schedule combinations (SC) based on a mix of two different sequences and slack positions between providers. For example, SC1 is the combination of the optimal heuristic schedule generated by the integer programming model studied in Chapter 4, which is SD/LC/HC sequence followed by slack and a double booking for the very first two appointments. SC2 is used the SD/LC/HC sequences for both providers, but the double booking is involved only for provider1 in order to locate slack in the different appointment slots. Since we have learned the importance of the slack position, we compare the performance when the slack locates in the same or different appointment slots between two providers. Next, we study the different sequence combinations in SC3 and SC4 - SD/LC/HC and LC/HC/SD. Again, the double booking only occurs in SC3. Note that we have examined the sequencing combination of SD/LC/HC and HC/LC/SD and found that the performance was worse than the sequence we present here since starting with the type HC appointment highly increases the idle time. Next, the sequences from SC5 to SC6 match those from SC1 to SC4. The only difference is that provider2 starts seeing patients a slot (15-min.) later than provider1, also allowing for different slack locations. Based on observation at the practice, the start time of providers can be sometimes different. We study the objective performance of schedule combinations, on average of six instances in Figure 5.21.

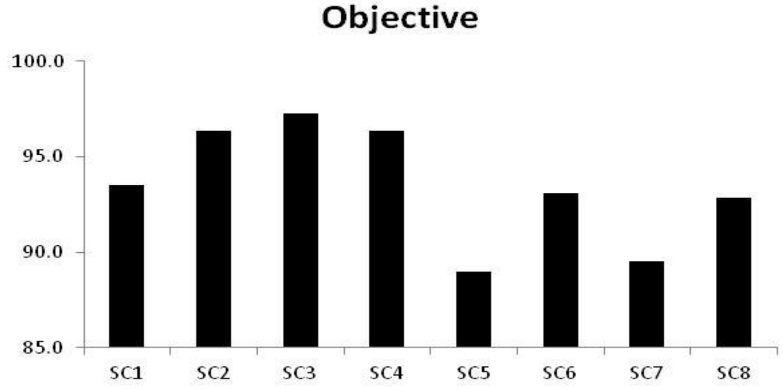


Figure 5.21: Objective of eight heuristic schedule combinations, on average of six instances

Based on the schedules in Figure 5.20 and the performance in Figure 5.21, we find schedule guidelines. First, both providers can schedule a double booking at the very first slot, but one of the providers starts one slot (15-min.) later, which allocates the slack in the different appointment slots between two providers. It is noticeably shown in Figure 5.21 that the schedules which one of providers starts to see patients a slot later (SC5 to SC8) performs 5%, 3%, 8%, and 4% better, compared to the schedules which both providers simultaneously begins (SC1 to SC 4), respectively. In addition, SC5 and SC7, which involves double booking for both providers provide significantly better performance. Second, sequences are sensitive when both providers start with a double booking. In other words, with a double booking at the beginning of the session for both providers, a mix of SD/LC/HC and SD/LC/HC performs better than a mix of SD/LC/HC and LC/HC/SD (for example. SC1 vs. SC3 and SC5 vs. SC7). However, without a double booking at the beginning for one of providers, the sequence does not significantly affect to the performance. Third, a double booking with HC appointment needs to be avoided when another provider already has a double booking. This observation is from the SC3 which provides the worst performance. It is because HC has the longest mean service time among appointment types and highly variable. Figure 5.22 displays the 90th percentile of wait time per patient and idle time per provider, on average of six instances.

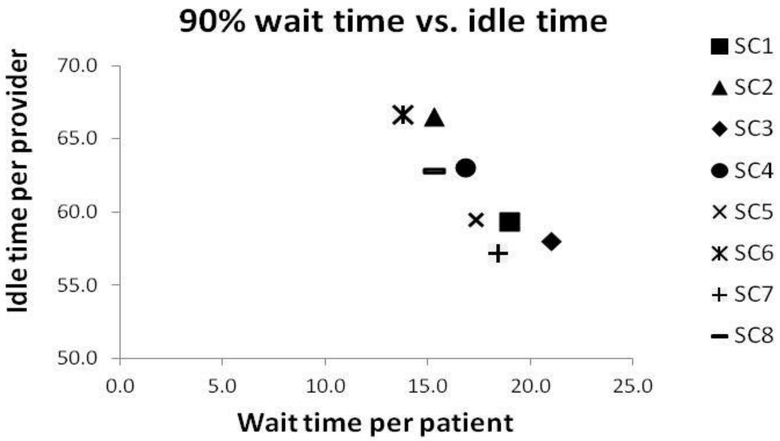


Figure 5.22: 90th percentile of wait time per patient and idle time per provider, on average of six instances

As shown in Figure 5.22, the 90th percentile of wait time and idle time also present the same results; the performance of schedules, in which one of the providers start a slot later, dominate other schedules which have the same starting time for both providers. Also, it is important to book the slack in the different appointment slots between two providers.

5.4 Conclusion

The team primary practice involves far more complex patient flows than the single-provider practice. Nurses work as a team taking care of patients of any of the providers in a flexible manner– flexible nurses. As a result, providers will see available patients from their panels according to the sequence in which they complete the nurse step, which may be different from their appointment sequence– patient crossover. In our study, thus, we consider the two sequential steps – nurse and provider, multiple resources at each step, and flexibility at each step. In addition, nurses and providers face wildly uncertain service times. All these factors compound to make the scheduling problem challenging.

We model a novel mixed integer program of the team primary care practice with the objective of minimizing a weighted measure of patient wait time and provider idle time. Since our proposed model is computationally expensive, we develop lower bounds and tightening constraints to solve the problem more effectively. We further consider special cases and additional constraints to strengthen the formulation. This significantly reduces the running time.

Our computational study shows that both an optimized schedule with identical appointment times for both providers, and an optimal staggered schedule significantly outperform the practice policy schedule. Comparing identical and staggered schedules, the staggering of slack between providers significantly reduces wait time per patient. In addition, the performance comparison of single-provider and team primary care practices shows that nurse flexibility and patient crossover in the team practice improves both patient wait time and provider idle time.

We can summarize the scheduling guidelines for team primary care practices as follows. A robust schedule for a team primary practice with flexible nurses and crossover, in the absence of no-shows, should: 1) stagger the slack of the two providers, 2) schedule a slack after every two HC appointments, except for three consecutive HC appointments at the very beginning of the session; 3) include no double-booking for any provider; and 4) give priority to the busier provider's patients with the same appointment time in the nurse step. In the presence of no-show rates in the range of 5-30%, double-booking the first two patients of one of the providers is optimal. When the no-show rate reaches 30%, both providers double-book their first two patients.

However, no double-booking occurs later in the schedule, in the sessions with 5 patients per provider tested.

This chapter focuses on a single appointment type, type HC, which includes complex patient conditions. There are practices that specialize in this type of patients, for which the models developed are fully appropriate. Other practices, however, will serve a mix of patients. We study multiple patient types using a scheduling simulation tool developed in Excel. This tool allows us to compare the single-provider and team practices considering empirical data and heuristic schedules based on our findings for the single-provider study. In the cases tested, we find that a combination schedule using the optimal schedule for one provider and a heuristic schedule for the other performs significantly better than the practice schedule. Flexible nurses and patient crossover do provide significant benefits. Practice schedules that use different heuristic schedules for each provider are also shown to perform well. We derive the following scheduling guidelines: both providers can schedule a double booking at the very first slot, but one of the providers should start the session one slot (15-min.) later; this allows the practice to allocate the slack in the schedules of the two providers on different 15-min time slots. In addition, a double booking that involves one HC appointment should be avoided in slots when the other provider has a double booking. This is because HC has the longest mean service time among appointment types, and highly variable.

The main advantage of the Excel scheduling tool is to dynamically calculate the on-going performance of the schedule as patients call in and their appointments are inserted in the schedule. The Excel simulation tool has the following features. First, by including a color-coded Gantt chart based on average service times, the tool provides a visual aid to the scheduler. Second, it allows the scheduler to dynamically test out different patient mix, patient sequence, and start time combinations in a format that resembles provider calendars. Finally, the Excel tool allows the scheduler to input random scenarios and calculate not just averages, but key percentiles of wait time, idle time, and session completion time. The Excel tool can be readily adjusted by the practices to incorporate their visit time profiles and needs. Practices can use their own patient classifications by changing the VBA code to link with data in the *nurse* and *provider* spreadsheets. They can also input their own observed service time data in *nurse* and *provider* spreadsheets.

CHPATER 6

CONCLUSIONS

Variability in primary care practices is significant: uncertain service time, different patient conditions/appointment types, multiple providers and non-provider staffs – nurse/medical assistants – and multiple stages in the patient flow. The variability makes patient flow inefficient: long wait time for patients and unnecessary provider idle time.

We collect data pertaining to the complete chronology of patient flow on nine separate workdays in a small family medicine practice in Massachusetts. Our collaborating practice is representative of primary care practices in U.S: the practice, like countless others all over the US, is small and consists of many patient conditions/appointment types that are prevalent nationally. In analyzing empirical data, we identify the inefficient components and bottlenecks in patient flow. We propose an easy-to-implement patient classification scheme: prescheduled appointments of high complexity (HC), prescheduled appointment of relatively low complexity (LC), and same-day appointments (SD). Also, we point out the importance of effective coordination between nurse and provider steps.

From the modeling perspective, we first formulate a stochastic integer program for a single-provider primary care practice, considering two sequential steps, the nurse step and the provider step, with random service times at both steps depending on patient type. Since most literature focuses on scheduling problems with only the provider stage, we compare performance and schedules suggested by our two-service-stage model with those that only consider the provider stage. We find a 21% difference on performance and significant difference in the structure of optimal schedules. Thus, it is important to include not only the provider step but also the nurse step in the model. Next, we develop a user-friendly Excel scheduling tool for schedulers to dynamically manage appointment schedules in real time, which includes more practical issues: multiple appointment types and human resources at multiple steps in the patient flow process which we call a team primary care practice. The tool can be easily customized to practice needs. While studying the performance of the team practice with the Excel simulation tool, our interest turned to the *optimal* schedule for team primary care practices. So, we formulated a novel stochastic integer programming formulation for team practices where multiple nurses can flexibly see patients while providers have their own dedicated panel appointments. We show that such a formulation can be solved in reasonable computation time while providing near optimal solutions.

From the operational perspective, we summarize the scheduling guidelines for a single-provider and a team primary care practice.

<u>Scheduling guidelines for a single-provider primary care practice:</u>

1. Different amounts of slacks in the schedule depending on the patient type - slack after two HC appointments, after four LC appointments, no slack needed in SD appointments.
2. Optimal DIP results in dome shape-like patterns and the SIP sequence is SPT-like.
3. Our easy-to-implement heuristic schedules (3AH) can provide time-of-day preferences for patients and be financially viable for the practice.
4. More double-booking and less slack is needed as no-show rates increase.

<u>Scheduling guidelines for a team primary care practice:</u>

1. Allowing flexible nurses and patient crossover (providers seeing earliest available patients after nurse steps) significantly improve wait time per patient.
2. Slack should be scheduled after two HC appointments.
3. Staggering slack: slack should be positioned in non-identical appointment slots (i.e. staggered) for the two providers.
4. With multiple appointments, both providers can double book at the very first slot (Bailey-Welch rule); one of providers should start one slot (15-min.) later, which assigns slack in different appointment slots between providers.

In conclusion, this research has developed novel mathematical programming formulations and also dealt with practical issues: factors causing high variability in patient flow; effective coordination of nurse and provider; uncertain service time dependent on patient type; three well differentiated patient conditions; and flexibility among nurses and patients. Our analysis results in easy-to-implement scheduling guidelines for primary care practices.

APPENDIX A

TEAM PRIMARY CARE PRACTICE MODEL

The model is included the reformulated constraints.

$$Min. \qquad \frac{1}{S}\left(\alpha\left[\sum_{s}\left(\left(z_{J_1,s}^{1,finish}-\sum_{j^1=1}^{J^1}\tau_{j,s}^{P_1}\right)+\left(z_{J_2,s}^{2,finish}-\sum_{j^2=1}^{J^2}\tau_{j,s}^{P_2}\right)\right)\right]\right.$$

$$+\beta\left[\sum_{s}\sum_{i=1}^{n}\left(y_{i,s}^{start}-15X_i\right)\right.$$

$$\left.\left.+\sum_{s}\left(\sum_{j=1}^{J^1}\left(z_{j,s}^{1,start}-t_{j,s}^{1}\right)+\sum_{j=1}^{J^2}\left(z_{j,s}^{2,start}-t_{j,s}^{2}\right)\right)\right]\right) \tag{1}$$

Subject to.

$$y_{1,s}^{start}=0 \qquad \forall s\in S \tag{2}$$

$$y_{2,s}^{start}=0 \qquad \forall s\in S \tag{3}$$

$$z_{0,k,s}^{finish}=0 \qquad \forall k\in K, s\in S \tag{4}$$

$$X_1=0 \tag{5}$$

$$X_2=0 \tag{6}$$

$$y_{3,s}^{start}\geq \min\left(y_{1,s}^{finish},y_{2,s}^{finish}\right) \qquad \forall s\in S \tag{7}$$

$$y_{3,s}^{start}\geq y_{1,s}^{finish}-M_{3,s}^{1}n_{3,s} \qquad \forall s\in S \tag{7-1}$$

$$y_{3,s}^{start}\geq y_{2,s}^{finish}-M_{3,s}^{1}(1-n_{3,s}) \qquad \forall s\in S \tag{7-2}$$

$$N_{3,s}^{max}\geq \max\left(y_{1,s}^{finish},y_{2,s}^{finish}\right) \qquad \forall s\in S \tag{8}$$

$$N_{3,s}^{max}\geq y_{1,s}^{finish} \qquad \forall s\in S \tag{8-1}$$

$$N_{3,s}^{max}\geq y_{2,s}^{finish} \qquad \forall s\in S \tag{8-2}$$

$$N_{i,s}^{max}\geq \max\left(N_{i-1,s}^{max},y_{i-1,s}^{finish}\right) \qquad \forall i\in 4..I, s\in S \tag{9}$$

$$N_{i,s}^{max}\geq N_{i-1,s}^{max} \qquad \forall i\in 4..I, s\in S \tag{9-1}$$

$$N_{i,s}^{max}\geq y_{i-1,s}^{finish} \qquad \forall i\in 4..I, s\in S \tag{9-2}$$

$$y_{i,s}^{start}\geq \min\left(N_{i-1,s}^{max},y_{i-1,s}^{finish}\right) \qquad \forall i\in 4..I, s\in S \tag{10}$$

$$y_{i,s}^{start}\geq N_{i-1,s}^{max}-M_{i,s}^{1}n_{i,s} \qquad \forall i\in 4..I, s\in S \tag{10-1}$$

$$y_{i,s}^{start}\geq y_{i-1,s}^{finish}-M_{i,s}^{1}(1-n_{i,s}) \qquad \forall i\in 4..I, s\in S \tag{10-2}$$

$$y_{i,s}^{finish}=y_{i,s}^{start}+\tau_{i,s}^{N} \qquad \forall i\in I, s\in S \tag{11}$$

$$y_{i,s}^{start} \geq 15X_i \qquad \forall i \in I, s \in S \tag{12}$$

$$t_{j,s}^{k} = y_{f[j,k],s}^{finish} \qquad \forall k \in K, j \in J_k, s \in S \tag{13}$$

$$z_{1,s}^{k,start} \geq \min\left(t_{1,s}^{k}, t_{2,s}^{k}\right) \qquad \forall k \in K, s \in S \tag{14}$$

$$z_{1,s}^{k,start} \geq t_{1,s}^{k} - M_{1,s}^{2,k} p_{1,s}^{k} \qquad \forall k \in K, s \in S \tag{14-1}$$

$$z_{1,s}^{k,start} \geq t_{2,s}^{k} - M_{1,s}^{2,k}\left(1 - p_{1,s}^{k}\right), \qquad \forall k \in K, s \in S \tag{14-2}$$

$$P_{2,s}^{k,max} \geq \max\left(t_{1,s}^{k}, t_{2,s}^{k}\right) \qquad \forall k \in K, s \in S \tag{15}$$

$$P_{2,s}^{k,max} \geq t_{1,s}^{k} \qquad \forall k \in K, s \in S \tag{15-1}$$

$$P_{2,s}^{k,max} \geq t_{2,s}^{k} \qquad \forall k \in K, s \in S \tag{15-2}$$

$$P_{j,s}^{k,max} \geq \max\left(P_{j-1,s}^{k,max}, t_{j,s}^{k}\right) \qquad \forall k \in K, j \in \{3..J_K\}, s \in S \tag{16}$$

$$P_{j,s}^{k,max} \geq P_{j-1,s}^{k,max} \qquad \forall k \in K, j \in \{3..J_K\}, s \in S \tag{16-1}$$

$$P_{j,s}^{k,max} \geq t_{j,s}^{k} \qquad \forall k \in K, j \in \{3..J_K\}, s \in S \tag{16-2}$$

$$z_{j_k,s}^{start} \geq \min\left(P_{j,s}^{k,max}, t_{j+1,s}^{k}\right) \qquad \forall k \in K, j \in \{2..J_k - 1\}, s \in S \tag{17}$$

$$z_{j,s}^{k,start} \geq P_{j,s}^{k,max} - M_{j,s}^{2,k} p_{j,s}^{k} \qquad \forall k \in K, j \in \{2..J_k - 1\}, s \in S \tag{17-1}$$

$$z_{j,s}^{k,start} \geq t_{j+1,s}^{k} - M_{j,s}^{2,k}\left(1 - p_{j,s}^{k}\right) \qquad \forall k \in K, j \in \{2..J_k - 1\}, s \in S \tag{17-2}$$

$$z_{J,s}^{k,start} \geq P_{J_k,s}^{k,max} \qquad \forall k \in K, s \in S \tag{18}$$

$$z_{j,s}^{k,finish} = z_{j,s}^{k,start} + \tau_{j,s}^{P_k} \qquad \forall k \in K, j \in J_k, s \in S \tag{19}$$

$$z_{j,s}^{k,start} \geq z_{j-1,s}^{k,finish} \qquad \forall k \in K, j \in J_k, s \in S \tag{20}$$

$$X, y^{start}, y^{finish}, z^{start}, z^{finish} \geq 0$$

APPENDIX B

PROOF OF THEOREM 1 AND 2

Proof of Theorem 1

M1 is based on the difference of finish time with nurse between patient *i* and patient *i*+1. We consider two cases for upper bounds of M1 for each patient *i* under each scenario *s*: Case1 is when finish time of patient *i*-1 with nurse is greater than and equal to maximum of the finish times of patient from 1 to *i*-1 with nurses; and Case2 is when maximum of the finish times up to patient *i*-1 with nurses is greater than the finish time of patient up to *i*-1 with nurses. For the constraints to be valid, we ensure that

$$M1_{i,s} \geq \left(N_{i-1,s}^{max} - y_{i-1,s}^{finish}\right)^{+} \qquad \forall i \in 4..I, s \in S$$

<u>Case 1</u>: $N_{i-1,s}^{max} \leq y_{i-1,s}^{finish}$

In this case, observe that

a. The appointment time of patient *i*-1 is at most 30 minutes after that of patient *i*-2, and thus $X_{i-1} \leq y_{i-2,s}^{start} + 30$.

b. By definition: $N_{i-1,s}^{max} \geq y_{i-2,s}^{finish} = y_{i-2,s}^{start} + \tau_{i-2,s}^{N}$, and thus $N_{i-1,s}^{max} - \tau_{i-2,s}^{N} \geq y_{i-2,s}^{start}$.

c. Combining the two, we get that patient *i*-1 is available at time:

$$X_{i-1} \leq y_{i-2,s}^{start} + 30 \leq N_{i-1,s}^{max} - \tau_{i-2,s}^{N} + 30.$$

d. A nurse will be available to serve patient *i*-1 at time $N_{i-1,s}^{max}$ or earlier.

e. The start time of patient *i*-1 with the nurse is

$$y_{i-1,s}^{start} \leq Max\{ N_{i-1,s}^{max}, N_{i-1,s}^{max} - \tau_{i-2,s}^{N} + 30\}.$$

Thus, the difference $y_{i-1,s}^{finish} - N_{i-1,s}^{max}$ is bounded by $\tau_{i-1,s}^{N} + Max\{0,30 - \tau_{i-2,s}^{N}\}$.

Case 2: $N_{i-1,s}^{max} > y_{i-1,s}^{finish}$

In this case, observe that while patient *i*-1 has finished with one nurse, say nurse1 w.l.o.g., the other nurse, nurse2, is still busy with an earlier patient. The difference between two can be calculated depending on which patient is still with nurse2. If patient *r* is still with nurse2, it means that patients *r*+1, *r*+2, …, through *i*-1 is seen by nurse1, we have that:

a. $N_{i-1,s}^{max} = y_{r,s}^{start} + \tau_{r,s}^{N}$

b. $y_{i-1,s}^{finish} \geq y_{r-1,s}^{start} + \tau_{r+1,s}^{N} + \tau_{r+2,s}^{N} + \cdots + \tau_{i+1,s}^{N}$

c. $y_{r,s}^{start} \geq y_{r-1,s}^{start}$ since patients are ordered according to their appointment times, $X_1 \leq X_2 \leq \cdots \leq X_I$.

d. Thus, the difference $N_{i-1,s-}^{max} - y_{i-1,s}^{finish} \leq \tau_{r,s}^{N} - \sum_{u=r+1}^{i-1} \tau_{u,s}^{N}$

$Max_{r=1,\ldots,i-2}\left\{\tau_{r,s}^{N} - \sum_{u=r+1}^{i-1} \tau_{u,s}^{N}\right\}$ will provide the tight bound.

The overall bound on the difference for both cases then is

$Max\{Case1, Case\ 2\} = Max\left\{\tau_{i-1,s}^{N} + Max\left\{0{,}30 - \tau_{i-2,s}^{N}\right\}, Max_{r=1,\ldots,i-2}\left\{\tau_{r,s}^{N} - \sum_{u=r+1}^{i-1} \tau_{u,s}^{N}\right\}\right\}$

Proof of Theorem 2

M2 is derived from the difference of nurse finish time between patient *j* and patient *j*+1 with provider *k,* for *j*=1, 2, …, *J_k*, for provider *k*. Observe that if $t_{j,s} = y_{i,s}^{finish}$ then $t_{j+1,s} = y_{i+2,s}^{finish}$ where *j* is the *j*th patient in provider's *k* panel, who is the *i*th patient in the practice. For the M2 constraints to be valid we must ensure that

$$M2_{j,s} \geq \left(P_{j,s}^{max} - y_{i+2,s}^{finish}\right)^{+}$$

We also consider two cases for upper bounds of M2: Case1 is when nurse finish time of patient *i*+2 is greater than and equal to maximum of the nurse finish times up to patient *j*; and

Case2 is when maximum of the nurse finish times up to patient *j* is greater than the finish time of patient *i*+2 with nurse.

Case 1: $P_{j,s}^{max} \leq y_{i+2,s}^{finish}$

In this case, observe that

a. The appointment time of patient *i*+2 is at most 30 minutes after that of patient *i*, thus

$$X_{i+2} \leq y_{i,s}^{start} + 30$$

b. By definition: $P_{j,s}^{max} \geq y_{i,s}^{finish} = y_{i,s}^{start} + \tau_{i,s}^{N}$, and thus $P_{j,s}^{max} - \tau_{i,s}^{N} \geq y_{i,s}^{start}$

c. Combining the two, we get that patient *i*+2 is available at time:

$$X_{i+2} \leq y_{i,s}^{start} + 30 \leq P_{j,s}^{max} - \tau N_{i,s} + 30$$

d. Patient *i*+1 (from the other provider's panel) will be seen by a nurse at a time no later than $Max\{\, P_{j,s}^{max}, P_{j,s}^{max} - \tau_{i,s}^{N} + 30\}$. This is using that consecutive patients arrive at most 30 minutes apart to the practice.

e. A nurse will be available for patient *i*+2 at time $Max\{P_{j,s}^{max}, P_{j,s}^{max} - \tau N_{i,s} + 30\} + \tau N_{i+1,s}$, or earlier. This is a bound on the time a nurse will be available if patient *i*+1 is scheduled to see a nurse right after patient *i* finishes.

f. The start time of patient *i*+2 with the nurse is

$$y_{i+2,s}^{start} \leq Max\big\{\, P_{j,s}^{max} - \tau_{i,s}^{N} + 30, Max\{\, P_{j,s}^{max}, P_{j,s}^{max} - \tau_{i,s}^{N} + 30\big\} + \tau_{i+1,s}^{N}\}$$

$$= \tau_{i+1,s}^{N} + P_{j,s}^{max} + \; Max\big\{\, 0, \tau_{i,s}^{N} + 30\big\}$$

Thus, the difference $y_{i+2,s}^{finish} - P_{j,s}^{max}$ is bounded by $\tau_{i+2,s}^{N} + \tau_{i+1,s}^{N} + Max\big\{\, 0, -\tau_{i,s}^{N} + 30\big\}$

Case 2: $P_{j,s}^{max} > y_{i+2,s}^{finish}$

In this case, observe that while patient *i*+2 has finished with one nurse, say nurse1 w.l.o.g., the other nurse, nurse2, is still busy with an earlier patient $r \leq i + 1$ from the same provider. The difference between the two can be calculated depending on which patient is still with nurse2. If patient *r* is still with nurse2, it means that patients *r*+1, *r*+2, …, through *i*+1 were seen by nurse1, we have that:

a. $P_{j,s}^{max} = y_{r,s}^{start} + \tau_{r,s}^{N}$

b. $y_{i+1,s}^{finish} \geq y_{r+1,s}^{start} + \tau_{r+1,s}^{N} + \tau_{r+2,s}^{N} + \cdots + \tau_{i+1,s}^{N}$

c. $y_{i+2,s}^{start} \geq y_{i+1,s}^{start}$ since patients are ordered according to their appointment times, $X_1 \leq X_2 \leq \cdots \leq X_J$.

d. Thus, the difference $P_{j,s}^{max} - y_{i+2,s}^{finish} \leq \tau_{r,s}^{N} - \sum_{u=r+1}^{i+1} \tau_{u,s}^{N}$

The maximum in $Max_{r=1,\ldots,j,r\ in\ provider's\ k\ panel}\left\{\tau_{r,s}^{N} - \sum_{u=r+1}^{i+1} \tau_{u,s}^{N}\right\}$ will give us the bound we are looking for in this case.

The overall bound on the difference for both cases then is

$Max\{Case1, Case\ 2\} =$

$$Max\left\{\tau_{i+2,s}^{N} + \tau_{i+1,s}^{N} + Max\{\,0, -\tau_{i,s}^{N} + 30\}, \underset{r=1,\ldots,j,r\ in\ provider's\ k\ panel}{Max}\left\{\tau_{r,s}^{N} - \sum_{u=r+1}^{i+1} \tau_{u,s}^{N}\right\}\right\}$$

BIBLIOGRAPHY

Anderson, R. T., Camacho, F. T., and Balkrishnan. R. (2007) Willing to wait?: The influence of patient wait time on satisfaction with primary care. *BMC Health Services Research*, 7(1), 31.

Bagust, A., Place, M., and Posnett, J. W. (1999) Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *BMJ: British Medical Journal*, 319, 7203: 155.

Bailey, N. (1952) A Study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society*, 14, 185–199.

Balasubramanian, H., Biehl, S., Dai, L., and Muriel, A. (2103) Dynamic scheduling of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments, online at *Health Care Management Science*, DOI: 10.1007/s10729-013-9242-2

Berg, B. P. (2012) *Optimal Planning and Scheduling in Outpatient Procedure Centers*. [Raleigh, North Carolina], North Carolina State University. http://www.lib.ncsu.edu/resolver/1840.16/7926.

Blumenthal, D., Causino, N., Chang, Y., Culpepper, L., Marder, W., Saglam, D., Stafford, R., and Starfield, B. (1999) The duration of ambulatory visits to physicians. *The Journal of family practice*, *48*(4), 264-271.

Bodenheimer,T., and Pham, H. H. (2010) Primary care: current problems and proposed solutions. *Health Affairs (ProjectHope)*, 29, 799–805.

Camacho, F., Anderson, R. T., Safrit, A., Jones, A. S., and Hoffmann, P. (2006) The relationship between patient’s perceived waiting time and office-based practice satisfaction. *NC Med J*, 67, 6, 409-413

Cayirli, T., and Veral, E. (2003) Outpatient scheduling in health care: A review of literature. *Production and Operations Management: an International Journal of the Production and Operations Management Society*, 12, 519-549.

Cayirli, T., Veral, E., and Rosen, H. (2006) Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9, 47–58.

Cayirli, T., Veral, E., and Rosen, H. (2008) Assessment of patient classification in appointment system design. *Production and Operations Management*, 17, 338–353.

Chien, C. F., Tseng, F. P., and Chen, C. H. (2008) An evolutionary approach to rehabilitation patient scheduling: A case study. *European Journal of Operational Research, 189*(3), 1234-1253.

Chakraborty, S., Muthuraman, K., and Lawley, M. (2012) Sequential clinical scheduling with patient no-show: The impact of pre-defined slot structures. *Socio-Economic Planning Sciences*, 47(3), 205–219.

Chakraborty, S., Muthuraman, K., and Lawley, M. (2010) Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions*, 42, 354–366.

Chew, S. F. (2011) Outpatient appointment scheduling with variable interappointment times. *Modelling & Simulation in Engineering*, 2011.

Denton, B., and Gupta, D. (2003) A Sequential Bounding Approach for Optimal Appointment Scheduling. *IIE Transactions*, 35, 1003–1016.

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

Erdogan, S. A., and Denton, B. (2013) Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing,25*(1), 116-132.

Fishman, G. S. (1996) *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, New York, NY.

Gilchrist, V., McCord, G., Schrop, S. L., King, B. D., McCormick, K. F., Oprandi, A. M., ... and Zaharna, M. (2005) Physician activities during time out of the examination room. *The Annals of Family Medicine*, *3*(6), 494-499.

Gottschalk, A., and Flocke, S. A. (2005) Time spent in face-to-face patient care and work outside the examination room. *The Annals of Family Medicine*, *3*(6), 488-493.

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

Gul, S., Fowler, J. W., Denton, B. T., and Huschka, T. (2011) Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management*, 20, 406–417.

Hahn-Goldberg, S., Carter, M. W., and Beck, J. C. (2012) Dynamic template scheduling to address uncertainty in complex scheduling problems: a case study on chemotherapy outpatient scheduling. In *Society for Health Systems Conference,. Las Vegas, NV*.

Halton, J. (1970) A retrospective and prospective survey of the Monte Carlo method. *SIAM Review*, 12, 1–63.

Hammersley, J.M. and Handscomb, D. C. (1964) *Monte Carlo Methods*. Methuen, London.

Harper, P. R., and Gamlin, H. M. (2003) Reduced outpatient waiting times with improved appointment scheduling: a simulation modeling approach. *Or Spectrum*, 25(2), 207-222.

Hassin, R., and Mendel, S. (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Science*, 54, 565–572.

Ho, C. and Lau, H. (1992) Minimizing Total Cost in Scheduling Outpatient Appointments, *Management Science*, 38, 1750–1764.

Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, MI.

Hsu, V. N., de Matta, R., and Lee, C. Y. (2003) Scheduling patients in an ambulatory surgical center. *Naval Research Logistics (NRL)*, *50*(3), 218-238.

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

Kis, T., and Pesch, E. (2005) A review of exact solution methods for the non-preemptive multiprocessor flowshop problem. *European Journal of Operational Research*, *164*(3), 592-608.

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

Klassen, K. J., and Yoogalingam, R. (2009) Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management,* 18, 447–458.

Kleywegt, A., Shapiro, A., and Homem de Mello, T. (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12, 479–502.

Lin, J., Muthuraman, K., and Lawley, M. (2011) Optimal and approximate algorithms for sequential clinical scheduling with no-shows. *IIE Transactions on Healthcare Systems Engineering*, 1, 20–36.

Mancilla, C. and Storer, R. (2012) A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions*, 44, 655–670.

Migongo, A. W., Charnigo, R., Love, M. M., Kryscio, R., Fleming, S. T., and Pearce, K. A. (2012) Factors relating to patient visit time with a physician. *Medical Decision Making, 32*(1), 93-104.

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

Oh, H.J., Muriel, A., and Balasubramanian H. (2014) A User-friendly Excel Simulation for Scheduling in Primary Care Practices. *Proceedings of the Winter Simulation Conference*.

Oh, H. J., Muriel, A., and Balasubramanian H., Atkinson, K., and Ptaszkiewicz, T. (2013) Guidelines for scheduling in primary care under different patient types and stochastic nurse and provider service times. *IIE Transactions on Healthcare Systems Engineering* 3, 4, 263-279.

Petterson, S. M., Liaw, W. R., Phillips, R. L., Rabin, D. L., Meyers, D. S., and Bazemore, A. W. (2012). Projecting US primary care physician workforce needs: 2010-2025. *The Annals of Family Medicine, 10*(6), 503-509.

Pham, D. N., and Klinkert, A. (2008) Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research,185*(3), 1011-1025.

Pérez, E., Ntaimo, L., Malavé, C. O., Bailey, C., and McCormack, P. (2013) Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine. *Health care management science*, *16*(4), 281-299.

Potisek, N. M., Malone, R. M., Shilliday, B. B., Ives, T. J., Chelminski, P. R., DeWalt, D. A., and Pignone, M. P. (2007) Use of patient flow analysis to improve patient visit efficiency by decreasing wait time in a primary care-based disease management programs for anticoagulation and chronic pain: a quality improvement study. *BMC health services research*, *7*(1), 8.

Qu, X., Peng, Y., Kong, N., and Shi, J. (2013) A two-phase approach to scheduling multi-category outpatient appointments–A case study of a women's clinic. *Health care management science*, *16*(3), 197-216.

Ribas, I., Leisten, R., and Framiñan, J. M. (2010) Review and classification of hybrid flow shop scheduling problems from a production system and a solutions procedure perspective. *Computers & Operations Research*, *37*(8), 1439-1454.

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

Robinson, L.W. and Chen, R. R. (2011) Estimating the implied value of the customer's waiting time. *Manufacturing Service Oper. Management.* 13:1, 53–57.

Rockafellar, R. T. and Wets, R. J.-B. (1991) Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16, 119–147.

Ruiz, R., and Vázquez-Rodríguez, J. A. (2010) The hybrid flow shop scheduling problem. *European Journal of Operational Research*, *205*(1), 1-18.

Rojas, S., Castaño, F. A., Velasco, N., and Amaya, C. A. (2011) A decision support tool to generate the monthly schedule for consulting rooms in a public hospital.

Saremi, A., Jula, P., Elmekkawy, T., and Wang, G. (2013) Appointment scheduling of outpatient surgical services in a multistage operating room department. *International Journal of Production Economics*, 141, 646–658.

Sawik, T. (2000) Mixed integer programming for scheduling flexible flow lines with limited intermediate buffers. *Mathematical and Computer Modelling*,*31*(13), 39-52.

Sawik, T. (2001) Mixed integer programming for scheduling surface mount technology lines. *International Journal of Production Research*, *39*(14), 3219-3235.

Sawik, T. (2005) Integer programming approach to production scheduling for make-to-order manufacturing. *Mathematical and Computer Modelling*, *41*(1), 99-118.

Sawik, T., Schaller, A., and Tirpak, T. M. (2002) Scheduling of printed wiring board assembly in surface mount technology lines. *Journal of Electronics Manufacturing*, *11*(01), 1-17.

Soriano, A. (1966) Comparison of two scheduling systems. *Operations Research*, 14, 388–397.

Stahl, J. E., Holt, J. K., and Gagliano, N. J. (2011) Understanding performance and behavior of tightly coupled outpatient systems using RFID: Initial experience. *Journal of medical systems*, *35*(3), 291-297.

Tai-Seale, M., McGuire, T. G., and Zhang, W. (2007) Time allocation in primary care office visits. *Health services research*, *42*(5), 1871-1894.

Tang, J., Yan, C., and Cao, P. (2014) Appointment scheduling algorithm considering routine and urgent patients. *Expert Systems with Applications*,*41*(10), 4529-4541.

Topaloglu, S. (2006). A multi-objective programming model for scheduling emergency medicine residents. *Computers & Industrial Engineering*, *51*(3), 375-388.

Turkcan, A., Zeng, B., Muthuraman, K., and Lawley, M. (2011) Sequential clinical scheduling with service criteria. *European Journal of Operational Research*, 214, 780–795.

Van Slyke, R. M. and Wets, R. J.-B. (1969) L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal of Applied Mathematics*, 17, 638–663.

Wang, H. (2005). Flexible flow shop scheduling: optimum, heuristics and artificial intelligence solutions. *Expert Systems*, *22*(2), 78-85.

Welch, J. D. (1964) Appointment systems in hospital outpatient departments. *Operational Research Quarterly*, 15, 224–232.

Wu, X. D., Khasawneh, M. T., Hao, J., and Gao, Z. T. (2013, January). Outpatient Scheduling in Highly Constrained Environments: A Literature Review. In *The 19th International Conference on Industrial Engineering and Engineering Management* (pp. 1203-1213). Springer Berlin Heidelberg.

Yarnall, K. S., Østbye, T., Krause, K. M., Pollak, K. I., Gradison, M., and Michener, J. L. (2009) Peer Reviewed: Family Physicians as Team Leaders:“Time” to Share the Care. *Preventing chronic disease*, *6*(2).

Yawn, B., Goodwin, M. A., Zyzanski, S. J., and Stange, K. C. (2003) Time use during acute and chronic illness visits to a family physician. *Family practice*, *20*(4), 474-477.