



University of  
Massachusetts  
Amherst

## Flexibility and Capacity Allocation under Uncertain Prescheduled (Non-urgent) Demand and Same- day (Urgent) Demand in Primary Care Practices

Item Type	Dissertation (Open Access)
Authors	Gao, Xiaoling
DOI	<a href="https://doi.org/10.7275/6177121.0">10.7275/6177121.0</a>
Download date	2026-05-17 21:05:04
Link to Item	<a href="https://hdl.handle.net/20.500.14394/19379">https://hdl.handle.net/20.500.14394/19379</a>

**FLEXIBILITY AND CAPACITY ALLOCATION UNDER  
UNCERTAIN PRESCHEDULED (NON-URGENT)  
DEMAND AND SAME-DAY (URGENT) DEMAND IN  
PRIMARY CARE PRACTICES**

A Dissertation Presented

by

XIAOLING GAO

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

DOCTOR OF PHILOSOPHY

February 2015

Mechanical and Industrial Engineering Department

© Copyright by Xiaoling Gao 2015

All Rights Reserved

**FLEXIBILITY AND CAPACITY ALLOCATION UNDER  
UNCERTAIN PRESCHEDULED (NON-URGENT)  
DEMAND AND SAME-DAY (URGENT) DEMAND IN  
PRIMARY CARE PRACTICES**

A Dissertation Presented

by

XIAOLING GAO

Approved as to style and content by:

---

Ana Muriel, Co-chair

---

Hari Balasubramanian, Co-chair

---

Senay Solak, Member

---

Donald Fisher, Department Chair  
Mechanical and Industrial Engineering Department

*I dedicate this dissertation to my Dad, Guangping Gao and Mom, Baomei Jia, for their constant support and unconditional love. I love you all dearly.*

## ACKNOWLEDGMENTS

Similar as almost every Ph.D., it has been a long journey for me to complete the dissertation. Although I was full of motivation and passion to start the Ph.D., it is still inevitable to frequently experience frustration and depression along this journey due to many reasons, not only the difficulty and uncontrollability of research but also pressure from personal life. I forgot counting how many times I lost confidence, got blocked with research and writing, or was overtired with working in last four years. However, it is true that ‘Life is tough, but I am tougher!’ We always grow better and learn experiences from failures. I am really proud of myself to overcome all these negative feelings and to have a stronger mind now. Fortunately, I was not alone along this challenging and long journey. Without the supporters, who I am about to mention, I could not arrive at the destination.

First and foremost I’d like to sincerely and deeply thank my advisors Professor Hari Balasubramanian and Professor Ana Muriel. Along this long journey, they were not only advisors in my research to provide insight and direction but also good friends in my life to encourage and support me. Both Professor Balasubramanian and Professor Muriel have spent much time on guiding and helping me. They always provide their students patience, genuine caring and concern. I am very fortunate to work with my advisors in last four years. This valuable experience and all the stimulating advices from them during my Ph.D. process will definitely play an important role in my future career. For all of these, I cannot thank them enough. I am always grateful for all my advisors did for me.

Then I also want to thank Professor Senay Solak, who is the remaining member of my dissertation committee. His valuable feedback with my research are greatly appreciated.

I want to give special thanks to all the staffs in our department, especially Dorothy Adams and Ellen Roberts. As an international student, I had much more difficulties due to culture, language, and communications in the first year of my Ph.D. life. Dorothy and Allen are always nice, warm-hearted, and patient to help me to get familiar to the new environment and to enjoy the life in Umass.

I also thank all my officemates, Asli, Joanne, Michael, Yan, Eddie and Longjie, who are all my good friends. In the last four year, we are like a big family to share the joy of successes and the sorrow of failures. Their personal cheerings are greatly appreciated.

Of course no acknowledgments would be complete without giving thanks to Professor Bixiang Wang, who is the advisor of my master thesis. I will never forget his encouragement and advice during my master period. Without his help, I would not have the opportunity to come to Umass to purse the Ph.D.

In addition, I'd like to thank my best friend, Yingying Duan. Our friendship began since we were in high school, which was 15 years ago. We shared so many stories and thoughts in our lives. Her messages and suggestions are always full of her care and support, and belief in me, which comfort me a lot when I met difficulties and felt frustrated. I am really indebted to her for being my friend.

Last but certainly not least, I must acknowledge my parents with deepest thanks. My parents always provides me unconditional love and care. They have given me love and have taught me how to give others love. Although I am the single child of my family and my parents always miss me, they still encouraged me to go aboard to have experiences in foreign countries. They are always proud of me regardless of my

successes or failures. Without their support, I would never make it this far. My most sincerely thanks and deepest love are for my parents.

## ABSTRACT

# FLEXIBILITY AND CAPACITY ALLOCATION UNDER UNCERTAIN PRESCHEDULED (NON-URGENT) DEMAND AND SAME-DAY (URGENT) DEMAND IN PRIMARY CARE PRACTICES

FEBRUARY 2015

XIAOLING GAO

B.Sc., UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

M.Sc., NEW MEXICO INSTITUTE OF MINING AND TECHNOLOGY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ana Muriel and Professor Hari Balasubramanian

In this dissertation, we are applying and extending well-established concepts of flexibility in manufacturing and service sectors to a healthcare setting: primary care. In the healthcare scenarios, appointments are booked over time and thus future resource capacity is sequentially being allocated under partial demand information. In manufacturing flexibility is typically presented as a technology choice that requires heavy investment for expensive flexible equipment, or highly cross-trained workers, but can then be used at little or no cost to satisfy demand. In primary care, however, the resources are inherently flexible, as primary care physicians are naturally able to see other panel's patients. There is therefore no long-term cost to the system for "installing" flexibility, but a cost for "using" this flexibility. This cost results from the

loss of patient-physician continuity which may induce patient dissatisfaction, require longer appointment durations as the physician needs to study the unfamiliar patient's history, and potentially lead to poorer medical outcomes.

Appointments in primary care are of two types: 1) prescheduled appointments, which are booked in advance of a given workday; and 2) same-day appointments, which are booked as calls come during the course of the workday. This creates two competing demand streams with different continuity needs. For same-day patients, the need for timely access often outweighs the need for continuity. Prescheduled appointments, on the other hand, include patients with chronic conditions who require regular monitoring and follow ups, and for whom continuity is essential.

Within this context, we address two interrelated problems: 1) the capacity allocation between prescheduled and same-day patients and how it is impacted by flexibility and the addition of extra resources; 2) the dynamic allocation of same-day patients to an existing schedule as they call over the day. The study of the former aggregate capacity allocation problem is based on a 3-stage framework. We assume different flexibility configuration to study the impact of flexibility in primary care practices. Our study of flexibility in primary care practices suggest that better management of the inherently flexibility inside primary care practices helps to balance prescheduled and same-day demand streams. We then study the latter dynamic allocation problem based on a simulation model, which captures several realistic issues like, patient' preferences, call-in frequency of same-day requests, and policies to reserve time blocks for prescheduled patients, etc. Our study provides guidelines for clinic to provide better quality of care for patients.

# TABLE OF CONTENTS

	Page
<b>ACKNOWLEDGMENTS</b> .....	v
<b>ABSTRACT</b> .....	viii
<b>LIST OF TABLES</b> .....	xiv
<b>LIST OF FIGURES</b> .....	xvi
 <b>CHAPTER</b>	
<b>1. INTRODUCTION, RESEARCH MOTIVATION AND LITERATURE REVIEW</b> .....	
<b>1</b>	
1.1 Introduction and research motivation .....	1
1.2 Literature review .....	8
1.2.1 Limited resource allocation among non-urgent and urgent demands in manufacturing and service context .....	8
1.2.2 Flexibility .....	10
1.2.3 Related operations research applications in the context of health care .....	12
1.3 Contributions .....	17
1.4 Dissertation overview .....	20
 <b>2. MATHEMATICAL MODELING AND FRAMEWORK</b> .....	
<b>22</b>	
2.1 Introduction .....	22
2.2 Modeling framework and assumptions .....	22
2.3 Model I: prescheduled patients are dedicated .....	28
2.4 Model II: prescheduled patients are flexibly shared .....	30
2.5 Model III: prescheduled patients are pooled .....	39
2.6 Computational effectiveness and scalability of formulations .....	43

<b>3. STRUCTURAL PROPERTIES AND ANALYTICAL RESULTS</b>	<b>46</b>
3.1 Introduction	46
3.2 Diminishing returns property	46
3.2.1 Diminishing returns property	46
3.2.2 Counterexamples: diminishing returns property does not always hold	48
3.2.3 Configurations that satisfy the diminishing returns property	50
3.2.4 Proofs of Theorem 1	51
3.3 Impact of flexibility on booking limits	58
3.4 Greedy search heuristic	66
3.4.1 Iterative search procedure	67
3.4.2 Performance of greedy search heuristic	68
3.4.2.1 Computational results	70
3.4.3 Analytical method to calculate expected performances	72
3.4.4 Example: apply the greedy heuristic to the capacity allocation problem	75
3.4.5 Computational efficiency of the greedy heuristic and the lower bound	77
<b>4. IMPACT OF FLEXIBILITY AND CAPACITY ALLOCATION IN PRIMARY CARE PRACTICES</b>	<b>79</b>
4.1 Introduction	79
4.2 Computational results	79
4.2.1 Impact of flexibility in primary care practices	79
4.2.2 Evaluation of booking policies	86
4.2.3 $N^p$ changes as a function of prescheduled flexibility or same-day flexibility	90
4.2.4 Negative correlation study	93
4.3 Summary and conclusions	94
<b>5. IMPACT OF FLEXIBILITY AND CAPACITY ALLOCATION IN PRIMARY CARE PRACTICES WITH ADDITIONAL PROVIDER</b>	<b>98</b>
5.1 Introduction	98

5.2	Computational results . . . . .	99
5.2.1	Expected performances: under symmetric case . . . . .	101
5.2.2	Expected performances: under asymmetric Case . . . . .	103
5.2.3	Impact of flexibility and additional provider on the optimal booking limit . . . . .	105
5.2.4	Expected performances: under asymmetric P/S values . . . . .	107
5.2.5	Expected performances: under asymmetric 4 physician practices . . . . .	108
5.2.6	Impact of extra capacity from additional provider on the probability of overtime . . . . .	110
5.2.7	Sensitivity analysis: flexibility vs. booking limit . . . . .	112
5.3	Summary and conclusions . . . . .	114
<b>6.</b>	<b>SIMULATION: APPOINTMENT SCHEDULING PROBLEM IN PRIMARY CARE PRACTICES UNDER DYNAMIC ARRIVALS . . . . .</b>	<b>117</b>
6.1	Introduction . . . . .	117
6.2	Simulation model and assumptions . . . . .	119
6.2.1	Modeling the appointment call-in process . . . . .	120
6.2.2	Modeling patient preferences . . . . .	122
6.2.2.1	Modeling time-of-day preferences . . . . .	122
6.2.2.2	Modeling physician preferences . . . . .	124
6.2.3	Modeling allocation process . . . . .	125
6.3	Computational results . . . . .	127
6.3.1	Impact of prescheduled patient time-of-day flexibility . . . . .	127
6.3.2	Impact of guiding prescheduled patient appointment times . . . . .	129
6.3.3	Evaluation of a threshold policy for the appointment scheduling problem under dynamic arrivals . . . . .	133
6.3.3.1	Impact of different ratios of $R^p$ and $R^s$ . . . . .	133
6.3.3.2	Validation of reserving optimal $N^P$ policy: single physician practices and multiple physician practices . . . . .	136
6.3.4	Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences . . . . .	141

6.4	Summary and conclusions	148
<b>7.</b>	<b>DISCUSSION</b>	<b>150</b>
7.1	Applications in other contexts	150
7.2	Implications for primary care practices	151
7.3	Future study	154
<b>APPENDICES</b>		
<b>A.</b>	<b>EXPECTED REVENUE BASED ON ANALYTICAL METHOD</b>	<b>156</b>
A.1	Prescheduled patients and same-day patients are both dedicated	156
A.2	Prescheduled patients are dedicated while same-day patients are fully flexibly shared	157
A.3	Prescheduled patients and same-day patients are both dedicated with their own physicians while one additional provider is added to serve all same-day patients	158
A.3.1	Analysis of dedicated with overflow system when $m=2$	158
A.3.2	Analysis of Dedicated with overflow system when $m=4$	160
<b>B.</b>	<b>HOW TO REDUCE VARIANCE OF OUTPUT FOR OUR SIMULATIONS?</b>	<b>166</b>
B.1	Simulations based on large-scaled random numbers	166
B.2	Simulations based on common random numbers	169
	<b>BIBLIOGRAPHY</b>	<b>173</b>

## LIST OF TABLES

Table	Page
2.1 Running time comparisons of formulations: single replication (unit:s).....	44
3.1 Optimal solutions under given amount of N : asymmetric 6/12 8/16 10/20, 120% workload .....	72
3.2 Comparisons of running time under different configurations and demand scenarios(time:/s) .....	78
4.1 Impact of flexibility in primary care practices.....	82
4.2 Prescheduled patients are fully flexibly shared vs. prescheduled patients are pooled while same-day patients are always fully flexibly shared .....	85
4.3 Performances of different booking policies: expected revenue and 85% percentile of patient overflow .....	89
4.4 Does diminishing return property holds when two types of demand are negatively correlated? .....	94
5.1 Numerical experiment setting .....	100
6.1 Comparisons of aggregate model and simulation model: single physician practice .....	136
6.2 Comparisons of reserving optimal $N^p$ policy and no threshold policy: single physician practice .....	138
6.3 Comparisons of reserving optimal $N^p$ policy and no threshold policy: multiple physicians practice under symmetric demands .....	140
6.4 Comparisons of reserving optimal $N^p$ policy and no threshold policy: multiple physicians practice under asymmetric demands .....	141

6.5	Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences: symmetric demands . . . . .	144
6.6	Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences: symmetric demands(continuited with Table 6.5) . . . . .	145
6.7	Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences: asymmetric demands . . . . .	146
6.8	Comparisons of two same-day scheduling policies: continuity considered vs. continuity not considered for same-day patients . . . . .	147

## LIST OF FIGURES

Figure	Page
1.1 Different flexibility configurations . . . . .	3
2.1 Performances of 2-chain and full flexibility configuration under an extreme prescheduled demand scenario . . . . .	32
3.1 Spanning tree of the search space of the capacity allocation problem : 2-physician practice . . . . .	69
3.2 Revenue of dedicated and fully-flexible systems as a function of the iteration number in the greedy heuristic for a 2-physician practice, with 120% workload and demand asymmetry. . . . .	76
4.1 Sensitivity analysis: expected revenue vs. booking limit in the entire system. . . . .	87
4.2 Sensitivity analysis: patient overflow vs. booking limit in the entire system. . . . .	88
4.3 $N^p$ changes as a function of same-day flexibility when prescheduled patients are dedicated: under asym 6/12 8/16 10/20 . . . . .	90
4.4 $N^p$ changes as a function of prescheduled flexibility when same-day patients are fully flexible: under sym 8/16 . . . . .	92
5.1 Expected revenue, prescheduled appointments satisfied, same-day patients satisfied and diversions for fully flexible and dedicated configurations as a function of $Y$ . . . . .	101
5.2 Expected revenue, prescheduled appointments satisfied, same-day patients satisfied and diversions for fully flexible and dedicated configurations as a function of $Y$ . . . . .	103
5.3 120% workload: comparison of optimal capacity allocation between dedicated with additional same-day provider and fully flexible with additional same-day provider system . . . . .	105

5.4	Expected revenue, prescheduled appointments satisfied, same-day patients satisfied and diversions for fully flexible and dedicated configurations as a function of $Y$ . . . . .	107
5.5	Comparison of impact of flexibility in 4 physicians' practice and 2 physicians' practice . . . . .	108
5.6	Impact of additional provider on the probability of overtime . . . . .	110
5.7	Performance of dedicated and fully-flexible systems as a function of the iteration number in our algorithm for a 2-physician practice, with $y = 0$ , 120% workload and demand asymmetry . . . . .	112
6.1	The simulation time-scales: example of two physicians practice . . . . .	121
6.2	Phone call frequency over a workday . . . . .	121
6.3	Patients' time-of-day preferences . . . . .	123
6.4	Patients' physician preferences . . . . .	124
6.5	The impact of number of time preferences allowed for prescheduled patients on revenue, prescheduled overflow, and same-day overflow: single physician under demand ratio 8/16, 100% workload . . . . .	128
6.6	The impact of number time preferences allowed for prescheduled patients on revenue, prescheduled overflow, and same-day overflow: single physician under demand ratio 16/8, 100% workload . . . . .	129
6.7	Expected revenue, prescheduled overflow and same-day overflow: single physician with demand ratio 8/16 under 100% workload; clinic guides the patients to schedule appointments with refusals . . . . .	131
6.8	Expected revenue, prescheduled overflow and same-day overflow: single physician with demand ratio 8/16 under 100% workload; clinic guides the patients to schedule appointments without refusals . . . . .	132
6.9	Expected revenue increment vs. $N^P$ under different ratios of $\frac{R^p}{R^s}$ : single physician under demand ratio 8/16, 100% and 120% workload . . . . .	134

6.10	Expected revenue increment vs. $N^P$ under different ratios of $\frac{R^p}{R^s}$ : single physician under demand ratio 16/8, 100% and 120% workload .....	135
B.1	Revenue vs. Number of replications : single physician 4/20, 100% workload .....	167
B.2	Revenue vs. Number of replications : single physician 8/16, 100% workload .....	167
B.3	Revenue vs. Number of replications : single physician 16/8, 100% workload .....	168
B.4	Revenue vs. Number of replications : single physician 20/4, 100% workload .....	168
B.5	Revenue vs. Number of replications : single physician 4/20, 100% workload under same seed .....	171
B.6	Revenue vs. Number of replications : single physician 8/16, 100% workload under same seed .....	171
B.7	Revenue vs. Number of replications : single physician 16/8, 100% workload under same seed .....	172
B.8	Revenue vs. Number of replications : single physician 20/4, 100% workload under same seed .....	172

# CHAPTER 1

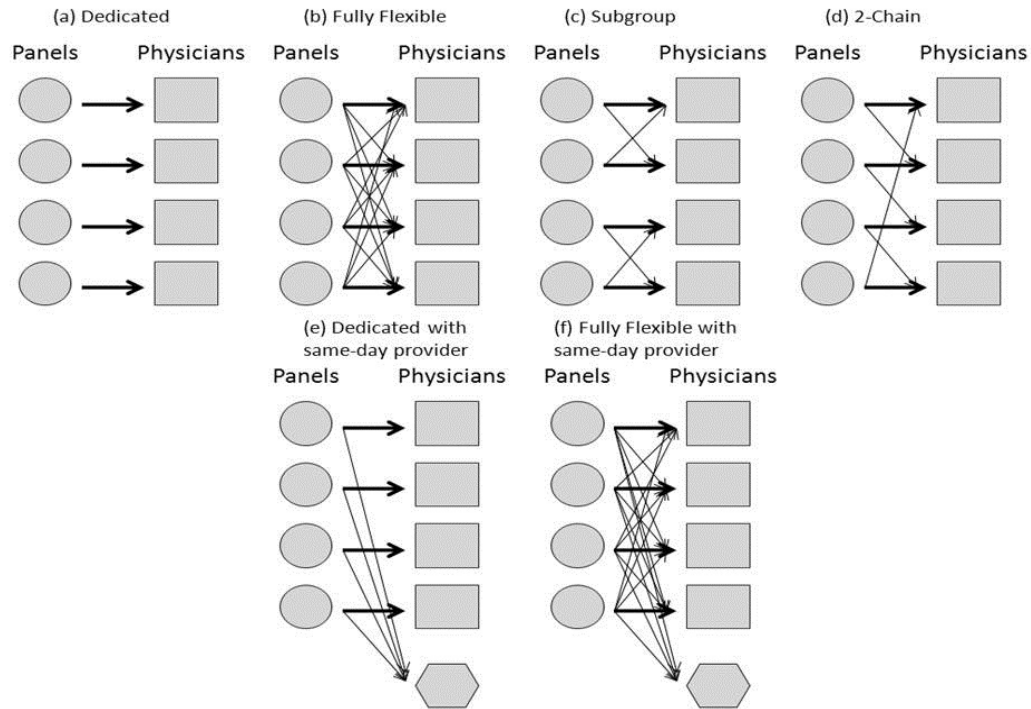
## INTRODUCTION, RESEARCH MOTIVATION AND LITERATURE REVIEW

### 1.1 Introduction and research motivation

Many manufacturing and service firms generally face important decisions on how to allocate limited resources/capacity among multiple demand classes. In this dissertation, motivated by a health care setting (outpatient primary care), we formulate a general model to study this problem. Throughout the dissertation, we focus on how to allocate primary care physicians' capacity to balance two successively realized demand streams: prescheduled (non-urgent) patients vs. same-day (urgent) patients; however, the model can be adapted to any other limited resource allocation problem with two demand classes, where one demand stream is realized and fulfilled in advance and the other demand stream arrives at short notice and needs to be urgently fulfilled. To avoid computational intractability, we first formulate higher-level aggregate model without considering the dynamic arrival of demand, to study capacity allocation involving two demand classes. Later we relax the aggregate assumption and evaluate the impact of dynamic arrivals. The most important decision variable in our capacity allocation problem is the booking limit for non-urgent demands, which arrive in advance. A booking limit seems necessary because otherwise there would not be sufficient capacity left for short-notice urgent demands, which arrive later.

Another key theme in this dissertation is the impact of flexibility on the capacity allocation process. The concept of flexibility (sharing or pooling of capacity) was initially proposed in manufacturing, as an effective way to balance demand for various

products with the capacity of different plants/resources. We are applying and extending well-established concepts of flexibility in manufacturing and service sectors to the primary care setting. In outpatient primary care, there are often multiple providers working in the same practice. These providers may choose to pool their capacity so as to effectively meet the variability in demand. In manufacturing flexibility is typically presented as a technology choice that requires heavy investment for expensive flexible equipment, or highly cross-trained workers, but can then be used at little or no cost to satisfy demand. In primary care, however, the resources are inherently flexible, as primary care physicians, due to their medical training, are naturally able to see their own patients as well as the patients of other physicians. There is therefore no long-term cost to the system for “installing” flexibility, but a cost for “using” this flexibility. Using flexibility results in the loss of patient-physician continuity. [22] shows that improving continuity could help patients be more satisfied with their care and could also help to improve the efficiency for both physicians and patients. [23] reports that reduced continuity could increase the likelihood of emergency department visits. Therefore flexibility must be used intelligently in primary care practices. We elaborate on this further in the discussion below.



**Figure 1.1.** Different flexibility configurations

A practice can achieve maximum continuity of care by mandating that patients should see only their own provider. This, however, hampers timely access to care. At the other extreme, a practice may allow patients to see any provider. This is ideal for timely access, but hampers continuity of care. The two extremes are shown in Figure 1.1(a) and Figure 1.1(b). In the first case, the providers are dedicated while in the second the providers are fully flexible. Figure 1.1(c) (d) (e) show partially flexible configurations that offer a middle ground between Figure 1.1(a) and Figure 1.1(b). In each of them, a patient sees only one other physician other than her own physician. Figure 1.1(d) is referred to as the 2-chain in the manufacturing flexibility literature [34] and allows demand variation to be absorbed effectively by the entire practice. While 2-chain is a new concept to health care, practices do use the subgroup configuration (Figure 1.1(c)). Physicians here may be divided into independent, self-contained teams (Mayo Clinic and other academic primary care

practices). The dedicated with additional provider configuration of Figure 1.1(e) is also common - here if the patient's own physician is unavailable, the patient tends to see an "overflow" physician or nurse practitioner (we have observed this setting at a small private practice as well as a community clinic in Western Massachusetts; academic medical centers also use this model). Figure 1.1(f) is actually another full flexibility configuration with additional provider, in which physicians could be seen by any physician or newly hired additional provider.

These flexibility configurations must be considered in light of the two different appointment types that a primary care practice schedules: 1) prescheduled appointments, which are booked in advance of a given workday; and 2) same-day appointments, which are booked as calls come during the course of the workday. This creates two demand streams competing for the practice's capacity, but with different continuity needs. Prescheduled appointments include patients with chronic conditions who require regular monitoring and follow ups, and for whom continuity (seeing their own provider) is more essential. For same-day patients, on the other hand, the need for quick access often outweighs the need to see one's own physician. With this in mind, how should a practice allocate the limited capacity of its multiple providers among prescheduled (non-urgent) patients and same-day (urgent) patients? What flexibility configurations are better suited to address demand effectively? These are the key questions we consider in this dissertation.

This dissertation also has policy relevance in the healthcare environment currently prevailing in the United States. The US faces a growing crisis of national wide shortage of primary care providers. For example, the rate of Massachusetts internists accepting new patients was 69% in 2006, and this rate was reduced to 52% in 2008 [4]. Other states like Texas and Vermont also have reported similar shortages in [17], [2], [45], [3] and [1]. To improve access, practices need to be more flexible and pool their capacities and yet provide continuity wherever necessary.

Our study in this dissertation can be broken down into two main parts:

I. In the models we propose, we allow for flexibility in satisfying both prescheduled demand and same-day demand; potentially a different flexibility configuration could be implemented for each of these demand streams. We investigate a number of questions regarding flexibility and capacity allocation. Are there analytical insights or properties that apply to particular flexibility configurations? What is the impact of same-day flexibility (the most realistic and useful setting in outpatient primary care) on the optimal booking limit? How do the optimal booking limits change with the additional flexibility in the system provided by the sharing of prescheduled slots while same-day patients are already fully flexibly shared? Will this result in significant improvements in timely access? Would these results hold true for other service applications where perhaps continuity is more important to same-day requests than to prescheduled ones? In a practice with high workloads, what is the impact of hiring a new provider who sees same-day patients? We use aggregate level stochastic optimization models for a single workday to answer these questions. We use these models to both derive analytical insights as well as to provide computational evidence.

II. In primary care practices, there are many realistic issues that impact appointment scheduling but are difficult to express mathematically in the framework of an optimization model. This includes the dynamic arrival of appointment requests (as opposed to the aggregate or lumped arrivals which we assume in part I); prescheduled patients' time of day preferences to schedule appointments; and prescheduled patients' willingness to be diverted to another physician. We use a discrete event simulation, applied to a single workday, to model these realistic features and answer questions such as the following. What is the impact of prescheduled patients' time-of-day flexibility? What is the impact of guiding prescheduled patient to certain reserved blocks of time so that same-day requests can be adequately fulfilled? And

what is the impact of flexibility under dynamic arrivals and patient preferences? Do the results we find the aggregate level stochastic optimization models still hold?

In part I and II, our principal performance measure is a weighted measure of the number of prescheduled and same-day patients seen by the practice during a workday. We call this measure “revenue”. We also report on a number of other relevant measures such as the number of unfulfilled prescheduled requests (prescheduled overflow) and number of unfulfilled same-day requests (same-day overflow, a proxy for overtime). Where necessary, we report not just the averages, but also higher percentiles of some of these measures.

In the entire dissertation, we assume that the prescheduled and same-day demands are independent of each other and are Poisson distributed. The Poisson assumption for same-day demand is certainly reasonable in health care. For prescheduled demand it requires some justification, since a part of the prescheduled demand is generated by follow-up visits that are scheduled by the clinic. However, when the demand for a particular workday in the future is considered, a Poisson rate appears to be a good approximation. Even if the clinic schedules follow-up visits, patient preferences for specific days imparts sufficient randomness to the demand process.

In this dissertation, computational experiments are based on a range of demand scenarios, which could represent most common situations in primary care practices. To describe demand scenarios, we use the term ‘workload’ and the term ‘P/S’ in the entire dissertation. ‘Workload’ is defined as the ratio of the expected total demand for the clinic and the total available capacity. For instance, in a practice with two physicians, suppose each physician has a daily demand rate of 8 for prescheduled appointment and 16 for same-day appointments. The total expected demand is  $2 \times 8 + 2 \times 16 = 48$ , and the total capacity is  $24 \times 2 = 48$ , therefore, workload of the clinic is 100%. In a 120% utilized clinic, the capacity would still be 48, but the expected demand would be  $(2 \times 8 \times 1.2) + (2 \times 16 \times 1.2) = 57.6$ . Here, each physician’s average

prescheduled demand and same-day demand is multiplied by 1.2 to create the 120% workload case. The term ‘P/S’ for one physician is used to refer to the ratio of prescheduled to same-day demand at the 100% workload level for this physician. For example, suppose a 2-physician practice is under 120% workload, and  $P/S$  ratio for physician 1 is  $6/12$ , then the expected prescheduled demand and the expected same-day demand for physician 1 are  $6 \times 1.2$  and  $12 \times 1.2$ , respectively. Workload reflects how busy the clinic is and  $P/S$  ratios reflect the type of clinic. For instance, family medicine clinics are likely to have a greater number of prescheduled appointments; at the other end of the spectrum, an urgent care center will have mostly same-day patients but very few prescheduled patients.

Computational experiments in the dissertation cover multiple symmetric scenarios and asymmetric scenarios. In the symmetric cases, each physician in the given practice is utilized with identical demand, i.e. the  $P/S$  ratio and the workload for individual physician is identical. In the asymmetric cases, we consider practices with asymmetry in the workload of individual physician as well as asymmetry in  $P/S$  ratios. For example, in a 3-physician practice under 100% workload, the  $P/S$  ratios of this practice are set to be  $6/12$   $8/16$   $10/20$  ( $6/12$  for physician 1,  $8/16$  for physician 2 and  $10/20$  for physician 3, we always apply this seriation for  $P/S$  ratios setting of a multiple physician practice throughout this dissertation). Then each physician is associated with same  $P/S$  ratio to be  $1/2$ , but each physician is under a different actual workload. On the other hand, for example, still in a 3-physician practice under 100% workload, we could assume the  $P/S$  ratios to be  $8/16$   $12/12$   $16/8$ , to construct a practice with asymmetry in  $P/S$  ratios for individual physicians but with identical workload for each physician. These asymmetries reflect situations where some senior physicians have greater number of patients than other physicians in the practice, or may have more patients with chronic conditions, with the result that their total prescheduled demand is higher in relation to their same-day demand.

The rest of this chapter is organized as follows: in Section 1.2, we present the literature review for each relevant concept, including resource allocation problems among multiple demand classes in manufacturing and service context, flexibility, and health-care applications. In Section 1.3, major contributions of our research are presented. Finally, in Section 1.4, we provide an overview of the dissertation.

## **1.2 Literature review**

In this dissertation, we limit our review to the most relevant papers in three topics. Firstly, we review quantitative results in limited resource allocation among non-urgent and urgent demands in manufacturing and service context. Next, we review the principal results in the area of flexibility most related to our research. Finally, we review the related operations research applications in the context of health care.

### **1.2.1 Limited resource allocation among non-urgent and urgent demands in manufacturing and service context**

Resource allocation problems among multiple customer classes - especially related to the allocation of non-urgent demands (scheduled in advance) and urgent demands (arrive at short notice) - can be commonly observed in many domains. We now provide a few examples of papers in this area.

[13] address the admission control and sequencing decision problem in a production system. In their study, one class of orders (long-term contract) is on a made-to-stock basis and another class of orders (arrive at short notice) is on a made-to-order basis. They use a simple two-class M/M/1 queue model to study the problem and characterize the structure of optimal policies under this model. [28] propose two models to address the admission control problem but with different characteristics. In the first model, all the orders are produced on a make-to-order basis while in the second model, the contractual orders (long-term contract) might be produced on a

make-to-stock basis. They propose a simple threshold policy in the make-to-order model and the optimal policy becomes complicated in the second model. Then they present a two-critical-number heuristic, which is shown to perform well. [40] propose a general model to study the limited capacity allocation problem over different demand classes when demand is stochastic and capacity is perishable, in a make-to-order manufacturing system.

Motivated by the challenging problem of scheduling crews in a gas utility company, [6] study a resource allocation problem to perform both non-urgent jobs (like non-urgent customer service, regular testing of instruments, construction work, and replacement work, all of which are usually scheduled in advance) and urgent jobs (like repairing a gas leak, which usually occurs randomly at short notice). The goal of the capacity allocation problem is to minimize overtime for the crew, while performing all non-urgent jobs before deadlines and solving all urgent jobs in required time. The resulting capacity allocation problem is very challenging. In their work, [6] allow the resources to be fully shared (i.e. full flexibility) to carry out the unpredictable emergent jobs, and consider the crew's overtime cost. The problem is decomposed into two phases in their model: job scheduling phase and crew assignment phase. They provide heuristics for each phase. Finally, they use simulation to validate the recommended strategies from the model and report a financial annual labor savings of 22.3%.

In healthcare, striking a balance between non-urgent and urgent demand occurs routinely. For example, limited operating room and surgeon capacity in hospitals needs to be allocated to balance elective surgeries demand while simultaneously accommodating emergency surgeries. [21] formulate a stochastic dynamic programming model to study the advance scheduling problem of elective surgery, given uncertain emergency surgery requests. [33] consider a similar problem to manage capacity over elective and emergency jobs, but in a multi-resource environment. A Markov Decision

Process formulation is used to study this problem and the demand information and resource availability are assumed to be immediately updated.

### 1.2.2 Flexibility

In our research, we focus on the inherent flexibility inside primary care practices to share physicians' capacity. In the manufacturing context, this is known as 'process flexibility'. The study of process flexibility started in 1980s. [50] survey several types of manufacturing flexibility and discuss the associated benefits, challenges, and trade-offs between no flexibility and full flexibility. Due to high installation cost of flexibility links, manufacturing firms usually prefer to maintain profit while introducing fewer flexibility links inside the system.

[34] firstly discuss the benefit of designing limited process flexibility inside system. They use a simulation method in numerical studies, and the results show that the long chain flexibility configuration can provide almost similar benefit as full flexibility configuration. Afterwards, quantities of applications arise with the concept of long chain and limited flexibility in other settings. For instance, [24] present a framework to address the flexibility benefit in multistage supply chains. Furthermore, they propose a flexibility measure to demonstrate the relationship between this measure and the inefficiencies of the supply chains. However, limited theoretical results explaining the effectiveness of long chain and limited flexibility have been reported. First, [5] show diminishing returns in benefit with the increased flexibility inside system. [16] develop a stochastic-programming-based method to quantify the performance of the long chain, using performance of full flexibility as baseline. The asymptotic analysis show that the long chain design performs almost as well as full flexibility when the system size is large. Furthermore, when the system size is finite, [52] established a theory to show the effectiveness of long chain design. While most literatures of long-chain design focus on the expected performances objectives, [15] study the problem

from the perspective of worse case analysis. The concept of graph expansion was introduced to study the problem. And they show that, under all demand scenarios, a variant of highly connected but sparse graphs is within  $\epsilon$ -optimality of the full flexibility graph.

Many papers study the application of process flexibility in different settings, for instance, queueing systems [51] and [30]. [41] study the impact of increasing flexibility under a make-to-order environment where flexibility is also used to hedge against operational variability. [11] use a newsvendor network model to study the classical capacity and flexible technology selection problem. In this study, each type of resource is associated with particular flexibility level (ability to process a given number of different product types). In the case of cross-training in serial production lines, [32] show that flexibility improves efficiency in two main ways: by balancing unevenness in capacity or workload between resources; and by handling the variability inherent in demand. Also in [32], they compare a strategy that balances capacity using the minimum amount of cross-training with the chaining of skills in the sequence of the serial line. They find that skill-chaining strategies are more robust, and more effective in variability buffering. [18] distinguish between range (the different demand scenarios that can be accommodated) and response (the cost of doing so; that is, the cost of using secondary rather than primary resources for production/service) of flexible systems. They show that upgrading system response outperforms improving system range. This result suggests that in the primary care settings, the benefits of restricting the number of doctors that can see a particular patient (resulting in lower cost of service because of familiarity and thus increased response) is likely to outweigh the higher range provided by a fully flexible team care practice where any doctor can see the patient.

### 1.2.3 Related operations research applications in the context of health care

The application of operations research to healthcare is a growing area of research. In fact, the limited resource allocation problem under multiple demand streams in healthcare has been widely observed, in settings as varied as primary care and surgical scheduling. In this section, we limit our review on appointment scheduling problem, which is an extension of the capacity allocation problem under dynamic demand arrivals.

The appointment scheduling problem in healthcare has attracted significant attention recently. [42] and [43] have recommended advanced access, which implies that physicians should “do today’s work today” rather than book appointments into the future. The adoption of open access, which promises patients same-day appointments, has prompted a series of questions. What is the impact of no-shows? How many patients can a physician have (panel size) to allow open access? What if patients have specific preferences to schedule appointments? How to deal with different type of patients? These questions have necessitated the use of queuing and stochastic optimization approaches that provide guidelines to practices.

[25] investigate the link between panel sizes and the probability of “overflow” or extra work for a physician under advanced access. They propose a simple probability model that estimates the number of extra appointments that a physician is expected to see per day as function of her panel size. The principal message of their work is that for advanced access to work, supply needs to be sufficiently higher than demand to offset the effect of variability.

[26] use a queuing model to determine the effect of no-shows on a physician’s panel size. They develop analytical queuing expressions that allow the estimation of physician backlog as a function of panel size and no-show rates. In their model, no show rates increase as the backlog increases; this results in the paradoxical situation

where physicians have low utilization even though backlogs are high - this is because patients that had to wait for long do not show up. [8] show that in addition to panel size, case-mix considerations are important when it comes to designing physician panels. Case-mix refers to the type of patients (older versus younger; healthy patients versus patients with chronic conditions) in a physician's panel. They propose that in the long term, panels can be redesigned to improve timely access and continuity. [38] represent a clinic's appointment system based on the framework of [26] and further assume a no-show risk for patients in the queue. This no-show risk is characterized with some probability decided by appointment delay. They show that, comparing to the impact of the magnitudes of patient show-up probabilities on the optimal decision of panel size, the impact of the sensitivity of these probabilities changing with delays has even more impact on the those optimal decisions.

Motivated with finding a proper appointment window to maximize the efficiency, [37] extend the framework in [38] by controlling the capacity of the queue (i.e., appointment window) to serve multiple types of patients who differ in arrival rates and show-up probabilities (i.e., heterogenous populations). The principle results of their work is, offering longer appointment window to patients with lower no-show rates is always optimal in a heterogenous population but may perform worse in a homogeneous population.

[27] conduct an empirical study of clinics in the Minneapolis metropolitan area that adopted open access. They provide statistics on call volumes, backlogs, number of visits with own physician (which measures continuity) and discuss options for increasing capacity at the level of the physician and clinic. [35] use discrete event simulation to study the effects of clinical characteristics in an open access scheduling environment on various performance measures such as continuity and overbooking. One of their primary conclusions is that continuity in care is affected adversely as the fraction of patients on open access increases. The authors mention provider group-

s (or physicians and support staff) working in teams as a solution to the problem. Numerous studies have studied the impact of no-shows and proposed overbooking strategies for single physician clinic sessions. Examples include [36], [44], and [14]. [49] compare the performance of two types of appointment-scheduling policies (traditional appointment scheduling and open access scheduling systems) for single provider. Though both of these two scheduling policies have substantial variability in the number of patients seen per day, the variability comes from different sources. In the traditional scheduling system, patients schedule their appointments well in advance, the variability in the number of patients seen per day results from no-shows within the fixed number of appointments for that day. In the open access system, the variability in the number of patients seen per day results from the day to day demand variability. Their numerical analysis reveals the open access generally outperforms the traditional appointment system when the objective function is a weighted average of patients' waiting time (lead time to appointment), the doctor's idle time, and the doctor's overtime. The traditional system works better than the open access system only when the patient wait time is held in little regard or when the probability of no-show is small. [39] propose new heuristic policies for dynamic scheduling of patient appointments under no-shows and cancelations. They find that open access works best when patient load is relatively low.

[53] consider an appointment booking problem for each workday separately to maximize clinic revenue. The patients' preferences are modeled explicitly by defining different acceptance probabilities for each physician and time-block combination, then are learned dynamically and updated to improve the booking decisions as booking preferences are different for each patient and they change over time for the same patient. A major criterion as patient-physician match rates measures patients' satisfaction in this paper. They also model two types of demand as advance-book (non-urgent) and same-day (urgent) demand which is similar in our model. For this multi-

physician practice, a sequential call-in process for prescheduled appointment request are considered while the same-day demand in their models are assumed to arrive just before the start of the workday. Their model is limited for low no-show rate practices.

The most closely related papers to our study are [29], [20], [48], [10], [19] and [7].

[29] explicitly model the capacity reservation problem under consideration of the patients' choice behavior to maximize the clinic revenue. The patients' choice behavior is characterized by a set of values to indicate the probability that a patient to request an appointment. The patients' preferences contain a specific slot in a day as well as a preference for their own physician's. They use a Markov Decision Process model to obtain booking policies that provide limits on when to accept or deny requests for appointments from patients. Both of one single physician practice and multiple physicians practice have been discussed to establish the optimal booking policies. The principal difference between their model and ours is that, in terms of flexibility, their clinic is fully flexible with regard to both non-urgent and urgent appointments while we consider different flexibility configurations and the associated impact in primary care practices.

[20] study the appointment scheduling problem with consideration of patient preferences regarding which they would like to be seen, based on an electronic appointment scheduling system. In their model, patients could choose a day to schedule appointment among the options provided by the system, or be refused if her preference does not match. No shows, revenues for shows, and service cost are considered in the model and the objective is to maximize the expected net profit for the proposed system. They initially propose an optimal policy with bounded optimality gap based on a static model. They then develop a dynamic model and propose a heuristic solution procedure. In our research, we also consider the appointment scheduling problem based on patients' preferences; however, the patients' preferences in our model is the preference for specific hours/times in a workday.

[48] consider an essential question for primary care practices: how many prescheduled appointments should a physician plan for in a workday given that the physician also has to see same-day patients? Their model considers different show rates for the two appointment types. They derive conditions under which a solution for the number of prescheduled appointments to reserve is locally optimal. [10] show a stronger result for the single physician problem, guaranteeing global optimality, by first showing that the revenue maximization function has diminishing returns under mild assumptions. They present a stochastic optimization model to determine the number of prescheduled appointments each physician should plan for and how this number changes depending on how flexible physicians are in seeing same-day patients of other physicians. An important conclusion of the study is that partial flexibility - where same-day patients are seen by a smaller subset of physicians, thus maintaining an acceptable level of continuity - comes very close to matching full flexibility with regard to the number of patients a practice is able to see per day. In our dissertation, we establish a similar framework as [10] and focus on the impact of flexibility in the resource allocation problem among multiple demand classes. We also extend the diminishing return property to hold under some particular flexibility configurations.

[19] examines the effect of reserving slots for urgent patients in health care practices on balancing long appointment queues with overtime to see urgent patients. Their model is based on a carve-out mixed with advance-access scheduling system. In their model, a fraction of the daily capacity is reserved for urgent patients, and the same-day appointments are scheduled on a first-come-first-serve basis. In addition, if the current day's capacity is full, the appointments of routine patients are carried over to the next day and all urgent overflow are satisfied on the same day by double booking, or referred to other physicians or emergency clinics. [7] focus on a hospital setting and formulate a general dynamic-programming model to study the problem of allocating fixed capacity among multiple customer classes. They assume that, for

some customer classes, the demand can be fully backlogged while for other customer classes, the demand will be lost if it is not fulfilled with available capacity. In this paper, they show that problem involving lost sales and backorders leads to a not so simple optimal structure yet exhibits desirable monotonicity properties. They propose a simple threshold heuristic policy in the computational study. In our research, we are interested in the impact of flexibility on resource allocation among multiple demand classes. To avoid computational intractability, we ignore the dynamically allocations in chapter 2 - 5, but use an aggregate model to capture the allocation process based on a simple threshold based heuristic policy, which performs well in [7] and is shown to be optimal in [19].

### 1.3 Contributions

In summary, our study of flexibility in primary care practices builds upon the extensive literature on manufacturing flexibility and its more recent application to service systems and worker training and allocation. There are, however, key operational differences that make the application of flexibility to primary care worthy of further analysis: (1) two demand streams associated with each resource, where one (prescheduled demands) gets realized before the other (same-day demands); (2) two conflicting objectives, timeliness and continuity of care; (3) no fixed cost associated with installing flexibility, but a loss in continuity for using it;(full flexibility configuration might be a good choice because it is easily implemented in practices with no installation cost) (4) appointments are booked over time and thus future resource capacity is sequentially being allocated under partial demand information. The latter point is mute in our aggregate analysis of the capacity allocation problem, but key to the dynamic clinic scheduling problem ([31]), which is considered in Chapter 6.

[10] propose a model to formulate the described capacity allocation problem in primary care practices, under two successively realized demand streams as a two stage

stochastic integer program. In their work, only same-day patients are allowed to be flexibly shared by different physicians while prescheduled patients are always dedicated. In our dissertation, we first extend the formulation in [10] to a more general case (Section 2.4), where both prescheduled patients and same-day patients could be flexibly shared. A practice can choose any flexibility configuration to serve patients. This extended formulation depends on the analysis of relationship between the random demand scenarios, the sequential realization of these demands (i.e. prescheduled demands are realized before same-day demands), and the actual allocations based on different flexibility configurations. Furthermore, we propose a new framework, in which prescheduled patients from different panels share a common booking limit while same-day patients are fully flexibly shared (Section 2.5). In this framework, prescheduled patients are always served by their own physicians as long as the actual demand does not exceed the corresponding physician’s total capacity. This new framework is particularly beneficial in primary care practices because it could help to maintain the continuity for prescheduled patients (for whom continuity is much more critical) while improving access of patients as a regular full flexibility configuration.

We analyze different flexibility configurations and find that the diminishing return property holds for the revenue function under two particular groups of flexibility configurations (Section 3.2). One is any flexibility configuration as long as same-day patients are fully flexibly shared; and the other one is any same-day flexibility configuration as long as prescheduled patients are seen by their own provider (the dedicated case). Furthermore, under the former configurations, we prove that the optimal booking limit is always non-increasing when prescheduled flexibility inside the system increases (Section 3.3). Under the latter configurations, the optimal booking limit is shown to be non-decreasing with increased same-day flexibility when system is low-utilized, but non-increasing with increased same-day flexibility when system is over-utilized (Section 3.3). In Section 3.4, we show that greedy heuristic leads to an

optimal solution under symmetric cases. Interestingly, greedy heuristic yields optimal solution in our capacity allocation problems under all tested demand scenarios under a subset of all possible configurations. With this computational evidence, we propose the greedy heuristic as an effective approach to reduce computational efforts (Section 3.4).

Next, we computationally study the impact of flexibility in primary care practices. We focus on the prescheduled dedicated and same-day fully flexible configurations, which are commonly observed in practices due to the needs of maintaining continuity for prescheduled patients and the needs of increasing access for same-day patients. Surprisingly, in this capacity allocation problem, we find that, prescheduled flexibility always produces only a marginal revenue gain as long as same-day patients are fully flexibly shared. We also show that although the expected revenue of the entire system is not sensitive to the prescheduled booking limit, beyond some point, operating a threshold policy for prescheduled appointments could help to reduce the risk of long overtime (Section 4.2.2).

We study the impact of a newly hired additional provider in primary care practices, where the additional provider is limited to serve same-day patients. We find that this configuration can increase access for prescheduled patients, for whom continuity is critical. This is yet more evidence to support the finding that prescheduled flexibility only produce marginal gain as long as same-day patients are fully flexibly served (Section 5.2.1-5.2.2).

Finally, we establish a simulation model under dynamic arrivals to study the capacity allocation problem (usually called the appointment scheduling problem in health care under dynamic environment). In this dynamic model (Section 6.2), we capture some realistic issues in primary care practices, such as patients' preferences for time of the day, patients' willingness to be diverted to another physician, dynamic and non-homogeneous same-day arrivals, etc. Till now, few papers have considered

these issues. We still allow flexibility in the physician’s capacity-sharing behavior in the simulation to study its impact.

We use the simulation to also quantify the impact of patient preference flexibility on a practice’s performance. We vary this flexibility by varying the number of time-of-day preferences the patient will allow. Surprisingly, we find that the practice performs quite well even if a patient has only limited time-of-day choices under different patient preference distributions (Section 6.3.1). [9] showed that the earlier in the day prescheduled appointments are scheduled, the better the practice’s ability of satisfy same-day appointments during the 8-hour workday. However, [9] did not consider the fact that prescheduled patients have time-of-day preferences. We therefore test cases where a practice blocks certain hours of the day for prescheduled appointments, based on the same-day call frequency from a small, private 3-provider practice in Amherst, Massachusetts. The results suggest a better policy for this clinic to always leave more slots for same-day patients in the afternoon (Section 6.3.2).

## 1.4 Dissertation overview

This dissertation consists of seven chapters. In Chapter 1, we discuss the motivation to study the impact of flexibility in limited capacity allocation problem under non-urgent demand (booked in advance) and urgent demand (arrive randomly) and then report the literatures in three related topics: 1) limited resource allocation among non-urgent and urgent demands in manufacturing and service context; (2) flexibility (3) related operations research applications in the context of health care. In Chapter 2, we present a general 3-stage framework to capture the capacity allocation problem and then propose the mathematical formulations under different flexibility configurations. Chapter 3 presents all the analytical results based on the analysis of different flexibility configurations. Chapter 4 focuses on a primary care practice to discuss the insights by adapting models introduced in Chapter 2. Chapter 5 studies the impact of

additional providers introduced to a primary care practice. In Chapter 6, we establish a single workday simulation model to study a real clinic, where patients' preferences, call-in frequency over a day, clinic's policy to guide prescheduled appointments, etc., are considered. Finally, in Chapter 7, we discuss applications of our study in other contexts and implications for primary care practices based on our findings. Then we present possible directions for future study.

## CHAPTER 2

### MATHEMATICAL MODELING AND FRAMEWORK

#### 2.1 Introduction

In this chapter, we introduce a 3-stage stochastic models to capture the physicians' capacity allocation problem in primary care practices. In Section 2.2, we state the framework and major assumptions to establish the model. The model in Section 2.3 is under the cases that prescheduled patients are seen by their own physicians and flexibility is only allowed for same-day patients. The models in Section 2.4 are under the cases that flexibility is allowed for both prescheduled patients and same-day patients. Finally, in Section 2.5 we model the cases that an overall system limit on prescheduled appointments is used, and same-day appointments are fully flexibly shared (thus pooling the capacity available to prescheduled patients).

#### 2.2 Modeling framework and assumptions

We consider the daily capacity allocation of a general primary care practice with  $m$  physicians, indexed by  $i = 1, 2, \dots, m$ , each physician  $i$  with  $C_i$  patient appointment slots available. Physicians have their own patient panels, indexed by  $j = 1, 2, \dots, m$ , to serve. Let  $\mathcal{A}$  be the set of all possible panel-physician links  $(j, i)$  such that patients in panel  $j$  can be served by physician  $i$ . The set  $\mathcal{A}$  represents the particular flexibility configuration under consideration; that is, the network of allowed patient redirections within the practice. In our general model, we assume physician flexibility can be used by both prescheduled and same-day patients. That is  $\mathcal{A} = \mathcal{A}^p \cup \mathcal{A}^s$ . We use  $\mathcal{A}^p$  to represent prescheduled flexibility configuration and  $\mathcal{A}^s$  to represent same-day

flexibility configuration. In some particular computational experiments in Chapter 4 and Chapter 5, we focus on the most relevant cases in practice, where physician flexibility can only be used for the time-sensitive same-day patients. This is because patient-physician continuity is highly beneficial to prescheduled appointments, in which major physicals or follow-ups of chronic conditions are performed.

Each patient panel  $j$  generates two demand streams:  $D_j^p$  for prescheduled appointments, for whom continuity is critical, and  $D_j^s$  for same-day appointments, for whom timely access is essential.  $D_j^p$  and  $D_j^s$  are general random variables. The demand for prescheduled and same-day appointments can be represented by a random vector  $D = (\mathbf{D}^p, \mathbf{D}^s) = (D_1^p, D_2^p, \dots, D_m^p, D_1^s, D_2^s, \dots, D_m^s)$ . Note that  $\mathbf{D}^p$  is realized before  $\mathbf{D}^s$ , since prescheduled appointments are scheduled far in advance of a workday, same-day appointments are typically requested over the course of a workday. At the aggregate planning level we assume that all prescheduled demand is realized at once and, subsequently, all same-day demand is realized at once, at the beginning of the day. We ignore the dynamic/sequential arrival over the course of the scheduling period for prescheduled appointments and throughout the day for same day patients.

Each prescheduled patient from panel  $j$  seen by physician  $i$ , for any  $(j, i) \in \mathcal{A}^p$ , brings the practice a revenue of  $R_{ji}^p$  and each same-day patient from panel  $j$  a higher revenue of  $R_{ji}^s$ , when being seen by physician  $i$  for  $(j, i) \in \mathcal{A}^s$ . We assume that the revenue associated with same-day patients is higher because these patients (1) have a lower no-show rate, and (2) if no appointment slot is available, will either go to an emergency room and be lost to the practice, or require overtime at the practice. The prescheduled patients, on the other hand, can be offered a later appointment date in most situations. Thus, the practice needs to reserve some capacity to satisfy the urgent requests, and curtail the number of slots offered to prescheduled appointments to some level  $N_i^p \leq C_i$ . Note that any leftover unused slots, after the prescheduled allocation is completed, can now be used by the same-day patients. In addition, some

practices may make use of the inherent flexibility that primary care physicians have to see patients from any panel to better accommodate the demand.

We use a 3-stage model to formulate the capacity allocation problem for any given flexible configuration.  $N_i^p$  ( $i = 1, 2, \dots, m$ ) are the first stage decision variables which denote the booking limits, that is, how many slots should be made available for prescheduled appointments of each physician. The second and third stage decision variables are denoted by  $x_{ji}^p$  and  $x_{ji}^s$ , the number of booked prescheduled appointments and the number of same-day patients from panel  $j$  assigned to physician  $i$ , respectively.

Let  $\mathbf{N}^p = [N_1^p, N_2^p, \dots, N_m^p]$ ,  $\mathbf{X}^p = [x_{11}^p, x_{12}^p, \dots, x_{mm}^p]$  and  $\mathbf{X}^s = [x_{11}^s, x_{12}^s, \dots, x_{mm}^s]$ .

The objective is to maximize the expected revenue of satisfying prescheduled and same-day appointments. The mathematical formulation is as follows:

### 3-stage General Model

#### First Stage

$$\max_{\mathbf{N}^p} \mathbb{E}[R(\mathbf{N}^p, \mathbf{D}^p, \mathbf{D}^s)] \quad (2.1)$$

$$R(\mathbf{N}^p, \mathbf{D}^p, \mathbf{D}^s) = R^p(\mathbf{N}^p, \mathbf{D}^p) + \mathbb{E}_{\mathbf{D}^p} [R^s(\mathbf{N}^p, \mathbf{X}^{p*}(\mathbf{N}^p, \mathbf{D}^p), \mathbf{D}^s)] \quad (2.2)$$

$$\text{s.t. } N_i^p \leq C_i, \quad \forall i = 1, 2, \dots, m \quad (2.3)$$

$$N_i^p \geq 0 \text{ and integer} \quad (2.4)$$

#### Second Stage

$$R^p(\mathbf{N}^p, \mathbf{D}^p) = \max_{\mathbf{X}^p} \sum_{i=1}^m \sum_{j:(j,i) \in \mathcal{A}^p} R_{ji}^p x_{ji}^p \quad (2.5)$$

$$\text{s.t. } \sum_{j:(j,i) \in \mathcal{A}^p} x_{ji}^p \leq N_i^p, \quad \forall i = 1, 2, \dots, m \quad (2.6)$$

$$\sum_{i:(i,j) \in \mathcal{A}^p} x_{ji}^p \leq D_j^p, \quad \forall j = 1, 2, \dots, m \quad (2.7)$$

$$x_{ji}^p \geq 0 \text{ and integer} \quad (2.8)$$

### Third Stage

$$R^s(\mathbf{N}^p, \mathbf{X}^{p*}(\mathbf{N}^p, \mathbf{D}^p), \mathbf{D}^s) = \max_{\mathbf{X}^s} \sum_{i=1}^m \sum_{j:(j,i) \in \mathcal{A}^p} R_{ji}^s x_{ji}^s \quad (2.9)$$

$$\text{s.t. } \sum_{j:(j,i) \in \mathcal{A}^s} x_{ji}^s \leq C_i - \sum_{j:(i,j) \in \mathcal{A}^p} x_{ji}^p, \quad \forall i = 1, 2, \dots, m \quad (2.10)$$

$$\sum_{i:(j,i) \in \mathcal{A}^s} x_{ji}^s \leq D_j^s, \quad \forall j = 1, 2, \dots, m \quad (2.11)$$

$$x_{ji}^s \geq 0 \text{ and integer} \quad (2.12)$$

2.1 - 2.4 describe the first stage of this capacity allocation problem: maximize the total revenue gained from the prescheduled appointments and from the same-day appointments.  $N_i^p$  ( $i = 1, 2, \dots, m$ ) are the only decision variables in this stage. And constraint 2.3 requires  $N_i^p$  do not exceed total capacity of physician  $i$ .

2.5 - 2.8 establish the second stage of this capacity allocation problem: given booking limit  $N_i^p$ , we maximize the total revenue gained from prescheduled appointments to find the allocation of prescheduled demand. The allocation is given by  $x_{ji}^p$  ( $i = 1, 2, \dots, m, j : (j, i) \in \mathcal{A}^p$ ). Constraint 2.6 and constraint 2.7 limit the number of prescheduled appointments to the allocated capacity  $N_i^p$  and the realized demand  $D_i^p$ , respectively. Note that the decision variable  $\mathbf{X}^p(\mathbf{N}^p, \mathbf{D}^p)$ , which describe the allocation of prescheduled patients, are determined only with consideration of reserved capacity  $N_i^p$  and prescheduled demand realization, without consideration of same-day realizations.

2.9 - 2.12 formulate the third stage of this capacity allocation problem: given booking limit  $N_i^p$  and allocations  $x_{ji}^p$  for prescheduled patients obtained from the second stage, we maximize the total revenue gained from same-day appointments to find the allocation of same-day demand, which are given by  $x_{ji}^s$  ( $i = 1, 2, \dots, m, j : (j, i) \in \mathcal{A}^s$ ). Constraint 2.10 ensures that the total same-day appointments for any physician  $i$  do not exceed remaining capacity and constraint 2.11 limits the total

number of same-day appointments scheduled from a panel to the realized demand for such appointments from that panel.

A computationally intensive (and intractable) approach is to solve the three stage problem using an exhaustive search. In this approach, for any given booking limits and prescheduled demand scenario, the allocation of prescheduled requests is made independently of the potential outcomes of the later realized same-day demand, as would occur in practice. The allocation of same-day demand to the remaining capacity is then optimized. To reduce computational efforts, we prefer to establish a simpler stochastic program instead of the 3-stage model presented earlier, while still capture all the natures of this capacity allocation problem.

Modeling the problem as a common stochastic program is deceptively difficult. The major challenge in formulating the three-stage capacity allocation problem is correctly capturing the second step, where prescheduled demand can be allocated up to the desired booking limit while ideally considering the expected revenues on the third stage (same-day demand allocation). The difficulty is in making sure if the booking limit  $N_i^p$  for physician  $i$  is reached under given prescheduled demand scenarios. In an inappropriate formulation, a physician may reject or divert her prescheduled requests to another physician without reaching her booking limit, so that the leftover prescheduled slots are used to fulfill higher revenue same-day requests in scenarios where same-day demand is high. However, this allocation definitely contradict with the natures that (1) prescheduled demand is realized and fulfilled before same-day demand is realized and (2) physicians always satisfy the prescheduled patients from their own panels as much as possible. To correctly capture these natural characteristics of the capacity allocation problem, we propose multiple formulations under different flexibility configurations.

First, under dedicated prescheduled scenarios, where prescheduled demand can only be satisfied by the patient's own physician, this challenge can be easily over-

come by defining binary variables for each scenario that indicate whether or not the booking limit has been reached as a function of the observed prescheduled demand (**Formulation I**).

Second, when prescheduled patients can be fully flexibly shared, we could define a single binary variable for each scenario to indicate whether or not the total booking limit has been reached. Then the allocation of prescheduled demand and same-day demand could be optimized by the model (**Formulation II**). However, when prescheduled patients are partially flexibly shared, such formulations become much more complicated to be established. Because the allocation of prescheduled demand depends on not only the actual demand but also the allowed flexibility links. We present an approach to establish the formulation for a case, under which prescheduled flexibility configuration is 2-chain (**Formulation III**). This approach could be extended to other prescheduled partially flexible configurations.

Finally, if same-day patients are fully shared by all the physicians in the practice, i.e. under a full flexibility configuration for same-day requests, then it makes sense to implement a single booking limit for prescheduled requests that applies to the entire practice (**Formulation IV**). Prescheduled requests from each panel will be assigned to their physician until the practice-wide booking limit is reached. Again here whether or not the limit is reached can be evaluated directly based on given prescheduled demand scenarios. Thus the stochastic program can be formulated in the usual way with the addition of a single binary variable under each scenario that indicates whether or not the total booking limit is reached.

In **Formulation I - Formulation IV**, to solve the capacity allocation problems, we use a sample average approximation method. As described before, the demand for prescheduled and same-day appointments can be represented by a random vector  $D = (\mathbf{D}^p, \mathbf{D}^s) = (D_1^p, D_2^p, \dots, D_m^p, D_1^s, D_2^s, \dots, D_m^s)$ . Let  $T_i$  be the size of physician  $i$ 's panel.  $\mathbf{D}^p$  follows a discrete distribution that assigns a probability  $q_t$  to each possible

realization of demand for prescheduled, indexed by  $t$ ,  $t = 1, 2, \dots, T$ , where  $T \equiv T_1 \times T_2 \times \dots \times T_m$ ; that is,  $P[\mathbf{D}^p = (d_{1t}^p, \dots, d_{mt}^p)] = q_t$ . Given a particular realization  $t$  of  $\mathbf{D}^p$ , let  $p_{tu}$  denote the conditional probability  $P[\mathbf{D}^s = (d_{1u}^s, \dots, d_{mu}^s) | \mathbf{D}^p = (d_{1t}^p, \dots, d_{mt}^p)]$ .

We introduce the following decision variables:

$N_i^p$ : number of slots allocated for prescheduled demand of physician  $i$ ,  $i = 1, \dots, m$ .

$x_{jit}^p$ : number of slots booked by prescheduled demand in panel  $j$  assigned to physician  $i$  under demand realization  $t$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, m$  and  $(j, i) \in \mathcal{A}^p$ .

$x_{jitu}^s$ : number of same-day patients in panel  $j$  assigned to physician  $i$  under demand realization  $t$  for prescheduled and demand realization  $u$  for same-day, for all  $i = 1, \dots, m$ ,  $j = 1, \dots, m$  and  $(j, i) \in \mathcal{A}^s$ .

### 2.3 Model I: prescheduled patients are dedicated

In primary care practices, patient-physician continuity is important both for efficiency of service and improved patient outcomes, especially for prescheduled patients with chronic conditions. Thus, restricting prescheduled patients to see their own physicians while allowing flexibility for urgent patients to see other physicians is an attractive configuration to balance timely visits while ensuring continuity to the patients for whom it is critical. Based on this requirement, our general 3-stage model can be written as a simpler stochastic programming model, which is similar to the one in [10].  $N_i^p$  are first stage decision variables, and  $x_{jit}^p$  and  $x_{jitu}^s$  are second stage decision variables. Note that  $x_{jit} = 0 \forall j \neq i$ .

We introduce binary variables  $\phi_{it}$  ( $\forall i = 1, 2, \dots, m$  and  $t = 1, \dots, T$ ) in order to indicate whether or not there is left-over capacity initially reserved for prescheduled demand to be used by same-day appointments.

$$\phi_{it} = \begin{cases} 1, & \text{if } D_{it}^p < N_i^p \\ 0, & \text{otherwise} \end{cases} \quad (2.13)$$

## Formulation I

$$\max \sum_{t=1}^{T_L} \sum_{i=1}^m q_t \times [R_{jj}^p x_{jzt}^p + \sum_{u=1}^{U_t} p_{tu} \sum_{(j,i) \in \mathcal{A}^s} R_{ji}^s x_{jitu}^s] \quad (2.14)$$

$$\text{subject to } N_i^p \leq C_i, \forall i = 1, 2, \dots, m \quad (2.15)$$

$$N_i^p \leq D_{it}^p + C_i \phi_{it}, \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (2.16)$$

$$N_i^p \geq D_{it}^p - C_i(1 - \phi_{it}), \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (2.17)$$

$$x_{iit}^p \leq N_i^p, \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (2.18)$$

$$x_{iit}^p \leq D_{it}^p, \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (2.19)$$

$$\sum_{j:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq C_i - D_{it}^p \phi_{it}, \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.20)$$

$$\sum_{j:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq C_i - N_i^p + C_i \phi_{it}, \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.21)$$

$$\sum_{i:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq D_{jtu}^s, \forall j = 1, 2, \dots, m, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.22)$$

$$\phi_{it} \text{ binary}, \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (2.23)$$

$$N_i^p, x_{jzt}^p, x_{jitu}^s \geq 0 \text{ and integer}, \forall i, j = 1, 2, \dots, m, \text{ and } (j, i) \in \mathcal{A}^s, \quad (2.24)$$

$$t = 1, 2, \dots, T, u = 1, 2, \dots, U_t$$

Constraints 2.16 - 2.17 ensure that  $\phi_{it} = 1$  if  $D_{it}^p \leq N_i^p$  and  $\phi_{it} = 0$  if  $D_{it}^p > N_i^p$ . Constraints 2.18 - 2.19 limit the number of prescheduled appointments to the allocated capacity and the realized demand, respectively. Constraints 2.20 and 2.21 ensure that the total same-day appointments for any physician  $i$  do not exceed remaining capacity, when  $\phi_{it} = 1$  and  $\phi_{it} = 0$  respectively. Constraint 2.22 limits the total number of same-day appointments scheduled from panel  $j$  not to exceed the demand from this panel.

## 2.4 Model II: prescheduled patients are flexibly shared

Some primary care practices do allow for some flexibility in the allocation of prescheduled patients to physicians beyond their physicians. To study the effect of the flexible policies on timely access, we could use a full-blown 3-stage model; however, as explained earlier, the model requires an exhaustive search and sequential computation of the solutions to the different stages resulting in prohibitive run times.

In Section 2.3, we reduce the original 3-stage model to a simpler formulation under the scenarios that prescheduled patients are dedicated. In fact, another special case is that prescheduled patients are fully flexibly shared. Under these scenarios, the general 3-stage model could also be simplified by introducing single binary variable  $\phi_t$  for each prescheduled demand scenario  $t$  ( $\forall t = 1, \dots, T$ ) in order to indicate whether or not the total booking limit is reached by prescheduled demand in the entire system.

$$\phi_t = \begin{cases} 1, & \text{if } \sum_{i=1}^m D_{it}^p < \sum_{i=1}^m N_i^p \\ 0, & \text{otherwise} \end{cases} \quad (2.25)$$

Similarly,  $N_i^p$  are first stage decision variables, and  $x_{jit}^p$  and  $x_{jitu}^s$  are second stage decision variables. Note that  $x_{jit} = 0$  ( $\forall j \neq i, i, j = 1, \dots, m$ ).

### Formulation II

$$\max \sum_{t=1}^{T_L} \sum_{i=1}^m q_t \times \left[ \sum_{j:(j,i) \in \mathcal{A}^p} R_{ji}^p x_{jit}^p + \sum_{u=1}^{U_t} p_{tu} \sum_{j:(j,i) \in \mathcal{A}^s} R_{ji}^s x_{jitu}^s \right] \quad (2.26)$$

$$\text{subject to } N_i^p \leq C_i, \quad \forall i = 1, 2, \dots, m \quad (2.27)$$

$$\sum_{j:(j,i) \in \mathcal{A}^p} x_{jit}^p \leq N_i^p, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.28)$$

$$\sum_{i:(j,i) \in \mathcal{A}^p} x_{jit}^p \leq D_{jt}^p, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.29)$$

$$\sum_{j=1}^m D_{jt}^p \leq \sum_{j=1}^m N_j^p + \sum_{j=1}^m C_j \times (1 - \phi_t), \quad \forall t = 1, 2, \dots, T \quad (2.30)$$

$$\sum_{j=1}^m D_{jt}^p \geq \sum_{j=1}^m N_i^p - \phi_t \times \sum_{j=1}^m C_i, \quad \forall t = 1, 2, \dots, T \quad (2.31)$$

$$\sum_{i=1}^m x_{jit}^p \geq D_{jt}^p \times \phi_t, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.32)$$

$$\sum_{j:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq C_i - N_i^p + C_i \phi_t, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.33)$$

$$\sum_{j:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq C_i - \sum_{j=1}^m x_{jit}^p + C_i(1 - \phi_t), \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.34)$$

$$\sum_{i:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq D_{jtu}^s, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.35)$$

$$\phi_t \text{ binary}, \quad \forall t = 1, 2, \dots, T \quad (2.36)$$

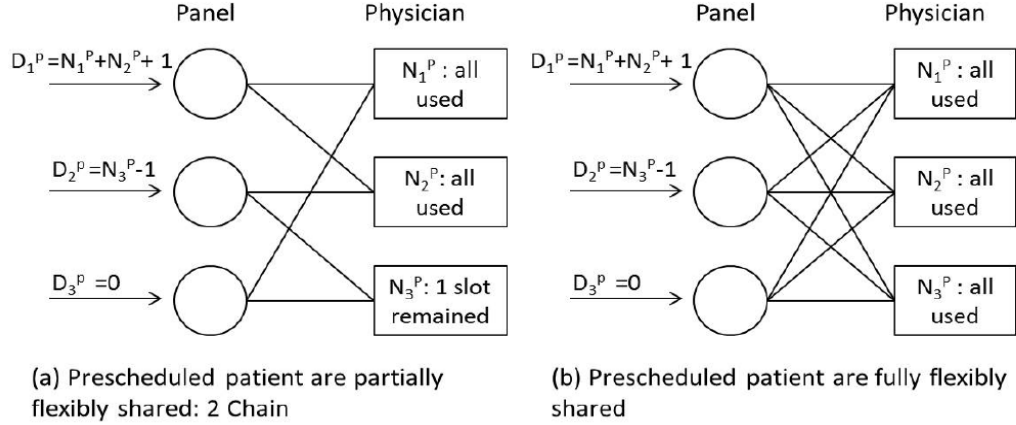
$$N_i^p, x_{jit}^p \geq 0 \text{ and integer}, \quad \forall i, j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.37)$$

$$x_{jitu}^s \geq 0 \text{ and integer}, \quad \forall i, j = 1, 2, \dots, m, \text{ and } (j, i) \in \mathcal{A}^s, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.38)$$

Constraints 2.28 - 2.29 limit the number of prescheduled appointments to the allocated capacity and the realized demand, respectively. Constraints 2.30 - 2.31 ensure that  $\phi_t = 1$  if  $\sum_{i=1}^m D_{it}^p \leq \sum_{i=1}^m N_i^p$  and  $\phi_t = 0$  if  $\sum_{i=1}^m D_{it}^p > \sum_{i=1}^m N_i^p$ . Constraint 2.32 is a key constraint only hold for prescheduled patients fully flexibly shared scenarios, which ensures that when  $\phi_t = 1$  (total prescheduled demand does not exceed total reserved capacity), all the prescheduled demand could be perfectly allocated. Constraints 2.33 and 2.34 ensure that the total same-day appointments for any physician  $i$  do not exceed remaining capacity, when  $\phi_{it} = 0$  and  $\phi_{it} = 1$  respectively. Constraint 2.35 limits the total number of same-day appointments scheduled from panel  $j$  not to exceed the demand from this panel.

By using **Formulation II**, computational effort could be reduced a lot under prescheduled patients fully flexibly shared scenarios. However, is it possible to establish similar formulation under prescheduled patients partially flexibly shared scenarios, by introducing appropriate decision variables? In fact, it is doable but very

distinctive for different flexibility configuration because not only actual demand but also flexibility configuration will impact on whether or not the booking limit of one physician is reached.



**Figure 2.1.** Performances of 2-chain and full flexibility configuration under an extreme prescheduled demand scenario

For example, Figure 2.1 shows the performances of a 2-chain configuration and a full flexibility configuration under one extreme prescheduled demand scenario. In this scenario, the demand from first panel is extremely high while the demand from third panel is zero. In fact, these two configurations have equal amount of prescheduled demand, which is  $N_1^p + N_2^p + N_3^p$ ; however, in the 2-chain configuration, one prescheduled request from panel one could not be served due to the limited flexibility in the system. In that case, the booking limit of physician 3 is not reached under the 2-chain configuration while it should be reached under the full flexibility configuration. Actually the 2-chain configuration performs almost same as the full flexibility configuration except under some extreme cases, for example,  $D_1^p > N_1^p + N_2^p$ ,  $D_1^p + D_2^p > N_1^p + N_2^p + N_3^p$ , etc. Note that, different partial flexibility configuration will have different optimal allocations to assign patients (certainly none of them could perform better than full flexibility). Then, for any prescheduled partially flexible configuration, we have

to analyze all the possibilities of relationships between the allocations and different demand realizations to establish the constraints.

In this dissertation, we illustrate how to formulate the capacity allocation problem based on a 3-physician practice, inside which 2-chain configuration is allowed for prescheduled patients and any flexibility configuration could be allowed for same-day patients. This approach can be extended to any other prescheduled partial flexibility configurations.

As described earlier in Section 2.2, the difficulty to establish the formulation is to create appropriate decision variables to indicate if booking limits are not reached (2.30 and 2.31 in **Formulation II**) and then to introduce constraints to guarantee correct amount of prescheduled patients are satisfied when booking limits is not reached (2.32 in **Formulation II**). In the prescheduled fully flexible configuration, single decision variable to indicate the situation in the entire system is sufficient because the full flexibility for prescheduled patients could perfectly redirect demand if there is available capacity from any other physician. However, when prescheduled flexibility is partial, we need such decision variable for each physician, because the links inside the system is not sufficient enough to adjust demand with available capacity under some demand scenarios, recall Figure 2.1.

Under the given prescheduled 2-chain configuration for a 3-physician practice, we use binary variables  $\phi_{it}$  for each physician  $i$  under prescheduled demand scenario  $t$  ( $\forall i = 1, \dots, 3, t = 1, \dots, T$ ) in order to indicate whether or not the individual booking limit is reached or not (0 reached, and 1 not reached). Different demand scenarios lead to different value of  $\phi_{it}$  and the conditions of demand scenarios are not as visualized as prescheduled full flexibility configuration. Based on our analysis, we propose 4 cases to jointly illustrate the relationships between demand scenarios and described decision variables. For each case, we introduce a group of decision variables  $\phi_{it}^k$  ( $\forall i = 1, \dots, 3, t = 1, \dots, T, k = 1, \dots, 4$ ) to indicate if any individual booking limit is reached

or not (always 0 reached, and 1 not reached in every case). We start analysis with the simplest case, then explore new case to take off the exceptions from known cases to revise the decision variables.

In addition, we introduce a group of important decision variables  $\phi_{it}^*$  for each physician  $i$  under prescheduled demand scenario  $t$  ( $\forall i = 1, \dots, 3, t = 1, \dots, T$ ):

$$\phi_{it}^* = \begin{cases} 1, & \text{if } D_{it}^p < N_i^p + N_k^p, k = (i + 1) \text{ mod } 3 \\ 0, & \text{otherwise} \end{cases} \quad (2.39)$$

to indicate if prescheduled demand for physician  $i$  exceed the potential maximum number of prescheduled patients could be seen (in this prescheduled 2-chain configuration, one panel is linked to two physicians, so the maximum number of satisfied prescheduled demand will never exceed the sum of booking limits of those two physicians).

The mentioned four different cases to decide if individual booking limit is reached or not are listed as follows:

(1) Although flexibility is allowed, prescheduled patients from one panel will be firstly served as much as possible by their own physicians to maintain continuity. That is, if the demand from panel  $i$  does not exceed the corresponding physician's booking limit  $N_i^p$ , all the prescheduled demand  $D_{it}^p$  will be fulfilled. We introduce binary variables  $\phi_{it}^1$  for each physician  $i$  ( $\forall i = 1, \dots, 3$ ) under prescheduled demand scenario  $t$  ( $\forall t = 1, \dots, T$ ):

$$\phi_{it}^1 = \begin{cases} 1, & \text{if } D_{it}^p < N_i^p \\ 0, & \text{otherwise} \end{cases} \quad (2.40)$$

to indicate if prescheduled demand from panel  $i$  exceed physician  $i$ 's booking limit.

(2) If prescheduled demand from panel  $i$  does not exceed physician  $i$ 's booking limit, is it still possible for this booking limit to be reached? The answer is yes because of the flexibility links. For example, if  $D_2^p < N_2^p$ , case 1 will decide  $\phi_{2t}^1$  to be

0; however, consider a prescheduled demand scenario  $t_1$   $[D_{1t_1}^p, D_{2t_1}^p, D_{3t_1}^p]$  that satisfied  $D_{1t_1}^p + D_{2t_1}^p > N_1^p + N_2^p$ , here then physician 2's booking limit is reached due to the flexibility link (1,2). To revise the decision variables, we introduce binary variables  $\phi_{kt}^2$  for physician  $k$  ( $k = (i+1) \bmod 3, i = 1, 2, 3$ ) under prescheduled demand scenario  $t$  ( $\forall t = 1, \dots, T$ ):

$$\phi_{kt}^2 = \begin{cases} 1, & \text{if } D_{it}^p + D_{kt}^p < N_i^p + N_k^p \\ 0, & \text{otherwise} \end{cases} \quad (2.41)$$

(3) If for physician  $i$ ,  $\phi_{it}^2$  is 1 (limit is not reached), is there any case to lead this physician's limit to be fulfilled? The answer is yes. Suppose  $D_2^p + D_3^p < N_2^p + N_3^p$  but  $D_1^p > N_1^p + N_2^p$ , then any demand scenario satisfying  $D_2^p + D_3^p \geq N_3^p$  leads physician 3's booking limit to be fulfilled. This is because, physician 1's reserved slots is much highly-utilized, then 2-chain flexibility tends to make the third physician's reserved capacity occupied by prescheduled demand from panel 2, in order to reduce prescheduled loss from panel 1. As long as  $D_2^p + D_3^p \geq N_3^p$ , physician 3's booking limit is always reached.

Recall the decision variable  $\phi_{it}^*$  to indicate if prescheduled demand for physician  $i$  exceed the potential maximum number of prescheduled patients could be seen. To revise decision variable based on above case, we introduce binary variables  $\phi_{k_1t}^3$  for physician  $k_1$  ( $k_1 = (i+2) \bmod 3, k_2 = (i+1) \bmod 3, i = 1, 2, 3$ ) under prescheduled demand scenario  $t$  ( $\forall t = 1, \dots, T$ ):

$$\phi_{k_1t}^3 = \begin{cases} 1, & \text{if } D_{k_1t}^p + D_{k_2t}^p < N_{k_1}^p + M_1 \phi_{it}^* \\ 0, & \text{otherwise} \end{cases} \quad (2.42)$$

$M_1$  is a sufficiently large number.

(4) Finally, if  $D_2^p + D_3^p < N_2^p + N_3^p$  but  $D_1^p \leq N_1^p + N_2^p$  (actually this case is a complement of case 3:  $D_2^p + D_3^p < N_2^p + N_3^p$  but  $D_1^p > N_1^p + N_2^p$ ). As long as  $D_1^p + D_2^p + D_3^p \geq N_1^p + N_2^p + N_3^p$ , physician 3's booking limit must be fulfilled, because the flexibility links could shuffle demand as perfectly as a full flexibility under this

case. Then, we introduce binary variables  $\phi_{k_1 t}^4$  for physician  $k_1$  ( $k_1 = (i + 2) \bmod 3$ ,  $k_2 = (i + 1) \bmod 3$ ,  $i = 1, 2, 3$ ) under prescheduled demand scenario  $t$  ( $\forall t = 1, \dots, T$ ):

$$\phi_{k_1 t}^4 = \begin{cases} 1, & \text{if } D_{it}^p + D_{k_2 t}^p + D_{k_1 t}^p < N_i^p + N_{k_2}^p + N_{k_1}^p + M_2(1 - \phi_{it}^*) \\ 0, & \text{otherwise} \end{cases} \quad (2.43)$$

$M_2$  is a sufficiently large number.

Till now, we could not explore any other cases to revise the proposed decision variables. The revising policy is, if under any one among above 4 cases, physician  $i$ 's limit is decided to be reached (decision variable is 0), then the final decision of physician  $i$  should be zero as well. Only when physician  $i$ 's limit is never reached under all the cases, the final decision of physician  $i$  is 1. That is,

$$\phi_{it} = \min_{k \in \{1, \dots, 4\}} \phi_{it}^k \quad (2.44)$$

In summary, for a given 3-physician practice, inside which prescheduled flexibility is 2-chain and sameday flexibility is any configuration, the mathematical formulation is as follow:

### Formulation III

$$\max \sum_{t=1}^{T_L} \sum_{i=1}^m q_t \times \left[ \sum_{j:(j,i) \in \mathcal{A}^p} R_{ji}^p x_{jit}^p + \sum_{u=1}^{U_t} p_{tu} \sum_{j:(j,i) \in \mathcal{A}^s} R_{ji}^s x_{jitu}^s \right] \quad (2.45)$$

$$\text{subject to } N_i^p \leq C_i, \quad \forall i = 1, 2, \dots, m \quad (2.46)$$

$$\sum_{j:(j,i) \in \mathcal{A}^p} x_{jit}^p \leq N_i^p, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.47)$$

$$\sum_{i:(j,i) \in \mathcal{A}^p} x_{jit}^p \leq D_{jt}^p, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.48)$$

$$D_{it}^p \leq N_i^p + N_k^p + \sum_{i=1}^m C_i \times (1 - \phi_{it}^*), \quad \forall i = 1, 2, \dots, m, \quad k = (i + 1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.49)$$

$$D_{it}^p \geq N_i^p + N_k^p - \sum_{i=1}^m C_i \times \phi_{it}^*, \quad \forall i = 1, 2, \dots, m, \quad k = (i+1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.50)$$

$$D_{it}^p \leq N_i^p + \sum_{i=1}^m C_i \times (1 - \phi_{it}^1), \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.51)$$

$$D_{it}^p \geq N_i^p - \sum_{i=1}^m C_i \times \phi_{it}^1, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.52)$$

$$D_{it}^p + D_{kt}^p \leq N_i^p + N_k^p + 2 \sum_{i=1}^m C_i \times (1 - \phi_{it}^2), \quad \forall i = 1, 2, \dots, m, \quad k = (i+1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.53)$$

$$D_{it}^p + D_{kt}^p \geq N_i^p + N_k^p - 2 \sum_{i=1}^m C_i \times \phi_{it}^2, \quad \forall i = 1, 2, \dots, m, \quad k = (i+1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.54)$$

$$D_{k_1t}^p + D_{k_2t}^p \leq N_{k_1}^p + 4 \sum_{i=1}^m C_i \times \phi_{it}^* + 2 \sum_{i=1}^m C_i (1 - \phi_{k_1t}^3)$$

$$\forall i = 1, 2, \dots, m, \quad k_1 = (i+2) \bmod 3, \quad k_2 = (i+1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.55)$$

$$D_{k_1t}^p + D_{k_2t}^p \geq N_{k_1}^p - 4 \sum_{i=1}^m C_i \times \phi_{it}^* - 2 \sum_{i=1}^m C_i \phi_{k_1t}^3$$

$$\forall i = 1, 2, \dots, m, \quad k_1 = (i+2) \bmod 3, \quad k_2 = (i+1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.56)$$

$$D_{it}^p + D_{k_2t}^p + D_{k_1t}^p \leq N_i^p + N_{k_2}^p + N_{k_1}^p + 4 \sum_{i=1}^m C_i \times (1 - \phi_{it}^*) + 2 \sum_{i=1}^m C_i (1 - \phi_{k_1t}^4)$$

$$\forall i = 1, 2, \dots, m, \quad k_1 = (i+2) \bmod 3, \quad k_2 = (i+1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.57)$$

$$D_{it}^p + D_{k_2t}^p + D_{k_1t}^p \geq N_i^p + N_{k_2}^p + N_{k_1}^p - 4 \sum_{i=1}^m C_i \times (1 - \phi_{it}^*) - 2 \sum_{i=1}^m C_i \phi_{k_1t}^4$$

$$\forall i = 1, 2, \dots, m, \quad k_1 = (i+2) \bmod 3, \quad k_2 = (i+1) \bmod 3, \quad t = 1, 2, \dots, T \quad (2.58)$$

$$\phi_{it} \leq \phi_{it}^k, \quad \forall i = 1, 2, \dots, m, \quad k = 1, 2, 3, 4, \quad t = 1, 2, \dots, T \quad (2.59)$$

$$\phi_{it} \geq \sum_{k=1}^4 \phi_{it}^k - 3, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.60)$$

$$\sum_{i:(j,i) \in \mathcal{A}^p} x_{jit}^p \geq D_{jt}^p \times \phi_{jt}^1, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.61)$$

$$\sum_{i:(j,i) \in \mathcal{A}^p} x_{jit}^p + \sum_{i:(k,i) \in \mathcal{A}^p} x_{kit}^p \geq (D_{jt}^p + D_{kt}^p) \times \phi_{jt}^2$$

$$k = (i+1) \bmod 3, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.62)$$

$$\sum_{i:(k_2,i) \in \mathcal{A}^p} x_{k_2it}^p + \sum_{i:(k_1,i) \in \mathcal{A}^p} x_{k_1it}^p \geq (D_{k_2t}^p + D_{k_1t}^p) \times \phi_{k_1t}^3 - 2 \sum_{i=1}^m C_i \phi_{jt}^* \\ k_2 = (j+2) \bmod 3, k_1 = (j+1) \bmod 3, \forall j = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (2.63)$$

$$\sum_{i:(j,i) \in \mathcal{A}^p} x_{jit}^p + \sum_{i:(k_2,i) \in \mathcal{A}^p} x_{k_2it}^p + \sum_{i:(k_1,i) \in \mathcal{A}^p} x_{k_1it}^p \geq (D_{jt}^p + D_{k_2t}^p + D_{k_1t}^p) \times \phi_{k_1t}^4 - 4 \sum_{i=1}^m C_i (1 - \phi_{jt}^*) \\ k_2 = (j+2) \bmod 3, k_1 = (j+1) \bmod 3, \forall j = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (2.64)$$

$$\sum_{j:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq C_i - N_i^p + C_i \phi_{it}, \quad \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.65)$$

$$\sum_{j:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq C_i - \sum_{j:(j,i) \in \mathcal{A}^p} x_{jit}^p + C_i (1 - \phi_{it}), \quad \forall i = 1, 2, \dots, m, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.66)$$

$$\sum_{i:(j,i) \in \mathcal{A}^s} x_{jitu}^s \leq D_{jtu}^s, \quad \forall j = 1, 2, \dots, m, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.67)$$

$$\phi_{it}, \phi_{it}^k, \phi_{it}^* \text{ binary}, \quad \forall t = 1, 2, \dots, T, k = 1, 2, 3, 4 \quad (2.68)$$

$$N_i^p, x_{jit}^p \geq 0 \text{ and integer}, \quad \forall i, j = 1, 2, \dots, m, \text{ and } (j, i) \in \mathcal{A}^p, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.69)$$

$$x_{jitu}^s \geq 0 \text{ and integer}, \quad \forall i, j = 1, 2, \dots, m, \text{ and } (j, i) \in \mathcal{A}^s, t = 1, 2, \dots, T, u = 1, 2, \dots, U_t \quad (2.70)$$

Constraints 2.47 - 2.48 limit the number of prescheduled appointments to the allocated capacity and the realized demand, respectively. Constraints 2.49 - 2.50 ensure that  $\phi_{it}^* = 0$  if the demand from panel  $i$  exceeds the sum of booking limit connected with this panel under given configuration; otherwise,  $\phi_{it}^* = 1$ . Constraints 2.51 - 2.58 correspond to the designed decision variables in those 4 cases explained earlier. Constraints 2.59 - 2.60 guarantee  $\phi_{it} = \min_{k \in \{1, \dots, 4\}} \phi_{it}^k$ . Constraints 2.61 - 2.64 require correct amount of prescheduled patients are all satisfied when booking limits is not reached, corresponding to those 4 cases, respectively. Constraints 2.65 and 2.66 ensure that the total same-day appointments for any physician  $i$  do not exceed remaining capacity, when  $\phi_{it} = 0$  and  $\phi_{it} = 1$  respectively. Constraint 2.67 limits the total number of same-day appointments scheduled from panel  $j$  not to exceed the demand from this panel.

## 2.5 Model III: prescheduled patients are pooled

In the general 3-stage model, each physician  $i$  allocates a number of slots  $N_i^p$  to see prescheduled patients. Alternatively, the total number  $N^p = \sum_{i=1}^m N_i^p$  of prescheduled slots could be shared by all physicians to see prescheduled patients from their own panels until the overall limit is reached. Under this policy, prescheduled patients are always assigned to their own physicians. This policy minimizes loss of continuity for these patients. Meanwhile, as the prescheduled capacity is pooled together, the coefficient of variation of the demand satisfied by the single capacity limit lowers, in much a similar way as when prescheduled flexibility is present.

In fact, when demand from one panel exceeds the total capacity of the associated physician, restriction of prescheduled patients to be dedicated might result in a small amount of loss for revenue. **Example 2.5.1** shows an extreme demand scenario to illustrate this rare case. In that case, we modify the policy to allow additional full flexibility to serve the excess prescheduled demand and avoid loss of revenue if the overall limit is not reached.

**Example 2.5.1** Suppose  $m = 3$ ,  $C_i = 24$ , the reserved capacity  $\mathbf{N}^p = [10, 10, 10]$  in the prescheduled fully flexible system, and  $N^p = 30$  for the prescheduled pooled system, the demand realization is  $\mathbf{D}^p = [25, 2, 2]$  and  $\mathbf{D}^s = [15, 14, 14]$ . The system is perfectly 100% utilized but different number of patients can be served by these two systems. In the prescheduled fully flexible system, all the patients can be seen while in the prescheduled pooled system, one prescheduled patient will be lost because  $D_1^p$  exceeds the physician's individual capacity.

In the following sections and chapters, for the prescheduled pooled model, we always allow the additional prescheduled full flexibility for the part of prescheduled demand that exceeds the total physician's capacity, which rarely happens. In other words, in our prescheduled pooled model, most of prescheduled patients are only served by their own physicians, except under a very extreme case (prescheduled re-

quests are larger than physician’s total capacity) in which they may be served by other physician in the clinic.

Comparing to a framework where prescheduled patients and same-day patients are both fully flexibly shared but with a separate booking limit for each physician, what is the benefit of the prescheduled pooled model? In fact, when same-day patients are fully flexible, the prescheduled fully flexible model and the prescheduled pooled model (additional prescheduled full flexibility allowed) with  $N^p = \sum(\mathbf{N}^p)$  will always see same number of prescheduled patients and same number of same-day patients, because all the capacity can be used in the most efficient way due to the pooling and full flexibility effects. The only difference between these two systems is in the number of diversions for prescheduled patients and the number of diversions for same-day patients. The pooled system usually will not divert any prescheduled patients (except under extreme cases that the pooled system may produce a small amount of prescheduled diversions, which is still much less than a prescheduled fully flexible configuration) but may result in further same-day diversions. Since the number of additional same-day redirections in the pooled system cannot exceed the number of additional prescheduled redirections in the fully flexible system, the pooled system will always give a better performance than the prescheduled fully flexible system, as long as the same-day diversion cost is smaller than the prescheduled diversion cost. Given a fully flexible same-day configuration, the following example illustrates the comparison of the performance of the prescheduled pooled model and the prescheduled fully flexible model.

**Example 2.5.2** Suppose we still have  $m = 3$ ,  $C_i = 24$ , the reserved capacity  $\mathbf{N}^p = [8, 8, 8]$  in the prescheduled fully flexible system, and  $N^p = 24$  for the prescheduled pooled system, the demand realization is  $\mathbf{D}^p = [20, 2, 2]$  and  $\mathbf{D}^s = [20, 14, 14]$ . The system is perfectly 100% utilized and all the patients can be served; however, different number of diversions occur in these two systems. In the prescheduled fully flexible

system, there are 12 prescheduled diversions and 4 same-day diversions, while in the prescheduled pooled system, there are no prescheduled diversions and 16 same-day diversions. As prescheduled diversion cost is higher than same-day diversion cost in primary care practices, the prescheduled pooled model would work better under this demand utilization.

Using a single booking limit, we introduce binary variables  $\phi_t$  ( $\forall t = 1, \dots, T$ ) in order to indicate whether or not the limit is reached in prescheduled demand scenario  $t$ .

$$\phi_t = \begin{cases} 1, & \text{if } \sum_{i=1}^m D_{it}^p < N^p \\ 0, & \text{otherwise} \end{cases} \quad (2.71)$$

#### Formulation IV

$$\max \sum_{t=1}^{T_L} \sum_{i=1}^m q_t \times \left[ \sum_{j=1}^m R_{ji}^p x_{jit}^p + \sum_{u=1}^{U_t} p_{tu} \sum_{(j,i) \in A^s} R_{ji}^s x_{jitu}^s \right] \quad (2.72)$$

$$\text{subject to } N^p \leq \sum_{i=1}^m C_i, \quad \forall i = 1, 2, \dots, m, \quad (2.73)$$

$$\sum_{j=1}^m D_{jt}^p \leq N^p + \sum_{i=1}^m C_i \times (1 - \phi_t), \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.74)$$

$$\sum_{j=1}^m D_{jt}^p \geq N^p - \sum_{i=1}^m C_i \times \phi_t, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.75)$$

$$\sum_{i=1}^m \sum_{j=1}^m x_{jit}^p \leq N^p, \quad \forall t = 1, 2, \dots, T \quad (2.76)$$

$$\sum_{i=1}^m x_{jit}^p \leq D_{jt}^p, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.77)$$

$$\sum_{j=1}^m x_{jit}^p \leq C_i, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T \quad (2.78)$$

$$\sum_{i=1}^m \sum_{j=1}^m x_{jitu}^s \leq \sum_{i=1}^m C_i - \phi_t \times \sum_{j=1}^m D_{jt}^p, \quad \forall t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.79)$$

$$\sum_{i=1}^m \sum_{j=1}^m x_{jitu}^s \leq \sum_{i=1}^m C_i - N^p + \phi_t \times \sum_{i=1}^m C_i, \quad \forall t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.80)$$

$$\sum_{i=1}^m x_{jitu}^s \leq D_{jt}^s, \quad \forall j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.81)$$

$$\sum_{j=1}^m x_{jit}^p + \sum_{j=1}^m x_{jitu}^s \leq C_i, \quad \forall i = 1, 2, \dots, m, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.82)$$

$$\phi_t \text{ binary}, \quad \forall t = 1, 2, \dots, T \quad (2.83)$$

$$N^p, x_{jit}^p, x_{jitu}^s \geq 0 \text{ and integer}, \quad \forall i, j = 1, 2, \dots, m, \quad t = 1, 2, \dots, T, \quad u = 1, 2, \dots, U_t \quad (2.84)$$

Constraints 2.74 - 2.75 ensure that  $\phi_t = 1$  if  $\sum_{i=1}^m D_{it}^p < N^p$  and  $\phi_t = 0$  if  $\sum_{i=1}^m D_{it}^p \geq N^p$ . Constraint 2.76 limits the overall number of prescheduled appointments to the allocated capacity. Constraints 2.77 - 2.78 require the individual prescheduled allocations to satisfy the corresponding demand limit and capacity limit, respectively. Constraints 2.79- 2.80 ensure that the total same-day appointments for any physician  $i$  do not exceed remaining capacity, and constraint 2.81 requires the total number of same-day appointments scheduled from panel  $j$  not to exceed the demand from this panel. Similarly, constraint 2.82 limits the sum of prescheduled and same-day allocations to one single physician not exceed the total individual capacity.

Note that, we only need to slightly modify **Formulation IV**, if prescheduled patients are restricted to be seen only by their own physicians. First, the binary variable  $\phi_t$  ( $\forall t = 1, \dots, T$ ) is changed to:

$$\phi_t = \begin{cases} 1, & \text{if } \sum_{i=1}^m \min[D_{it}^p, C_i] < N^p \\ 0, & \text{otherwise} \end{cases} \quad (2.85)$$

to ensure the excess prescheduled demand is lost. Meanwhile we need to add the constraint  $x_{jit}^p = 0$  ( $\forall j \neq i, i, j \in \{1, 2, \dots, m\}$ ) to restrict prescheduled patient all dedicated.

## 2.6 Computational effectiveness and scalability of formulations

To test the computational effectiveness and scalability of the four formulations presented in Section 2.3 - Section 2.5, we focus on a 3-physician practice and summarize the running time (in seconds) in Table 2.1.

We allow 5 different flexibility configurations for this practice. In this table, ‘P-D P-D’ and ‘P-D P-F’ denote ‘prescheduled is dedicated and same-day is dedicated’ and ‘prescheduled is dedicated and same-day is fully flexible’, respectively, which are both computed based on Formulation I. ‘P-F P-F’ denotes ‘prescheduled is fully flexible and same-day is fully flexible’, and results under this configuration are calculated based on Formulation II. ‘P-C P-F’ denotes ‘prescheduled is 2-chained and same-day is fully flexible’, which is calculated based on Formulation III. Finally, ‘P-P P-F’ denotes ‘prescheduled is pooled and same-day is fully flexible’, and results under this configuration are computed based on Formulation IV.

Note that, the size of input for each run is decided by number of prescheduled demand scenarios  $\times$  number of same-day demand scenarios, based on sample average approximation method. We increase this size of input from  $10 \times 10$  to  $50 \times 50$ . The tested cases are all symmetric cases (each physician has identical P/S ratio and workload, refer to section 1.1 for more details about definitions of P/S ratio and workload). We tested four different  $P/S$  ratios (4/20, 8/16, 16/8 and 20/4) and workload is always 100%. All the experiments are computed using cplex 12.3. Running environment is [Intel(R) Core(TM) i7-3770M CPU@ 3.40GHz, RAM 32.00GB]. Note

that all the results in Table 2.1 are based on single replication.  $\geq 99999$  means the running time is longer than the breakpoint we set, which is 99999 seconds.

**Table 2.1.** Running time comparisons of formulations: single replication (unit:s)

	size	sym 4/20	sym 8/16	sym 16/8	sym 20/4
P-D S-D	10×10	1.06	3.15	2.64	2.6
	20×20	3.18	3.12	3.14	3.17
	30×30	4.7	5.43	3.74	3.16
	40×40	9.07	9.74	10.17	6.22
	50×50	17.13	18.14	15.74	8.76
	size	sym 4/20	sym 8/16	sym 16/8	sym 20/4
P-D S-F	10×10	2.23	3.64	3.1	3.19
	20×20	9.59	7.94	12.57	3.73
	30×30	20.45	15.39	15.75	6.89
	40×40	38.97	44.69	71.51	22.97
	50×50	91.81	98.08	173.76	41.3
	size	sym 4/20	sym 8/16	sym 16/8	sym 20/4
P-F S-F	10×10	3.67	3.08	2.61	3.16
	20×20	3.76	3.79	3.69	3.66
	30×30	10.32	11.62	9.72	9.81
	40×40	17.17	22.46	22.65	20.81
	50×50	40.37	57.37	42.1	37.72
	size	sym 4/20	sym 8/16	sym 16/8	sym 20/4
P-C S-F	10×10	6.25	6.86	3.75	3.65
	20×20	234.59	262.6	17.42	4.24
	30×30	676.98	15817.88	66.41	4.26
	40×40	66293.98	$\geq 99999$	129.05	6.44
	50×50	$\geq 99999$	$\geq 99999$	2785.03	15.21
	size	sym 4/20	sym 8/16	sym 16/8	sym 20/4
P-P S-F	10×10	2.53	2.51	2.6	3.05
	20×20	2.61	2.65	2.68	3.07
	30×30	4.21	3.76	3.16	3.76
	40×40	3.87	4.43	3.85	3.34
	50×50	4.36	4.76	4.84	4.41

Generally, Formulation I, Formulation II and Formulation IV are computationally quite efficient. When the number of scenarios is increased from 100 to 2500, the

running time for a replication is not significantly increased. The only exception is prescheduled dedicated and same-day fully flexible configuration (also computed based on Formulation I). This configuration needs time to be solved than a fully dedicated configuration and the running time is also significantly impacted by the size of input.

Due to the large number of constraints, the computational intractability of Formulation III is not a surprise. We find that, under most scenarios, the running time of Formulation III is highly impacted by the increase in the size of the input. This impact varies depending on the different demand scenarios. For example, under the symmetric case 8/16, the impact of increasing size of input on the running time is the highest (an input size of  $40 \times 40$  results in running time to be greater than 99999 seconds while other cases can still be solved within this time limit). In addition, we observe that, under the symmetric case 20/4, the impact of increasing size of input on the running time is almost negligible.

# CHAPTER 3

## STRUCTURAL PROPERTIES AND ANALYTICAL RESULTS

### 3.1 Introduction

In this chapter, we study structural properties of the practice's capacity allocation problem and derive analytical results comparing the optimal booking limits under various flexibility configurations. In Section 3.2, we define the diminishing returns property, and show under what flexibility configurations this property holds. In Section 3.3, we determine the impact of flexibility on the optimal booking limit under some particular flexibility configurations, for which the diminishing returns property holds. The diminishing returns property suggests that a greedy procedure (adding one unit of prescheduled capacity at a time to the doctor with the greatest expected revenue increase) should work well. We introduce such a greedy heuristic in Section 3.4, and show it to work well for small practices: 1) it provides the exact optimal solution in all the cases tested under a wide range of parameters, and 2) it is computationally effective. For larger practices we will approach the problem numerically in later chapters and use sample average approximation.

For simplicity, we ignore diversion costs throughout this section.

### 3.2 Diminishing returns property

#### 3.2.1 Diminishing returns property

Based on the framework of the 3-stage model proposed in Chapter 2, the expected revenue of a given flexibility configuration for a practice with  $m$  physician-

s is  $\mathbb{E}[R(\mathbf{N}^p, \mathbf{D}^p, \mathbf{D}^s)]$ , where  $\mathbf{N}^p = [N_1^p, N_2^p, \dots, N_m^p]$ ,  $\mathbf{D}^p = [D_1^p, D_2^p, \dots, D_m^p]$ , and  $\mathbf{D}^s = [D_1^s, D_2^s, \dots, D_m^s]$ . Naturally, we are interested in identifying properties of the revenue function that we can exploit to derive analytical results and develop effective solution approaches.

In this dissertation, we find that the Property of Diminishing Returns, defined below, holds for the configurations most common in practice.

**Definition 1** (Diminishing Returns Property). *Let  $\Delta_k(N_1^p, \dots, N_m^p) = ER(N_1^p, \dots, N_k^p + 1, \dots, N_m^p) - ER(N_1^p, \dots, N_k^p, \dots, N_m^p)$  denote the difference in revenue associated with increasing the number of slots offered to prescheduled patients of physician  $k$  by 1. The capacity allocation problem has the Property of Diminishing Returns if (i) when  $\Delta_k(N_1^p, \dots, N_m^p) \geq 0$ , then  $\Delta_k(N_1^p, \dots, N_m^p)$  is non-increasing in  $N_1^p, \dots, N_m^p$  and (ii) when  $\Delta_k(N_1^p, \dots, N_m^p) < 0$ , then  $\Delta_k(\tilde{N}_1^p, \dots, \tilde{N}_m^p) \leq 0$  for any vector  $(\tilde{N}_1^p, \dots, \tilde{N}_m^p)$  such that  $\tilde{N}_i^p \geq N_i^p$  for all  $i = 1, 2, \dots, m$ .*

Observe that we only require the difference in revenue to not increase while it is still positive. If it is negative, it may increase, but will never become positive. As a result, an optimal search can stop once the returns are negative, since further increasing the booking limits will never result in a better solution.

In other words, the diminishing returns property requires that each time a component of  $\mathbf{N}^p$  is increased by one unit, the associated revenue change function is non-increasing in any component of  $\mathbf{N}^p$ , as long as the previous revenue change was still positive. This property suggests that a greedy algorithm that increases  $\mathbf{N}^p$  once component at a time following the path of myopic maximum revenue increase at each step will work well. Before we develop a greedy heuristic, we need to understand when the diminishing returns property holds, and further analyze its ramifications.

Will the diminishing returns property always hold for a general flexibility configuration? The counterexamples in the next section show that the answer is no in general; in Section 3.2.3, however, we show that the diminishing returns property holds for

any flexibility configuration of same-day patients given a dedicated prescheduled appointment strategy, and also for any prescheduled flexibility configuration given a fully flexible same-day appointment strategy.

### 3.2.2 Counterexamples: diminishing returns property does not always hold

To illustrate when the diminishing returns property does not hold we describe some simple examples based on demand points, rather than distributions, where the revenue difference increases at some points when the vector  $\mathbf{N}^P$  increases.

#### Example 3.2.1

Consider a 2-physician practice where each physician has 8 slots of available capacity. The physicians face deterministic demand: demand for prescheduled and same-day appointments is  $(8,0)$  for physician 1, and  $(0,8)$  for physician 2, respectively. Prescheduled demand is fully flexible, but same-day demand is dedicated. If  $N_1^P = 4$  and  $N_2^P = 4$ , adding one more prescheduled slot to physician 1, i.e. making  $N_1^P = 5$ , will actually allow us to release one more slot for same day patients of physician 2 (since still  $N_2^P = 4$  but only 3 slots will be used). Then,  $R(5, 4) - R(4, 4) = R^s > R(4, 4) - R(3, 4) = R^p$ .

In general, this example shows that if flexibility for prescheduled appointments is allowed and same-day patients are not equally flexible, additional capacity for prescheduled appointments may result in a better allocation of the realized demand for prescheduled appointments to the various physicians, releasing precious capacity to same-day appointments who are not flexible to see the other physicians. As a result, an increase in the capacity allocated to prescheduled appointments may allow the system to serve an additional same-day patient with the consequent increase in revenue. Therefore, same-day patients need to also have access to flexibility in order for diminishing returns to hold. The issue is, how much flexibility would be sufficient?

The following example shows that diminishing returns would not always hold unless same-day patients enjoy full flexibility (given some prescheduled flexibility).

**Example 3.2.2**

Consider a 3-physician practice where 2-chain flexibility is used for both patient types, so that patients of panel  $i$  can see physicians  $i$  and  $(i+1)$  ( $i = 1, 2$ ) and patients of panel 3 can see physicians 3 and 1. Each physician has capacity 8 and the initial  $N^p$  is 4 slots for each of them. Demands are  $(3,12)$ ,  $(4,0)$  and  $(5,0)$ . In the current setting, all prescheduled demand is served (with physicians 1, 2, and 3, seeing 4 prescheduled patients each and one patient from 1 diverted to 2, and one patient from 2 diverted to 3), but only 8 of the 12 same-day patients can be seen, as they cannot reach the capacity of physician 3. Let's now increase  $N_3^p$  to 5. In that case, we can still satisfy the 12 units of prescheduled demand, with physician 1, 2, and 3 seeing 3, 4, and 5 patients respectively. This frees one additional slot in physician 1 that can be used to serve one more of her same-day patients, so that now 9 can be seen. We have:  $R(4, 4, 4) - R(4, 4, 3) = R^p < R(4, 4, 5) - R(4, 4, 4) = R^s$ .

In the next example, we show that even for the case of a single physician after the revenue change becomes negative, we no longer necessarily have non-increasing returns.

**Example 3.2.3**

Consider a single physician with capacity  $2B$  and demand for prescheduled always equal to  $B$  and demand for same-day patients equal to  $2B$ . Then as we increase the number of slots offered to prescheduled patients from 0 to  $B$ , the revenue difference associated with each unit increase is negative, equal to  $R^p - R^s$ . Once  $N^p > B$  then the additional slots will not be used by prescheduled appointments and thus result in no change in the number of prescheduled or same-day patients seen, with a revenue difference of zero. For instance,  $R(B) - R(B - 1) = R^p - R^s < 0$ , but

$R(B + 1) - R(B) = 0$ . Therefore, we don't necessarily have decreasing returns once the returns are negative.

**Example 3.2.1 - Example 3.2.3** show that for general flexibility configurations, where prescheduled patients can be seen by multiple physicians and same-day patients are not fully flexible to see all of the physicians in the practice, the expected revenue does not always exhibit diminishing returns.

### 3.2.3 Configurations that satisfy the diminishing returns property

To establish a general analysis, let us consider any flexibility configuration containing prescheduled flexibility and same-day flexibility,  $\mathbf{A} = \mathbf{A}^p \cup \mathbf{A}^s$ . For any physician  $k$ , if we increase the reserved capacity  $N_k^p$  by one unit, let  $\Delta_k(N_1^p, \dots, N_m^p) = ER(N_1^p, \dots, N_k^p + 1, \dots, N_m^p) - ER(N_1^p, \dots, N_k^p, \dots, N_m^p)$  denote the expected revenue difference. From the point of view of revenue change for a particular demand realization, four possible events can happen when offering one additional slot to prescheduled patients of physician  $k$ :

- (i). No change in the number of prescheduled patients and the number of same-day patients seen, that is  $\Delta_k = 0$ .
- (ii). 1 more prescheduled patient is gained and no change in the number of same-day patients seen.  $\Delta_k = R^p$ .
- (iii). 1 more prescheduled patient is gained while 1 more same-day patient is lost.  $\Delta_k = R^p - R^s$ .
- (iv). No change in the number of prescheduled patients seen while 1 more same-day patient is gained.  $\Delta_k = R^s$ .

Observe, however, that under either no prescheduled flexibility or full same-day flexibility, only events (i)-(iii) can happen. The reason is as follows: (a) Under no prescheduled flexibility, the additional slot  $N_k^p + 1$  can only be used by patients of panel  $k$ . In that case, the available slots to same-day patients of any of the physicians

can never increase, and thus no additional patients can be seen. (b) Under full same-day flexibility, all available slots in the practice after prescheduled demand is fulfilled can be used by any of the same-day patients. Since the practice-wide number of available slots to same-day patients can never increase by increasing  $N_k$ , the number of same-day patients seen can never increase in this case.

In any other case, we can construct counter-examples in the same spirit as those above where event (iv) would happen and the diminishing returns property is violated. Fortunately, the two special cases, no prescheduled flexibility and full same-day flexibility, are the most commonly used in practices due to the critical requirement of continuity for prescheduled patients and the shortage of capacity in the whole clinic to provide quick access for acute needs. Let's thus focus on those two important cases and show that the objective function exhibits diminishing returns.

**Theorem 1.** *In settings with a dedicated configuration for prescheduled patients (given any configuration for same-day patients), and in settings with a fully flexible configuration for same-day patients (given any configuration for prescheduled patients), the expected practice-wide revenue exhibits diminishing returns. That is, the gains associated with increasing the booking limit of any of the physicians by one unit decrease as the initial booking limit vector grows.*

The proof of Theorem 1 is shown in the next section.

### 3.2.4 Proofs of Theorem 1

In this section, to provide intuition, we first analyze two simple cases (1) dedicated configuration for both prescheduled patients and same-day patients and (2) dedicated configuration for prescheduled patients and full flexibility for same-day patients. Then we extend the analysis to the general cases: (3) dedicated configuration for prescheduled patients and any flexibility configuration for same-day patients; (4)

any flexibility configuration for prescheduled patients and full flexibility for same-day patients.

To show that the diminishing returns property holds in the four cases below, we need to study how the function  $\Delta_k(N_1^p, \dots, N_m^p)$  changes as  $N_1^p, \dots, N_m^p$  increase.

**Case 1. Dedicated configuration for both prescheduled patients and same-day patients**

*Proof.* As explained before, if we increase  $N_k^p$  by one unit, only events (i)-(iii) can happen. We thus have the following mathematical expression of the difference in revenue:

$$\begin{aligned} \Delta_k &= 0 \times P[(i)] + R^p \times P[(ii)] + (R^p - R^s) \times P[(iii)] \\ &= R^p \times P[D_k^p > N_k^p \cap D_k^s < C_k - N_k^p] + (R^p - R^s) \times P[D_k^p > N_k^p \cap D_k^s \geq C_k - N_k^p] \end{aligned} \quad (3.1)$$

After simplifying, we have

$$\Delta_k = P[D_k^p > N_k^p] \{R^p - R^s \times P[D_k^s \geq C_k - N_k^p | D_k^p > N_k^p]\} \quad (3.2)$$

The first term,  $P[D_k^p > N_k^p]$ , is the complement of the cumulative distribution function and thus always non-increasing in  $N_k^p$ . Similarly,  $P[D_k^s \geq C_k - N_k^p]$  is non-decreasing in  $N_k^p$ , and thus the second term is non-increasing in  $N_k^p$  for any joint demand distribution where the prescheduled and same-day demands are independent, or more broadly as long as they are not strongly negatively correlated. Thus the revenue difference  $\Delta_k$ , if positive, is non-increasing in  $N_k^p$ . When the revenue difference get to be negative, it may increase to zero when  $P[D_k^p > N_k^p] = 0$  but is otherwise non-increasing.

For  $i \neq k$ , the revenue difference in  $k$ ,  $\Delta_k$  will not change as  $N_i^p$  increases for this dedicated configuration. The returns are thus non-increasing in any component of  $\mathbf{N}^p$ . □

From the above proof, we know that, when the  $\Delta_k > 0$  is positive, then  $\Delta_k > 0$  is a non-increasing function in any  $N_i^p$ . On the other hand, when the  $\Delta_k > 0$  is negative, it satisfies the diminishing returns property until it jumps up to 0 because additional prescheduled capacity will never be used. As a result, once the negative revenue difference is produced, it can never become positive again.

The above proof, holds when prescheduled and same-day demand are independent, or not heavily negatively correlated. However, we have not characterized the degree of negative correlation as the answer also depends on the demand distribution and the flexibility configuration.

**Case 2. Dedicated configuration for prescheduled patients and full flexibility for same-day patients**

*Proof.* Similarly, we have

$$\begin{aligned}
& \Delta_k(N_1^p, \dots, N_m^p) \\
&= R^p \times P[D_k^p > N_k^p \cap \sum_{j=1}^m D_j^s < \sum_{i=1}^M C_i - \sum_{\substack{i=1 \\ i \neq k}}^M \min[N_i^p, D_i^p] - N_k^p] \\
&+ (R^p - R^s) \times P[D_k^p > N_k^p \cap \sum_{j=1}^m D_j^s \geq \sum_{i=1}^M C_i - \sum_{\substack{i=1 \\ i \neq k}}^M \min[N_i^p, D_i^p] - N_k^p] \\
&= P[D_k^p > N_k^p] \left\{ R^p - R^s \times P\left[ \sum_{j=1}^m D_j^s \geq \sum_{i=1}^M C_i - \sum_{\substack{i=1 \\ i \neq k}}^M \min[N_i^p, D_i^p] - N_k^p \mid D_k^p > N_k^p \right] \right\}
\end{aligned}$$

Similarly to **case 1**, if we assume prescheduled demand and same-day demand are independent, or at least not heavily negatively correlated, the second term is non-increasing in  $N_i^p$  for all  $i = 1, 2, \dots, m$ . The diminishing returns property thus follows as in case 1. □

**Case 3. Dedicated configuration for prescheduled patients and any flexibility configuration for same-day patients**

Here we extend analysis to a more general case: prescheduled patients enjoy no flexibility and same-day patients follow any given flexibility configuration.

First, we recall some notation used in Chapter 2. Let  $\mathcal{A}$  denote the network to illustrate a given flexibility configurations, where link  $(j, i)$  belongs to  $\mathcal{A}$ , if patients from panel  $j$  can see physician  $i$ .

Second, we introduce the shortfall expression derived in [34] for a multi-plant multi-product flexible production system. The optimal value of shortfall associated with a flexibility configuration  $\mathcal{A}$  is shown to be

$$V(A) = \max_M \left\{ \sum_{j \in M} D_j - \sum_{i \in P(M)} C_i \right\}$$

where  $M$  is any subset (including the null set) of the index set  $\{1, 2, \dots, m\}$  and  $P(M)$  is the subset of plants that can produce at least one of the products in  $M$ . Extending this to our health care problem,  $M$  is any subset of panels and  $P(M)$  is the subset of physicians that can see patients of at least one of the panels in  $M$ . Thus,  $i \in P(M)$  if and only if there is at least one panel  $j \in M$  such that  $(j, i) \in A$ . Each term within the maximization is the difference between the demand for some subset of panels and the maximum capacity available from the physicians which are connected to that subset of panels.

The shortfall of same-day patients under a configuration  $\mathcal{A}$  only containing same-day flexibility is given by,

$$V_{\mathcal{A}}^s(N_1^p, \dots, N_m^p) = \max_{M_{\mathcal{A}}} \left\{ \sum_{j \in M_{\mathcal{A}}} D_j^s - \sum_{i \in P(M_{\mathcal{A}})} (C_i - \min[N_i^p, D_i^p]) \right\} \quad (3.3)$$

*Proof.* If we increase  $N_k^p$  by one unit, by 3.3, the shortfall of same-day patients becomes

$$V_{\mathcal{A}}^s(N_1^p, \dots, N_k^p + 1, \dots, N_m^p) = \max_{M_{\mathcal{A}}} \left\{ \sum_{j \in M_{\mathcal{A}}} D_j^s - \sum_{\substack{i \in P(M_{\mathcal{A}}) \\ P(M_{\mathcal{A}}): k \notin P(M_{\mathcal{A}})}} (C_i - \min[N_i^p, D_i^p]), \right. \\ \left. \sum_{j \in M_{\mathcal{A}}} D_j^s - \sum_{\substack{i \in P(M_{\mathcal{A}}), i \neq k \\ P(M_{\mathcal{A}}): k \in P(M_{\mathcal{A}})}} (C_i - \min[N_i^p, D_i^p]) - (C_k - \min[N_k^p + 1, D_k^p]) \right\}$$

Observe that the shortfall of same day patients,  $V_{\mathcal{A}}^s(N_1^p, \dots, N_m^p)$ , is non-decreasing in  $(N_1^p, \dots, N_m^p)$  since the number of open slots available to same day patients for any physician will never increase when increasing the booking limits of any of the physicians in the case of dedicated prescheduled patients. When increasing one of the booking limits by one, the shortfall will either go up by one or stay the same. For convenience, we introduce the function  $L_k^s(\mathcal{A}, \mathbf{N}^p)$  to illustrate the event that if we lose one same-day patient when we increase  $N_k^p$  to  $N_k^p + 1$ .

$$L_k^s(\mathcal{A}, \mathbf{N}^p) = \begin{cases} 1, & \text{if } V_{\mathcal{A}}^s(N_1^p, \dots, N_k^p + 1, \dots, N_m^p) > V_{\mathcal{A}}^s(N_1^p, \dots, N_m^p) \\ 0, & \text{otherwise} \end{cases} \quad (3.4)$$

Note that if the shortfall goes up by one when increasing a particular booking limit  $k$  from  $N_k^p$  to  $N_k^p + 1$ , given a vector of limits  $\mathbf{N}^p = (N_1^p, \dots, N_m^p)$ , then it will continue to go up by one for any vector of booking limits  $\mathbf{N}^{p'}$ , such that  $\mathbf{N}^{p'} > \mathbf{N}^p$ .  $L_k^s(\mathcal{A}, \mathbf{N}^p)$  is thus non-decreasing in  $N_i^p$ ,  $i = 1, 2, \dots, m$ .

Based on this notation, for this general case that there is no flexibility for prescheduled patients and full flexibility for same-day patients, we have

$$\begin{aligned} \Delta_k(\mathcal{A}^s, N_1^p, \dots, N_m^p) &= R^p \times P[D_k^p > N_k^p \cap L_k^s(\mathcal{A}, \mathbf{N}^p) = 0] \\ &+ (R^s - R^p) \times P[D_k^p > N_k^p \cap L_k^s(\mathcal{A}, \mathbf{N}^p) > 0] \\ &= R^p \times P[D_k^p > N_k^p] - R^s \times P[D_k^p > N_k^p \cap L_k^s(\mathcal{A}, \mathbf{N}^p) > 0] \\ &= P[D_k^p > N_k^p] \left\{ R^p - R^s \times P[L_k^s(\mathcal{A}, \mathbf{N}^p) > 0 | D_k^p > N_k^p] \right\} \end{aligned}$$

Since  $L_k^s(\mathcal{A}, \mathbf{N}^P)$  is non-decreasing in  $N_i^p$ ,  $i = 1, 2, \dots, m$ , we have as in the previous cases that the revenue difference  $\Delta_k$  is non-increasing in  $N_i^p$ ,  $i = 1, 2, \dots, m$ , for any distributions of prescheduled and same-day demands that are independent, or in general not heavily negatively correlated.  $\square$

In summary, if there is no flexibility for prescheduled patients, the non-increasing property is always true when the difference associated with one more reservation for physician  $k$  is greater than zero, regardless of the flexibility configuration of same-day patients.

**Case 4. Any flexibility configuration for prescheduled patients and full flexibility for same-day patients**

In this case, the shortfall of prescheduled patients under the setting containing any prescheduled flexibility configuration and full same-day flexibility  $\mathcal{A} = (\mathcal{A}^p, \mathcal{A}^s)$  is given by  $V_{\mathcal{A}}^p(N_1^p, \dots, N_m^p)$ , here  $V_{\mathcal{A}}^p(N_1^p, \dots, N_m^p) = \max_{M_{\mathcal{A}}} \{ \sum_{j \in M_{\mathcal{A}}} D_j^p - \sum_{i \in P(M_{\mathcal{A}})} N_i^p \}$ .

*Proof.* If we increase  $N_k^p$  by one unit, then the shortfall of prescheduled patients becomes

$$\begin{aligned} & V_{\mathcal{A}}^p(N_1^p, \dots, N_k^p + 1, \dots, N_m^p) \\ &= \max_{M_{\mathcal{A}}} \left\{ \sum_{j \in M_{\mathcal{A}}} D_j^s - \sum_{\substack{i \in P(M_{\mathcal{A}}) \\ P(M_{\mathcal{A}}): k \notin P(M_{\mathcal{A}})}} N_i^p, \sum_{j \in M_{\mathcal{A}}} D_j^s - \sum_{\substack{i \in P(M_{\mathcal{A}}) \\ P(M_{\mathcal{A}}): k \in P(M_{\mathcal{A}})}} N_i^p - 1 \right\} \end{aligned}$$

For convenience, we introduce the indicator function  $G_k^p(\mathcal{A}, \mathbf{N}^P)$  to signal when an additional prescheduled patient can be seen as  $N_k^p$  is increased to  $N_k^p + 1$ .

$$G_k^p(\mathcal{A}, \mathbf{N}^P) = \begin{cases} 1, & \text{if } V_{\mathcal{A}}^p(N_1^p, \dots, N_k^p + 1, \dots, N_m^p) < V_{\mathcal{A}}^p(N_1^p, \dots, N_m^p) \\ 0, & \text{otherwise} \end{cases} \quad (3.5)$$

Let  $uC(N_1^p, N_2^p, \dots, N_m^p)$  denote the capacity used by prescheduled patients under a given booking limit vector  $\mathbf{N}^p = [N_1^p, N_2^p, \dots, N_m^p]$  and a particular prescheduled demand realization; then we have

$$\begin{aligned}
& \Delta_k(\mathcal{A}^p, N_1^p, \dots, N_m^p) \\
&= R^p \times P[G_k^p(\mathcal{A}, \mathbf{N}^p) > 0 \cap \sum_{j=1}^m D_j^s < \sum_{i=1}^m C_i - uC] \\
&+ (R^p - R^s) \times P[G_k^p(\mathcal{A}, \mathbf{N}^p) > 0 \cap \sum_{j=1}^m D_j^s \geq \sum_{i=1}^m C_i - uC] \\
&= R^p \times P[G_k^p(\mathcal{A}, \mathbf{N}^p) > 0] - R^s \times P[G_k^p(\mathcal{A}, \mathbf{N}^p) > 0 \cap \sum_{j=1}^m D_j^s \geq \sum_{i=1}^m C_i - uC] \\
&= P[G_k^p(\mathcal{A}, \mathbf{N}^p) > 0] \times \{R^p - R^s \times P[\sum_{j=1}^m D_j^s \geq \sum_{i=1}^m C_i - uC | G_k^p(\mathcal{A}, \mathbf{N}^p) > 0]\}
\end{aligned}$$

The probability of gaining one further prescheduled patient,  $P[G_k^p(\mathcal{A}, \mathbf{N}^p) > 0]$ , is clearly non-increasing in all  $N_i^p$  for  $i = 1, 2, \dots, m$ . Observe also that  $P[\sum_{j=1}^m D_j^s \geq \sum_{i=1}^m C_i - uC]$  is always non-decreasing in  $N_i^p$ , because increasing any  $N_i^p$  by one unit will never decrease the number of prescheduled patients seen,  $uC$ . In that case, when  $\Delta_k > 0$ ,  $\Delta_k(N_1^p, \dots, N_m^p)$  is always the product of two terms, both of which are always non-increasing in  $N_i^p$ . Therefore, the non-increasing diminishing returns property holds also for this case, as long as the returns are positive and the prescheduled and same-day demands are not heavily negatively correlated.  $\square$

The proofs of the four cases above follow the same argument. First, the probability of gaining one additional prescheduled patient when offering them one more of the slots of physician  $k$  can never increase as the original number of slots available to them through each of the physicians in the practice increases; that is,  $P[G_k^p(\mathcal{A}, \mathbf{N}^p) > 0]$  is non-increasing in  $N_1^p, N_2^p, \dots, N_m^p$ . Second, the probability of losing one additional same-day patient as the number of slots that they can use decreases when one more prescheduled is seen will never decrease as long as the demands of prescheduled and

same-day patients are independent or positively correlated. That is, when  $\mathbf{D}^p, \mathbf{D}^s$  are not negatively correlated,  $P[L_k^s(\mathcal{A}, \mathbf{N}^p) > 0 | G_k^p(\mathcal{A}, \mathbf{N}^p) > 0]$  will never decrease as  $N_i^p$  increases.

In the proof of Theorem 1, we assume independence or positive correlation of prescheduled and same-day demands to prove the diminishing returns property; however, when demands are heavily negatively correlated, there are cases where the conditional probability  $P[L_k^s(\mathcal{A}, \mathbf{N}^p) > 0 | G_k^p(\mathcal{A}, \mathbf{N}^p) > 0]$  decreases, as shown in the following example.

**Example 3.2.4**

Negative correlation makes increments potentially increasing, even for a single physician when the increments are positive. Consider the same setting as in **Example 3.2.3**, where a single physician has capacity  $2B$ , but with two possible demand realizations;  $(B, 2B)$  with probability  $p$  and  $(2B, 0)$  with probability  $1 - p$ . The revenue increments behave just as in **Example 3.2.3** for the first demand realization. For the second, every time  $N^p$  is increased, the revenue grows by  $R^p$ . Thus  $ER(B) - ER(B - 1) = p(R^p - R^s) + (1 - p)R^p < (1 - p)R^p = ER(B + 1) - ER(B)$ . Observe that for  $p$  sufficiently small the expected revenue difference is positive. In this case the demands are negatively correlated.

**3.3 Impact of flexibility on booking limits**

In the previous section, we prove that the diminishing returns property holds for two cases with a fixed level of flexibility: (i) configurations that allow only same-day flexibility [that is, no prescheduled flexibility and any same-day flexibility], or (ii) configurations that allow any prescheduled flexibility and full same-day flexibility. In this section, we study the impact of flexibility on the optimal booking limits just under those two sets of configurations, for which the diminishing return property holds.

In fact, these two special groups of configurations are the most meaningful for primary care practices. Due to the shortage of primary care providers, same-day patients are frequently assigned to see other physicians when their own physicians are not available when the call comes in. Moreover, prescheduled patients are sometimes assigned across panels due to multiple factors such as patients' preferences, time of available slots, the urgency of the request of appointment, etc. Although these patients usually require higher continuity, some do not mind to trade continuity for timely access temporarily. Consequently, both prescheduled flexibility and same-day flexibility can exist in primary care clinics. What is the impact of these two types of flexibility on the optimal capacity allocation to prescheduled vs. same-day patients? Prescheduled flexibility and same-day flexibility are designed similarly in our model. Nevertheless, our analysis shows different trends in the optimal booking limits  $\mathbf{N}^{\mathbf{P}}$  as the amount of flexibility increases, depending on whether it is prescheduled or same-day flexibility that is added. This is because of the sequential assignment, first of prescheduled demand up to the booking limits, and then of same-day demand to all remaining capacity, including unused prescheduled capacity.

**Definition 2.** *A configuration  $\mathcal{A}'$  is said to be more flexible than a configuration  $\mathcal{A}$  if it contains all the panel-physician links of  $\mathcal{A}$ ; that is,  $\mathcal{A} \subseteq \mathcal{A}'$ .*

**Theorem 2.** *Given full flexibility for same-day patients, the optimal  $\mathbf{N}^{\mathbf{P}}$  is non-increasing as the flexibility for prescheduled patients increases.*

*Proof.* Recall that  $G_k^p(\mathcal{A}, \mathbf{N}^{\mathbf{P}})$  and  $L_k^s(\mathcal{A}, \mathbf{N}^{\mathbf{P}})$  represent the events of gaining one more prescheduled patient and losing one more same-day patient, respectively, when the number of slots of physician  $k$  available to prescheduled patients increases from  $N_k^p$  to  $N_k^p + 1$  in a system under configuration  $\mathcal{A}$ .

Given any flexibility configuration available to prescheduled patients combined with full flexibility for same-day patients, we have the following expression of the

revenue difference function  $\Delta_k$  associated with the increase from  $N_k^p$  to  $N_k^p + 1$ .

$$\begin{aligned}\Delta_k &= P(G_k^p(\mathcal{A}, \mathbf{N}^p) > 0) \times [R^p - R^s P(L_k^s(\mathcal{A}, \mathbf{N}^p) > 0 | G_k^p(\mathcal{A}, \mathbf{N}^p) > 0)] \\ &= P(G_k^p(\mathcal{A}, \mathbf{N}^p) > 0) \times [R^p - R^s Q]\end{aligned}$$

where  $Q = P(L_k^s(\mathcal{A}, \mathbf{N}^p) > 0 | G_k^p(\mathcal{A}, \mathbf{N}^p) > 0)$ . Given full flexibility for same-day patients,

$$Q = P\left(\sum_{i=1}^m D_i^s > \sum_{i=1}^m C_i - uC | G_k^p(\mathcal{A}, \mathbf{N}^p) > 0\right),$$

where  $uC$  is the used or unavailable capacity after prescheduled demand has been assigned. We will keep the notation  $\Delta$  for the original flexibility configuration  $\mathcal{A}$  and use  $\Delta'$  to denote the incremental revenue under a more flexible configuration  $\mathcal{A}'$ .

Observe that:

(1) In the fully flexible same-day configuration, the probability of losing a patient only depends on the system-wide number of slots left over after the allocation of capacity to prescheduled patients, that is,  $\sum_{i=1}^m C_i - uC$ . The total number of slots  $uC$  used by prescheduled patients under a set of booking limits  $\mathbf{N}^p$  is always no lower in a more flexible configuration, where more allocation options are available. Therefore, for a more flexible configuration  $\mathcal{A}'$ , such that  $\mathcal{A} \subseteq \mathcal{A}'$ , we always have  $Q_{\mathcal{A}} \leq Q_{\mathcal{A}'}$ .

(2) As the vector  $\mathbf{N}^p$  increases, the term  $P(G_k^p(\mathcal{A}, \mathbf{N}^p) > 0)$  could become 0 before the left term  $[R^p - R^s Q]$  becomes non-positive; in this case, further increasing the booking limit will not affect the revenue, and a solution with larger  $N_k^p$  would continue being optimal. This implies that if increasing one component of the booking limit has a positive revenue for  $\mathcal{A}'$ , then it will have a non-negative revenue for  $\mathcal{A}$ .

Assume now that booking limits  $\mathbf{N}^{p*}$  and  $\mathbf{N}^{p'}$  with  $\mathbf{N}^{p*} < \mathbf{N}^{p'}$  are optimal for  $\mathcal{A}$  and  $\mathcal{A}'$ , respectively, and  $\mathbf{N}^{p*}$  is a maximal optimal vector (meaning that for any  $\mathbf{N}^p > \mathbf{N}^{p*}$  the revenue will be strictly lower and thus  $\Delta_k(\mathbf{N}^{p*}) < 0$  for all

physicians  $k = 1, 2, \dots, m$ ). Using the facts observed above, we can easily arrive to a contradiction:

Since  $\mathbf{N}^{\mathbf{P}^*} < \mathbf{N}^{\mathbf{P}'}$ , and the revenue function exhibits decreasing returns, there must be a physician  $k$  such that  $\Delta'_k(\mathbf{N}^{\mathbf{P}^*}) > 0$ . This means that  $P(G_k^p(\mathcal{A}', \mathbf{N}^{\mathbf{P}^*}) > 0) > 0$  and  $Q_{\mathcal{A}'} < R^p/R^s$ . In turn, for  $\mathcal{A}$  this means that either  $P(G_k^p(\mathcal{A}, \mathbf{N}^{\mathbf{P}}) > 0) = 0$  or  $(P(G_k^p(\mathcal{A}, \mathbf{N}^{\mathbf{P}}) > 0) > 0$  and  $Q_{\mathcal{A}} < Q_{\mathcal{A}'} < R^p/R^s$ ). In both cases, increasing  $N_k^{p^*}$  to  $N_k^{p^*} + 1$  will result in greater or equal revenue, contradicting the maximality of the optimal solution  $\mathbf{N}^{\mathbf{P}^*}$ .

□

When utilization is high, the more flexible system will be able to accommodate the prescheduled demand with fewer slots, and will have a higher probability of filling the available prescheduled slots because of the additional allocation options flexibility affords. Thus, it must set tighter booking limits to ensure enough capacity is available for the same-day patients. When utilization is low, on the other hand, having extra capacity available to prescheduled patients in the less flexible system comes at no cost, since those slots will simply be left over and available to same-day patients.

The following theorem shows that the optimal booking limits exhibit very different behavior for the case where only same-day patients enjoy flexibility.

**Theorem 3.** *Assume no flexibility is allowed in serving prescheduled patients and normal distributions of same-day demands, independent from each other, from prescheduled demands, and across symmetric doctors, with equal capacities and identically distributed same-day demands. As the flexibility to accommodate same-day patients increases from dedicated to fully flexible, we have that:*

1) *If  $R^p/R^s \leq 0.5$  the optimal booking limit  $\mathbf{N}^{\mathbf{P}}$  is non-decreasing regardless of the level of system utilization. 2) *If  $R^p/R^s > 0.5$  the optimal booking limit  $\mathbf{N}^{\mathbf{P}}$  is non-increasing when the system utilization is sufficiently high, and non-decreasing when the system utilization is low.**

*Proof.* Let the same-day demands be i.i.d. normal distributions with mean  $\mu$  and standard deviation  $\sigma$ . We will denote with a  $D$  the dedicated configuration and all its associated parameters, and with a  $F$  the same-day fully flexible configuration and all its parameters. For instance,  $\mathbf{N}_D^{\mathbf{P}^*}$  and  $\mathbf{N}_F^{\mathbf{P}^*}$  are the optimal booking limits for the dedicated and fully flexible configurations respectively. Since the setting is symmetric, we assume that all doctors have the same capacity,  $C$ , and will be given the same optimal booking limit,  $N$ ; that is,  $N_1^p = N_2^p = \dots = N_m^p = N$ . This is not true in general due to the integrality requirements of the booking limit, but we make this continuous approximation to simplify the problem into having a single booking limit decision  $N$ .

Let the optimal booking limits be  $\mathbf{N}_D^{\mathbf{P}^*} = (N_D, N_D, \dots, N_D)$  and  $\mathbf{N}_F^{\mathbf{P}^*} = (N_F, N_F, \dots, N_F)$ . Because of the diminishing returns property, they will occur at the points where the revenue difference becomes zero, that is, for  $N_D$  and  $N_F$  such that

$$Q_D = P[D_k^s > C - N_D] = \frac{R^p}{R^s}$$

$$Q_F = P\left[\sum_{i=1}^m D_i^s > \sum_{\substack{i=1 \\ i \neq k}}^m (C - \min\{D_i^p, N_F\}) + (C - N_F)\right] = \frac{R^p}{R^s}$$

respectively.

Applying the usual transformation into standard normal distributions we get that  $N_D$  and  $N_F$  must satisfy:

$$P\left[Z = \frac{D_k^s - \mu}{\sigma} > \frac{C - N_D - \mu}{\sigma}\right] = \frac{R^p}{R^s}$$

$$P\left[Z = \frac{\frac{\sum_{i=1}^m D_i^s}{m} - \mu}{\sigma/\sqrt{m}} > \frac{\sum_{\substack{i=1 \\ i \neq k}}^m (C - \min\{D_i^p, N_F\}) + (C - N_F) - m\mu}{\sqrt{m}\sigma}\right] = \frac{R^p}{R^s}$$

respectively.

Observe that the only difference between the expressions of  $Q_D$  and  $Q_F$  after standardizing the probability terms is their right hand side (RHS). At optimality the RHS's need to be equal since they lead to the same standard normal probability. Comparing the two RHS's we will be able to establish the relationship between  $N_D$  and  $N_F$ .

**Case 1:**  $\frac{R^p}{R^s} \leq 0.5$  In this case,  $P[Z > RHS] = \frac{R^p}{R^s} \leq 0.5$  implies that  $RHS \geq 0$  at optimality, and we can compare the right hand sides as follows:

$$\begin{aligned} \frac{\sum_{\substack{i=1 \\ i \neq k}}^m (C - \min\{D_i^p, N\}) + (C - N) - m\mu}{\sqrt{m}\sigma} &> \frac{\sum_{i=1}^m (C - N) - m\mu}{\sqrt{m}\sigma} \\ &= \frac{C - N - \mu}{\sigma/\sqrt{m}} > \frac{C - N - \mu}{\sigma} > 0 \end{aligned}$$

Therefore, for the same booking limit  $N$  the RHS for the dedicated system is positive and lower than the RHS of the flexible system. Since the RHS's are decreasing in  $N$ , and at optimality they need to be equal, we have that  $N_F > N_D$ .

**Case 2:**  $\frac{R^p}{R^s} > 0.5$  In this case,  $P[Z > RHS] = \frac{R^p}{R^s} > 0.5$  implies that  $RHS < 0$  at optimality, and we need to consider different subcases in order to compare the right hand sides.

**Case 2a: High System Utilization** In this case, with high probability, all the slots allocated to prescheduled patients will be used and thus the capacity available to same-day patients is  $\sum_{i=1}^m (C_i - N_i^p) = m(C - N)$ .

This allows us to simplify the expression of  $Q_F$  at the optimal booking limit,  $N_F$ :

$$Q_F = P\left[Z > \frac{C - N_F - \mu}{\sigma/\sqrt{m}}\right] = \frac{R^p}{R^s}$$

respectively.

Since  $\frac{R^p}{R^s} > 0.5$  and thus the RHS's are negative, we have that  $\frac{C-N-\mu}{\sigma} > \frac{C-N-\mu}{\sigma/\sqrt{m}}$ , and thus  $N_D > N_F$  to make the terms equal.

**Case 2b: Low utilization** As the system utilization decreases, the probability of prescheduled slots going unfilled and available for same day patients increases. Then we have that  $D_i^p < N$ , for any  $i = 1, 2, \dots, m$ , with a high probability, and the capacity available to same day patients in the flexible system is  $\sum_{\substack{i=1 \\ i \neq k}}^m (C - \min\{D_i^p, N\}) + C_k - N_k^p \geq m(C - N)$  with an even greater probability. As a result for sufficiently low utilization, so that the  $\sum(N - D_i^p)^+$  is large, we have

$$\begin{aligned} \frac{\sum_{\substack{i=1 \\ i \neq k}}^m (C - \min\{D_i^p, N\}) + (C - N) - m\mu}{\sqrt{m}\sigma} &= \frac{C - N - \mu}{\sigma/\sqrt{m}} + \frac{\sum_{\substack{i=1 \\ i \neq k}}^m (N - D_i^p)^+}{\sqrt{m}\sigma} \\ &> \frac{C - N - \mu}{\sigma} \end{aligned}$$

At optimality then  $N_F > N_D$  must hold.  $\square$

When the relative revenue of same-day patients is very high, then flexibility always allows the allocation of more slots to prescheduled patients. When the revenue of prescheduled patients is no less than 50% of that of same the patients, then it depends on the load on the system. As the utilization of the system increases, there is a greater demand for same-day appointments, which still produce higher revenue than prescheduled appointments. This prompts the practice, under any flexibility configuration for same-day appointments, to further restrict the number of prescheduled appointments offered, i.e., to reduce their booking limits, so that more same-day appointments can be seen. The more flexible configurations will offer even fewer prescheduled appointments, thus reserving more capacity for open access, since there is a higher probability of fully using the additional capacity when it is shared across same-day appointments in the practice. When system utilization is low, the need

for booking limits decreases, especially when same-day patients are flexibly shared; thus, higher optimal booking limits for prescheduled patients are optimal for the more flexible configurations.

Note that, Theorem 3 requires same-day demands to be i.i.d. normal distributions and only shows the property of the optimal booking limit under dedicated configuration and full flexibility configuration. In addition, computational results shown in Section 4.2.3 support the generality of Theorem 3 to other partial flexibility configurations and to other demand distributions (Poisson distributions assumed for both prescheduled and same-day demands).

These monotonicity patterns, allow us to identify bounds on the booking limits of any general flexibility configuration. As mentioned in the discussion of our modeling approach, the booking limits are hard to find for general configurations but relatively easy for the extreme ones.

**Corollary 1.** *The optimal booking limit associated with any flexibility configurations that have either dedicated prescheduled or fully flexible same-day patients is higher than or equal to the minimum of the booking limits associated with the two extreme configurations:*

- I. *No flexibility is allowed for either patient group.*
- II. *All patients can be flexibly shared.*

*Proof.* Recall that D denotes dedicated configuration, A denotes any flexibility configuration and F denotes full flexibility configuration. For example,  $(\mathbf{NP})_{\mathbf{DF}}^*$  denotes the optimal capacity allocation for a practice with dedicated flexibility for prescheduled patients and full flexibility for same-day patients.

By Theorem 2, we have  $(\mathbf{NP})_{\mathbf{FF}}^* \leq (\mathbf{NP})_{\mathbf{AF}}^* \leq (\mathbf{NP})_{\mathbf{DF}}^*$ .

By Theorem 3 and computational results shown in Section 4.2.3, when utilization is sufficiently high and  $R^p/R^s \geq 0.5$ ,  $(\mathbf{NP})_{\mathbf{DF}}^* \leq (\mathbf{NP})_{\mathbf{DA}}^* \leq (\mathbf{NP})_{\mathbf{DD}}^*$ , that is,

$(\mathbf{N}^{\mathbf{P}})_{\mathbf{FF}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{AF}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{DF}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{DA}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{DD}}^*$ . Under this case,  $(\mathbf{N}^{\mathbf{P}})_{\mathbf{FF}}^*$  always provides a lower bound for both  $(\mathbf{N}^{\mathbf{P}})_{\mathbf{DA}}^*$  and  $(\mathbf{N}^{\mathbf{P}})_{\mathbf{AF}}^*$ .

When the utilization decreases beyond a threshold point (sufficiently low) or  $R^p/R^s \leq 0.5$ , then  $(\mathbf{N}^{\mathbf{P}})_{\mathbf{DD}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{DA}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{DF}}^*$ . As  $(\mathbf{N}^{\mathbf{P}})_{\mathbf{FF}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{AF}}^* \leq (\mathbf{N}^{\mathbf{P}})_{\mathbf{DF}}^*$  always holds,  $\min\{(\mathbf{N}^{\mathbf{P}})_{\mathbf{DD}}^*, (\mathbf{N}^{\mathbf{P}})_{\mathbf{FF}}^*\}$  must provide a lower bound for  $(\mathbf{N}^{\mathbf{P}})_{\mathbf{DA}}^*$  and  $(\mathbf{N}^{\mathbf{P}})_{\mathbf{AF}}^*$ .

In summary, for any configuration  $\mathcal{A}$  with either dedicated flexibility for prescheduled patients or full flexibility for same-day patients,  $\min\{(\mathbf{N}^{\mathbf{P}})_{\mathbf{DD}}^*, (\mathbf{N}^{\mathbf{P}})_{\mathbf{FF}}^*\}$  must provide a lower bound for  $(\mathbf{N}^{\mathbf{P}})_{\mathcal{A}}^*$ .  $\square$

Corollary 1 leads to an easy-to-calculate bound, which provides us with a good starting point for possible heuristic procedures discussed in next section.

### 3.4 Greedy search heuristic

In this section, we propose a greedy search heuristic to solve the capacity allocation problem, and discuss the generated solution quality and computational performance. The greedy algorithm starts with no slots available to prescheduled patients and will myopically add one extra slot at a time to the physician for whom it will result in the maximum increase in revenue. It will stop when no positive returns can be obtained by further adding slots to any of the physicians. The diminishing returns property shown to hold for the most practical cases, where either prescheduled patients are dedicated or same-day patients are fully flexible, implies that the greedy algorithm provides the exact optimal solution in the case of a single physician, and suggests that it should work well in general. While we have not been able to prove that the greedy algorithm always finds the optimal solution, it did identify the optimal solution to the capacity allocation problem in every single case tested in our extensive computational experiments, under a wide range of parameter values; please see Section 3.4.2.1 for details.

The greedy heuristic provides us with an efficient alternative to solve moderate size problems, whereas for large instances we will use sample average approximation methods to solve the mathematical formulations presented in Chapter 2. In what follows we investigate the performance of the greedy search heuristic.

### 3.4.1 Iterative search procedure

The greedy search heuristic iteratively searches the possible capacity allocation decisions as follows.

*Step 1: Initialization.* Set  $N_k^p = 0$  for all  $k = 1, 2, \dots, m$  and  $N = N_1^p + N_2^p + \dots + N_m^p = 0$

*Step 2: Iteration.*

a) Increase  $N$  by 1

b) Calculate  $ER(N_1^p, \dots, N_k^p + 1, \dots, N_m^p)$  for each  $k = 1, 2, \dots, m$ .

c) Find  $\Delta_N = \max_k \Delta_k(N_1^p, \dots, N_m^p)$ , where  $\Delta_k(N_1^p, \dots, N_m^p) = ER(N_1^p, \dots, N_k^p + 1, \dots, N_m^p) - ER(N_1^p, \dots, N_k^p, \dots, N_m^p)$ , and let  $i$  be the maximum argument.

d) If  $\Delta_N \leq 0$ , go to Step 3.

e) Increase  $N_i^p$  by 1 and go to Step 2.

*Step 3: Termination.* The current solution  $\mathbf{N}^p = (N_1^p, N_2^p, \dots, N_m^p)$  is given.

The number of available prescheduled appointment slots is increased by one unit at a step, greedily for the physician that results in the highest revenue increase, until the system revenue can no longer be improved. If a lower bound on the number of prescheduled slots to make available can be derived, then it could be used in the initialization step. The analysis of the behavior of  $N_p$  as flexibility increases could be used to derive lower bounds; for instance, we could use the knowledge of the booking limits under say a dedicated configuration to derive initial setting for other flexibility configurations.

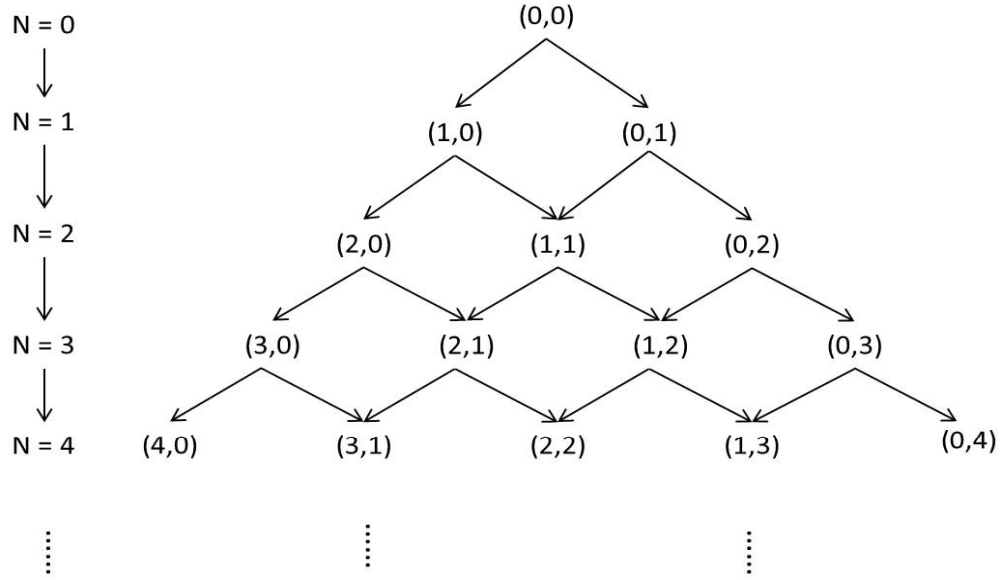
### 3.4.2 Performance of greedy search heuristic

The benefit of the greedy algorithm is the reduction in computational effort. In fact, a full search of the solution space would entail considering  $C_1 \times C_2 \times \dots \times C_m$  possible vectors  $N^p$ . However, the greedy algorithm will only visit  $m$  booking limits in each iteration, for at most  $C = \sum_{i=1}^m C_i$  (the total number of slots available system-wide) iterations, which significantly reduces the search space.

We could also see the full or exhaustive search algorithm as iterating from  $N = 0, \dots, C$ . At each step, there are  $N$  slots that can be distributed in any way among the  $m$  physicians. The number of  $\mathbf{N}^p$  vectors to explore at each iteration can be calculated as follows. For any  $N > 0$ , let  $i$  be the number of physicians with no available prescheduled slots, where  $i$  may vary from  $i = 0$  to  $i = m - 1$ . The remaining  $m - i$  physicians have to share the  $N$  slots. The number of ways in which the  $N$  slots may be assigned to the  $(m - i)$  physicians is  $\frac{(N-1)!}{(m-i-1)!(N-m-i)!}$  (combinations of  $N-1$ , taken in groups of  $m-i-1$ ) [This is because we can pose the problem as finding  $m - i - 1$  breakpoints in  $1, 2, \dots, N - 1$  that represent the last slot out of the  $N$  that is assigned to physician  $k$  for each of the first  $m - i - 1$  physicians. The remaining slots will be assigned to the last physician.] Then the total number of vectors to visit at each iteration then is  $\sum_{i=1}^{m-1} \left[ \frac{m!}{(m-i)!i!} \times \frac{(N-1)!}{(m-i-1)!(N-m-i)!} \right]$ .

Greedy algorithms may yield potentially bad solutions for some problems, as they may get stuck in a local optimum and never get close to optimality. We thus need to evaluate the performance of greedy algorithm for our capacity allocation problem. Under a fully dedicated configuration for both prescheduled and same-day patients, the greedy algorithm safely yields the optimal solution since the problem can be split into  $m$  problems in a single dimension. The risk of missing a global optimum only occurs under multiple dimensions, when the various resources interact under a flexible configuration. To illustrate our discussion, Figure 3.1 shows the spanning tree of the

search space for a 2-physician practice. We assume the practice allows same-day patients to be flexibly shared.



**Figure 3.1.** Spanning tree of the search space of the capacity allocation problem : 2-physician practice

Under a particular value of  $N$ , the nodes on the corresponding line represent all the possible vector solutions  $(N_1^p, N_2^p)$  at this step. Instead of checking all the nodes at each step  $N$ , the greedy algorithm only tours the nodes branched from the previous local optimum at step  $N - 1$ . As a result the greedy algorithm might miss the global optimal node. For example, suppose  $\Delta_1(0,0) > \Delta_2(0,0)$ ,  $\Delta_1(1,0) > \Delta_2(1,0)$  and  $\Delta_1(2,0) > \Delta_2(2,0)$ , the greedy algorithm reports locally optimal solutions to be  $(1,0) \rightarrow (2,0) \rightarrow (3,0)$ . Some nodes, such as  $(0,3)$  will never be visited in the greedy search. Could such a node be the global optimal solution? In fact, under symmetric cases, this is not possible. Under symmetric configurations and demands (the amount of demands for each physician are identical), for node  $(0,0)$ ,  $\Delta_1(0,0) = \Delta_2(0,0)$ , the greedy algorithm will increase the first component by one unit, then due to symmetry and diminishing returns,  $\Delta_1(1,0) < \Delta_2(1,0)$ , so the locally optimal solution obtained from the greedy algorithm is  $(1,1)$ , which is also global optimal solution at this step.

In other words, the greedy algorithm never deviates from the middle axle of the spanning tree, where the global optimal solution lies in the symmetric case.

For asymmetric physician demands, however, it is not so clear that the optimum at a particular step, for a given value of  $N$ , will always be found in the branches out of the previous optimal node at step  $N - 1$ . The diminishing returns property suggests that this should be the case in general. Furthermore, an exhaustive search over all nodes for each value of  $N$  results in the same path through the tree as that followed by the greedy algorithm in all of our computational tests. We thus conjecture that the greedy algorithm is in fact exact, and always finds the optimal solution to the capacity allocation problem. In what follows, we describe the wide range of parameters tested in our computational experiments and provide a sample of the results, which provide insight as to how the solution evolves at each step  $N$ .

### 3.4.2.1 Computational results

To test the conjecture that the greedy algorithm always finds the optimal solution to the capacity allocation problem, we run computational experiments under a range of asymmetric demand scenarios, which cover most practice situations.

We focus on a 3-physician practice and assume that prescheduled patients are dedicated and same-day patients are fully flexibly shared. Capacity for each physician is identical to be set to 24 slots per day (roughly 8 hours because one slot generally takes 20 mins). To test the performances of greedy algorithm under different asymmetries in demand scenarios, we run computational experiments over four different P/S ratio settings: (1) 6/12 8/16 10/20, (2) 12/6 16/8 20/10, (3) 8/16 12/12 16/8, and (4) 4/20 12/12 20/4. Each P/S ratio setting is repeated under 3 different workloads (80%, 100% and 120%) [Refer to section 1.1 for more details about the definitions of P/S ratios and workload, and how we create the asymmetry of practices].

In fact, all the computational results support the conjecture that greedy algorithm yields optimal solution for this capacity allocation problem. In that case, it is reasonable to suggest greedy algorithm as an efficient heuristic for our capacity allocation problems. Interestingly, not only the final destination (optimal solution) of a full search and that of a greedy search are same, the search path under these two search method always matches.

For example, Table 3.1 shows the results under the case that P/S ratio setting is 6/12 8/16 10/20 and workload is 120%. The columns of ‘NP\*’ present the optimal solutions under restricted N ( $N = 0, 1, \dots, 72$ ) and the columns of ‘revenue’ show the associated revenue. A greedy search will start from  $\mathbf{N}^{\mathbf{P}} = [0, 0, 0]$ . Due to the demand input,  $\Delta_3(0, 0, 0) > \Delta_1(0, 0, 0)$  and  $\Delta_3(0, 0, 0) > \Delta_2(0, 0, 0)$ , the third component of  $\mathbf{N}^{\mathbf{P}}$  needs to be increased by one, resulting a solution just matches the optimal solution (from a full search) with restriction  $N = 1$ . Then, as  $\Delta_2(0, 0, 1) > \Delta_1(0, 0, 1)$  and  $\Delta_2(0, 0, 1) > \Delta_3(0, 0, 1)$ , the second component of  $\mathbf{N}^{\mathbf{P}}$  needs to be increased by one due to greedy algorithm, resulting a solution matches the optimal solution (from a full search) with restriction  $N = 2$  again. Continue the analysis, these observations hold until the global optimal solution is reached ( $N = 23$  and  $\mathbf{N}^{\mathbf{P}^*} = [5, 8, 10]$  in this case). Beyond global optimal point, there is no such regularity.

**Table 3.1.** Optimal solutions under given amount of N : asymmetric 6/12 8/16 10/20, 120% workload

N	NP*			Revenue	N	NP*			Revenue	N	NP*			Revenue
0	0	0	0	51.22397										
1	0	0	1	51.94617	25	7	8	10	60.81347	49	24	19	6	60.65315
2	0	1	1	52.6597	26	9	8	9	60.79365	50	24	20	6	60.65267
3	0	1	2	53.36243	27	10	8	9	60.78013	51	24	21	6	60.65244
4	0	2	2	54.05165	28	12	7	9	60.77311	52	24	22	6	60.65234
5	0	2	3	54.72485	29	13	7	9	60.77006	53	24	23	6	60.65229
6	1	2	3	55.37871	30	14	7	9	60.76849	54	24	24	6	60.65228
7	1	2	4	56.00928	31	15	7	9	60.76774	55	24	24	7	60.64787
8	1	3	4	56.61369	32	16	7	9	60.76742	56	24	24	8	60.62078
9	2	3	4	57.18599	33	17	7	9	60.76728	57	24	24	9	60.57699
10	2	3	5	57.72528	34	18	7	9	60.76723	58	24	24	10	60.52326
11	2	4	5	58.22696	35	19	7	9	60.76721	59	24	24	11	60.46636
12	2	4	6	58.68454	36	20	7	9	60.7672	60	24	24	12	60.4121
13	3	4	6	59.09718	37	21	7	9	60.7672	61	24	24	13	60.3646
14	3	5	6	59.46737	38	22	7	9	60.7672	62	24	24	14	60.32603
15	3	5	7	59.78735	39	23	7	9	60.7672	63	24	24	15	60.29677
16	3	6	7	60.05745	40	24	7	9	60.7672	64	24	24	16	60.27591
17	4	6	7	60.28286	41	24	8	9	60.7586	65	24	24	17	60.26189
18	4	6	8	60.46567	42	24	9	9	60.73391	66	24	24	18	60.25296
19	4	7	8	60.60318	43	24	11	8	60.70468	67	24	24	19	60.24757
20	4	7	9	60.70303	44	24	13	7	60.6838	68	24	24	20	60.24446
21	5	7	9	60.77438	45	24	14	7	60.67038	69	24	24	21	60.24275
22	5	8	9	60.81369	46	24	15	7	60.66108	70	24	24	22	60.24185
*23	5	8	10	60.82961	47	24	17	6	60.65597	71	24	24	23	60.2414
24	6	8	10	60.82929	48	24	18	6	60.65412	72	24	24	24	60.24118

Generally, under all the tested scenarios, we find the optimal solution obtained under restricted N is always branched from the optimal solution from previous step (i.e. if two components of the optimal solution from previous step are changed, the yielded solution will never be optimal for current step). A rigorous analysis to further analyze this observation could be an interesting topic for our future study.

### 3.4.3 Analytical method to calculate expected performances

Once again we need to make use of Jordan and Graves' term to capture the shortfall associated with a given configuration and set of demands. It was also used

in Section 3.2 to prove diminishing return property. Given a system with  $m$  panels and  $m$  physicians, for any flexible configuration  $A$ , the optimal value of shortfall is given by

$$V(A) = \max_M \left\{ \sum_{j \in M} D_j - \sum_{i \in P(M)} C_i \right\}$$

where  $M$  is any subset (including the null set) of the index set  $\{1, 2, \dots, m\}$ . For any given subset of panels  $M$ ,  $P(M)$  is the subset of physicians that can serve at least one of the panels in set  $M$ . Thus,  $j \in P(M)$  if and only if there is at least one panel  $j \in M$  such that  $(j, i) \in A$ . Each term within the maximization is the difference between the demand for some subset of panels and the maximum capacity available from the physicians which are connected to that subset of panels.

Using the result of [34], we can analytically calculate expected overflow for a given flexibility configuration. Here overflow is equivalent to the unfilled demand of a given configuration. Then expected satisfied demand can be calculated based on the results of expected overflow. Particularly, for a fully flexible configuration  $\mathcal{A}$ , the expected number of diversions can be calculated as the expected overflow of the corresponding dedicated configuration minus the expected overflow from this configuration  $\mathcal{A}$ .

We assume all the prescheduled and same-day demand follow Poisson distributions (this method could be easily applied to any other discrete probability distribution).  $\lambda_j^p$  and  $\lambda_j^s$  denote the poisson distribution rates for prescheduled and same-day demand in panel  $j$ . By adapting Jordan and Graves' terms, we analyze multiple flexibility configurations for small practices ( $m \leq 4$ ).

For illustrative purposes, here we consider the 2-physician case for dedicated with one additional provider system. Refer to the appendix A to see results based on other flexibility configurations we explored.

Given one demand realization  $[D_1^p, D_2^p, D_1^s, D_2^s]$ , the overflow for prescheduled demand is just the sum of the overflow for each panel,  $\max\{0, D_1^p - N_1^p\} + \max\{0, D_2^p - N_2^p\}$ , because physicians are not allowed to share prescheduled demand. Here we only

have two terms inside the maximization, then the expected value of overflow is easy to calculate. However, for same-day demand, the situation becomes more complex. The overflow under a particular demand realization will be

$$\max\{0, D_1^s - [C_1 - \min(N_1^p, D_1^p) + Y], D_2^s - [C_2 - \min(N_2^p, D_2^p) + Y], \\ D_1^s + D_2^s - [C_1 + C_2 - \min(N_1^p, D_1^p) - \min(N_2^p, D_2^p) + Y]\}$$

which is the maximum of four terms. It is not easy to calculate the expected value for this maximization directly. Based on the analysis of these four terms, we find each of them corresponds to a disjoint condition or case. For example, the cases correspond to the 2-physician practice for dedicated with one additional provider system is as following four cases:

- (i)  $D_1^s - [C_1 - \min(N_1^p, D_1^p) + Y]$  is max  $\Leftrightarrow D_1^s \geq C_1 - \min(N_1^p, D_1^p) + Y$  and  $D_2^s < C_2$
- (ii)  $D_2^s - [C_2 - \min(N_2^p, D_2^p) + Y]$  is max  $\Leftrightarrow D_2^s \geq C_2 - \min(N_2^p, D_2^p) + Y$  and  $D_1^s < C_1$
- (iii)  $D_1^s + D_2^s - [C_1 + C_2 - \min(N_1^p, D_1^p) - \min(N_2^p, D_2^p) + Y]$  is max  $\Leftrightarrow D_1^s + D_2^s \geq C_1 + C_2 - \min(N_1^p, D_1^p) - \min(N_2^p, D_2^p) + Y, D_1^s \geq C_1$  and  $D_2^s \geq C_2$
- (iv) 0 is max, otherwise

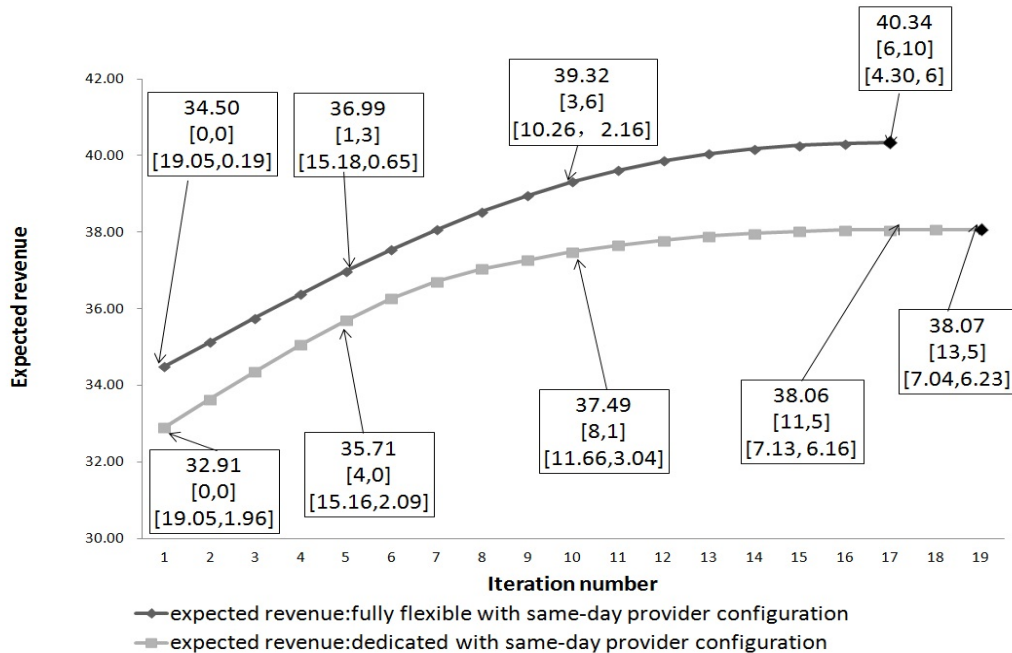
We can thus calculate the expected value of overflow based on the conditional probabilities under each of these conditions. For settings with more physicians, there will be more terms inside the maximization but not all of them need to be considered because some terms dominate others. However, in this dissertation, we only explore this method to calculate expected revenue for small practices such as  $m \leq 4$ . We didn't explore for larger practices because the number of terms involved in calculation exponentially increases.

### 3.4.4 Example: apply the greedy heuristic to the capacity allocation problem

To clarify how the heuristic works, we first consider a 2-physician practice where the P/S ratio for each physician is 8/16 and workload of the practice is 100%. If each physician has a capacity of 24, what should a nice solution of  $\mathbf{N}^P$  of each physician be? Since the physicians are identical with regard their workloads, our heuristic would start with  $\mathbf{N}^P$  values of  $[0, 0]$  and increase the  $\mathbf{N}^P$  value of each physician by 1, and calculate the total expected revenue for the clinic. Using the difference in expected revenues, the algorithm either stops or increments the appropriate physician's  $\mathbf{N}^P$  value by 1. Since the physicians are identical, there is no reason to favor one physician over another in the search. The trajectory or path taken by the algorithm would be fairly straightforward:  $[0,0]$ ,  $[1,0]$ ,  $[1,1]$ ,  $[2,1]$ ,  $[2,2]$  and so on until the optimal pair of values is reached. For this example, the optimal pair of  $\mathbf{N}^P$  values for the dedicated case is  $[9,9]$  and for the fully flexible case is  $[8,8]$ , assuming 0.75 and 0.9 as revenue for seeing one prescheduled and one same-day patient under 120% workload. In our observation,  $\mathbf{N}^P$  is not sensitive in a certain range beyond one point as the revenue function produces diminishing returns and when revenue function produces a positive but small value, the optimal  $\mathbf{N}^P$  does change while not significant impact on total revenue. Due to these observation, we establish a heuristic to analyze 4 physicians practices in order to reduce computational complexity. Given a arbitrary small  $\epsilon = 0.001$ , when the diminishing return produced by revenue function is equal or smaller than the  $\epsilon$ , stop searching for the locally optimal  $\mathbf{N}^P$ . This heuristic provides good enough  $N^p$  solution to achieve high revenue for the whole clinic.

If individual physician workloads are not identical (the asymmetric case, which we formally describe in the results section), then the path traversed by the algorithm will not be so straightforward. To illustrate, still consider a 2-physician practice in which P/S ratios for physician 1 is 6/12 and for physician 2 is 10/20. The entire practice

is under 120% workload. Then Physician 1 has a mean prescheduled demand of 7.2 and same-day demand of 14.4, while Physician 2 has a mean prescheduled demand of 12 and same-day demand of 24, see Figure 3.2. Each physician has a capacity of 24. If we define utilization as the ratio of expected total demand (prescheduled + same-day) and the capacity, then Physician 1 has a utilization of 90% and Physician 2 has a utilization of 150%. The overall utilization of the clinic or practice is  $(7.2 + 14.4 + 12 + 24)/(24 + 24) = 120\%$ . There are no additional same-day provider slots in this example.



**Figure 3.2.** Revenue of dedicated and fully-flexible systems as a function of the iteration number in the greedy heuristic for a 2-physician practice, with 120% workload and demand asymmetry.

The path of  $\mathbf{NP}$  values traversed by our heuristic for the fully-flexible and dedicated cases is shown in the Figure 3.2 . At selected points on this path, we provide the expected revenue; the  $\mathbf{NP}$  values for the two physicians (pair of values in the second parenthesis); and then the expected missed prescheduled and same-day demands (pair of values in the first parenthesis).

### 3.4.5 Computational efficiency of the greedy heuristic and the lower bound

In Section 3.4.2, we propose the greedy heuristic to be an efficient heuristic for this capacity allocation problem because in most practical scenarios, the greedy heuristic yields optimal solutions. In this section, we discuss the computational efficiency of the greedy heuristic and a lower bound we constructed.

**How to establish a lower bound** Under the case of dedicated prescheduled and dedicated same-day configuration, recall that

$$\Delta_k(N_1^p, \dots, N_m^p) = P[D_k^p > N_k^p] \times \{R^p - R^s \times P[D_k^s \geq C_k - N_k^p | D_k^p > N_k^p]\}$$

,  $(\mathbf{NP})_{\mathbf{DD}}^*$  is given by the first point to make  $\Delta_k(N_1^p, \dots, N_m^p) < 0$ . In fact  $P[D_k^s \geq C_k - N_k^p | D_k^p > N_k^p] = \frac{R^p}{R^s}$  can provide a good estimate of  $(\mathbf{NP})_{\mathbf{DD}}^*$ .

Similarly, under the case of fully flexible prescheduled and fully flexible same-day configuration, the difference in revenue can then be written as

$$\begin{aligned} \Delta_k(N_1^p, \dots, N_m^p) = & P\left[\sum_{j=1}^m D_j^p > \sum_{i=1}^m N_i^p\right] \times \{R^p \\ & - R^s \times P\left[\sum_{j=1}^m D_j^s \geq \sum_{i=1}^m C_i - \sum_{i=1}^m N_i^p \mid \sum_{i=1}^m D_i^p > \sum_{i=1}^m N_i^p > 0\right]\} \end{aligned}$$

By same logic, we know that  $P[\sum_{i=1}^m D_i^s \geq \sum_{i=1}^m C_i - \sum_{i=1}^m N_i^p \mid \sum_{i=1}^m D_i^p > \sum_{i=1}^m N_i^p > 0] = \frac{R^p}{R^s}$  can provide a good estimate of  $(\mathbf{NP})_{\mathbf{FF}}^*$  under symmetric cases.

Corollary 1 shows that  $\min\{(\mathbf{NP})_{\mathbf{DD}}^*, (\mathbf{NP})_{\mathbf{FF}}^*\}$  is always a lower bound for the optimal capacity allocation  $(\mathbf{NP})_{\mathcal{A}}^*$ , here  $\mathcal{A}$  belongs to the two sets we discuss in previous section. In fact, under these two extreme situations (dedicated for both prescheduled and same-day patients, fully flexible for both prescheduled and same-day patients), our model is similar to a newsvendor model. If  $R^p \geq R^s$  holds, we need to increase  $\mathbf{NP}$  until reaching the threshold  $\frac{R^p}{R^s}$ .

Table 3.2 summarizes the running time of Formulation I based on sample average approximation, greedy heuristic, and greedy heuristic with the provided lower bound.

The tested scenarios are all symmetric cases (each physician has identical P/S ratio and workload, refer to section 1.1 for more details about symmetric cases). We tested three different  $P/S$  ratios (8/16, 12/12 and 16/8) and each ratio is repeated under three different workloads (80%, 100%, and 120%). Formulation I is computed by using cplex 12.3. Greedy heuristic and greedy heuristic with lower bound are both computed by using Matlab 7.11.0. Running environment is [Intel(R) Core(TM) i5-2520M CPU@ 2.50GHz, RAM 6.00GB]. In this table, ‘SAA’ denotes ‘sample average approximation method’, ‘GA’ denotes ‘greedy algorithm’, and ‘GA.LB’ denotes ‘greedy algorithm with constructed lower bound’.

**Table 3.2.** Comparisons of running time under different configurations and demand scenarios(time:/s)

cases	Dedicated			Subgroup			Fully		
	SAA	GA	GA.LB	SAA	GA	GA.LB	SAA	GA	GA.LB
80% 8/16	337.4	0.2	0.3	325.8	3.2	2.8	971.4	545.2	300.9
100% 8/16	1640.5	0.2	0.2	4655.5	1.8	1.2	5951.1	346.1	136.0
120% 8/16	2946.9	0.1	0.1	42787.4	1.0	0.6	45441.4	165.2	22.2
80% 12/12	452.4	0.2	0.2	410.5	3.6	2.8	345.5	566.8	242.9
100% 12/12	1395.3	0.2	0.2	4982.2	2.7	1.8	6365.3	458.4	179.0
120% 12/12	2312.2	0.2	0.2	25190.8	1.7	1.0	45912.4	291.2	58.7
80% 16/8	292.2	0.3	0.3	352.2	4.0	3.5	211.4	579.8	166.9
100% 16/8	901.1	0.3	0.2	1672.2	3.6	2.2	1597.4	552.0	173.7
120% 16/8	866.4	0.2	0.2	14755.3	2.5	1.7	21576.9	384.9	59.8

From Table 3.2, we observe that, greedy heuristic sufficiently reduces the computational time for all tested cases. Furthermore, lower bound is advantage to cut searching space for complicated cases. Note that, the greedy heuristic is based on our analytical method to compute the expected revenue for one particular configuration, which only works for small cases ( $m \leq 4$ ) now. Because the number of terms involved in formula to compute revenue exponentially increases with number of physicians and the complexity of flexibility configurations.

# CHAPTER 4

## IMPACT OF FLEXIBILITY AND CAPACITY ALLOCATION IN PRIMARY CARE PRACTICES

### 4.1 Introduction

In this chapter, we apply the framework and models in Chapter 2, and study the impact of flexibility and capacity allocation problem in primary care practices. Results in this chapter address the following research questions:

(i) What is the impact of flexibility on average performance in a primary care practice?

(ii) Is the optimal threshold policy essential for clinics to consider in primary care practices? If yes, what is the benefit of this optimal threshold policy?

(iii) What is the impact of flexibility on the optimal booking limit? Are the computational results consistent with analytical results in Section 3.3?

### 4.2 Computational results

#### 4.2.1 Impact of flexibility in primary care practices

Although we focus on a primary care setting, our model is general and can be extended under a service setting with two demand classes: urgent demand vs. non-urgent demand. To understand the impact of different types of flexibility, we run computational experiments based on **Formulation I**, **Formulation II** and **Formulation IV**.

In all mathematical experiments, we have Poisson samples for both prescheduled demand and same-day demand. Each practice has 3 physicians and each physician

has total 24 available slots per workday, since a typical appointment takes about 20 minutes and a physician's workday may be up to 8 hours. Note that, most clinics will serve same-day patients with the physicians' overtime. In our computational study, we consider the unsatisfied part of demand to be patient overflow, which could be served with physicians' overtime or be refused.

Due to the findings from [12], we follow the assumption (also made in [10]) that prescheduled patients exhibit a no show rate of 25% while same-day patients have a significantly smaller no show rate of 10%. Therefore, we set revenue achieved from same-day as 0.9 and revenue from prescheduled as 0.75. Note that in the analytical part, we assume that there is no deduction cost for referrals for convenience. However, in order to test the impact of referrals on the total revenue, we assume there will be a deduction cost if the physician sees a patient not from her own panel. The deduction cost for same-day referrals is assumed to be 0.05 while that for prescheduled referrals is 0.15, as losing continuity for prescheduled patients will result a higher loss in efficiency.

For convenience, we introduce following definitions of three different types of flexibility:

**Baseline:** prescheduled patients dedicated, and same-day patients are dedicated.

**Type I flexibility:** prescheduled patients dedicated, and same-day patients are fully flexible.

**Type II flexibility:** prescheduled patients fully flexible, and same-day patients are dedicated.

**Type III flexibility:** Prescheduled patients are pooled, and same-day patients are fully flexible.

By using **Formulation I**, we test the performances of baseline and Type I flexibility. The difference in revenue between these two configurations provides the benefit of same-day flexibility (type I flexibility).

We test the performances of Type II flexibility by using **Formulation II**. We also test the performance of Type III flexibility by using **Formulation IV**. We quantify the benefit of introducing flexibility to serve prescheduled patient (Type II flexibility). We also compute the benefit of introducing additional flexibility (in the form of pooling prescheduled patients together to share a single booking limit) to serve prescheduled patient when same-day patients are fully flexible (Type III flexibility - Type I flexibility).

We run eight cases: each with different  $P/S$  ratios (recall Section 1.1,  $P/S$  ratio is  $\frac{\text{prescheduled demand rate}}{\text{sameday demand rate}}$ ) for physicians in the practice. The mean of prescheduled demand is  $P$  and the mean of same-day demand is  $S$ . Note that the system is perfectly balanced to have 100% workload. We multiply the demand by 0.8 and 1.2 to create over-worked and under-worked practices. Case 1 - case 4 are all symmetric cases: each physician has same  $P/S$  ratio, and we increase the prescheduled demand rates over the cases. Case 5 - case 6 are asymmetric cases with two groups of mixed  $P/S$  ratios (first physician has a  $P/S$  ratio less than one, second physician has a  $P/S$  ratio of one, and the last physician has a  $p/s$  ratio greater than one). Case 7 - case 8 are asymmetric cases with same  $P/S$  ratio (case 7 to be  $\frac{1}{2}$  and case 8 to be 2) but each physician is differently utilized. These cases are motivated based on our interactions with small primary care practices as well as larger academic practices. The results are summarized in Table 4.1.

**Table 4.1.** Impact of flexibility in primary care practices

80% utilization		Revenue[Improvement in %]				
Cases	baseline	Type I		Type II		Type III
sym 4/20	49.09	49.81	[1.45%]	49.11	[0.04%]	49.81 [1.45%]
sym 8/16	47.73	48.41	[1.44%]	47.80	[0.17%]	48.41 [1.44%]
sym 16/8	44.45	45.10	[1.45%]	44.69	[0.54%]	45.10 [1.46%]
sym 20/4	43.29	43.85	[1.31%]	43.63	[0.79%]	43.91 [1.45%]
asym 4/20,12/12,20/4	46.79	47.54	[1.62%]	46.99	[0.44%]	47.56 [1.65%]
asym 8/16,12/12,16/8	46.87	47.71	[1.80%]	47.09	[0.48%]	47.71 [1.80%]
asym 6/12,8/16,10/20	47.25	49.07	[3.85%]	47.98	[1.56%]	49.07 [3.85%]
asym 12/6,16/8,20/10	44.40	46.16	[3.95%]	45.42	[2.28%]	46.18 [4.01%]
100% utilization		Revenue[Improvement in %]				
Cases	baseline	Type I		Type II		Type III
sym 4/20	57.79	59.87	[3.60%]	57.92	[0.21%]	59.87 [3.60%]
sym 8/16	55.90	57.98	[3.72%]	56.21	[0.55%]	57.98 [3.72%]
sym 16/8	52.56	54.50	[3.68%]	53.29	[1.38%]	54.55 [3.79%]
sym 20/4	50.92	52.64	[3.37%]	51.92	[1.96%]	52.81 [3.71%]
asym 4/20,12/12,20/4	54.30	56.24	[3.58%]	54.84	[1.00%]	56.26 [3.61%]
asym 8/16,12/12,16/8	54.17	56.16	[3.67%]	54.67	[0.93%]	56.18 [3.71%]
asym 6/12,8/16,10/20	53.76	57.91	[7.73%]	55.67	[3.56%]	57.92 [7.74%]
asym 12/6,16/8,20/10	50.55	54.30	[7.42%]	52.76	[4.38%]	54.37 [7.56%]
120% utilization		Revenue[Improvement in %]				
Cases	baseline	Type I		Type II		Type III
sym 4/20	61.46	62.83	[2.22%]	61.56	[0.16%]	62.83 [2.23%]
sym 8/16	59.44	60.76	[2.22%]	59.53	[0.15%]	60.79 [2.27%]
sym 16/8	55.97	57.09	[2.00%]	56.13	[0.29%]	57.18 [2.16%]
sym 20/4	54.13	55.07	[1.75%]	54.49	[0.67%]	55.26 [2.10%]
asym 4/20,12/12,20/4	57.59	58.82	[2.15%]	57.74	[0.27%]	58.88 [2.24%]
asym 8/16,12/12,16/8	57.54	58.83	[2.24%]	57.64	[0.18%]	58.88 [2.33%]
asym 6/12,8/16,10/20	57.58	60.62	[5.28%]	58.80	[2.11%]	60.69 [5.40%]
asym 12/6,16/8,20/10	53.76	56.72	[5.52%]	55.41	[3.09%]	56.90 [5.85%]

Table 4.1 shows that, the highest benefit achieved from type I flexibility (only allowing flexibility to serve same-day patients in the system), is observed under 100% utilization. Under symmetric cases, this benefit is usually in the range of 3%-4%, even though the same-day average demand is very different. Note that, the benefit

of type I flexibility is not only achieved from same-day patients: case 3 (symmetric: 16/8) has less same-day patients than case 2 (symmetric: 8/16) but has a higher benefit due to type I flexibility. Similar benefit can be observed under asymmetric cases with mixed  $p/s$  ratio. In fact, the highest benefit from type I flexibility is obtained from the asymmetric cases with differently utilized physicians. Although case 7 (asymmetric: 6/12, 8/16, 10/20) and case 8 (asymmetric: 12/6, 16/8, 20/10) are different in the expected number of same-day patients, the benefit from type I flexibility for these two cases is larger than 7%. That is, this type I flexibility performs significantly better in the asymmetric cases than in the symmetric cases, i.e. when some physicians have higher demand in relation to others. In this case, flexibility is not only used to hedge against the variability in arriving same-day patient demands, but also to balance expected demand and available supply of each of the physicians. In the flexible system, the busier physician reserves more slots to satisfy prescheduled patient demands, while the lower utilized physician picks up the extra same-day appointment burden. Thus while flexibility implies a loss of continuity for same-day patients (who need it less anyway), it improves a physician's ability to provide more prescheduled appointments. These additional appointments can then be used for non-urgent but important follow-ups for patients with chronic conditions who have a in greater need for continuity.

Type II flexibility (only allowing flexibility to serve prescheduled patients in the system), also generates most benefit under 100% workload. However, different from type I flexibility, the benefit achieved from type II flexibility is closely connected with prescheduled demand rate. The more prescheduled patients, the greater the benefit from type II flexibility. Under symmetric cases, benefit from type II flexibility ranges from 0.20%-2%. No significant benefit can be observed under asymmetric cases with mixed  $p/s$  ratio. The highest benefit from type II flexibility is still obtained for the asymmetric cases with differently utilized physicians. The benefit from type II

flexibility under case 7 (asymmetric: 6/12, 8/16, 10/20) is 3.56% and the benefit from type II flexibility under case 8 (asymmetric: 12/6, 16/8, 20/10) is 4.38%, as case 8 has more prescheduled patients than case 7. Both of these two values are significantly larger than the benefit from type II flexibility under symmetric cases - again due to the ability to pool demand and accommodate unbalanced demand with available capacity.

Suppose a busy primary care practice is inherently flexible to serve same-day patients, let us consider following question: is it beneficial to introduce additional flexibility to serve prescheduled patients? Intuitively, there must be some gain in revenue as flexibility can pool demand to reduce demand variance and balance demand with capacity. However, in all 8 cases, our observation is quite surprising: when same-day patients are fully flexible, the introduced additional flexibility to serve prescheduled patients (type III flexibility - type I flexibility) is usually marginal (0.01%-0.40%). Why is the impact of this additional prescheduled flexibility small? In fact, same-day flexibility can balance demand well enough to get a higher revenue. With same-day flexibility, the additional prescheduled flexibility only generates benefits under the case that demand for prescheduled is over 24 for one physician and there are unused prescheduled slots still available to be flexibly used by the system. As this benefit only occurs in very extreme cases, the corresponding probability is small, resulting in small benefit. Based on this observation, suppose a practice already allows flexibility to serve urgent demand, the decision to introduce additional flexibility should be carefully considered based on necessity, because this flexibility results in a very marginal improvement in revenue. Another surprising observation is that, unlike type I and type II flexibility, the highest benefit of additional prescheduled flexibility is not obtained under 100% workload but observed under 120% workload. In other words, as we explained above, the benefit is higher when the system is busier.

Note that, **Type III flexibility** tested in Table 4.1 is not based on a real prescheduled fully flexible and same-day fully flexible configuration but based on its transformation, in which prescheduled patients from different panels share a common booking limit while prescheduled patients are always served by their own physicians as long as the actual demand does not exceed the corresponding physician’s total capacity. Note that, the sharing common booking limit structure in this transformation (no fixed individual booking limit) could work similar as a real prescheduled flexibility configuration (individual booking limit is fixed), see more details in Section 2.5. In fact, this special transformation is particularly beneficial in primary care practices because it could help to maintain the continuity for prescheduled patients (for whom continuity is much more critical) while improving access of patients as a regular full flexibility configuration. Then, question comes, what is the difference between these two configurations?

To answer this question, we use Formulation II to test a prescheduled and same-day both fully flexible configuration and use Formulation IV to test the transformation (prescheduled pooled and same-day fully flexible configuration). For each configuration, we still run previous described scenarios (8 different P/S ratios combined with 3 different workloads). All the comparisons are presented in Table 4.2.

**Table 4.2.** Prescheduled patients are fully flexibly shared vs. prescheduled patients are pooled while same-day patients are always fully flexibly shared

Cases	P-Pooled S-Full			P-Full S-Full		
	80%	100%	120%	80%	100%	120%
sym 4/20	49.81	59.86	62.80	49.81	59.87	62.85
sym 8/16	48.67	58.02	60.69	48.67	58.04	60.80
sym 16/8	45.41	54.46	57.03	45.41	54.57	57.19
sym 20/4	44.00	52.63	55.11	44.01	52.81	55.26
asym 4/20,12/12,20/4	47.67	56.27	58.75	47.67	56.33	58.88
asym 8/16,12/12,16/8	47.72	56.12	58.74	47.72	56.18	58.89
asym 6/12,8/16,10/20	49.14	57.94	60.53	49.14	57.96	60.69
asym 12/6,16/8,20/10	46.14	54.25	56.69	46.14	54.37	56.92

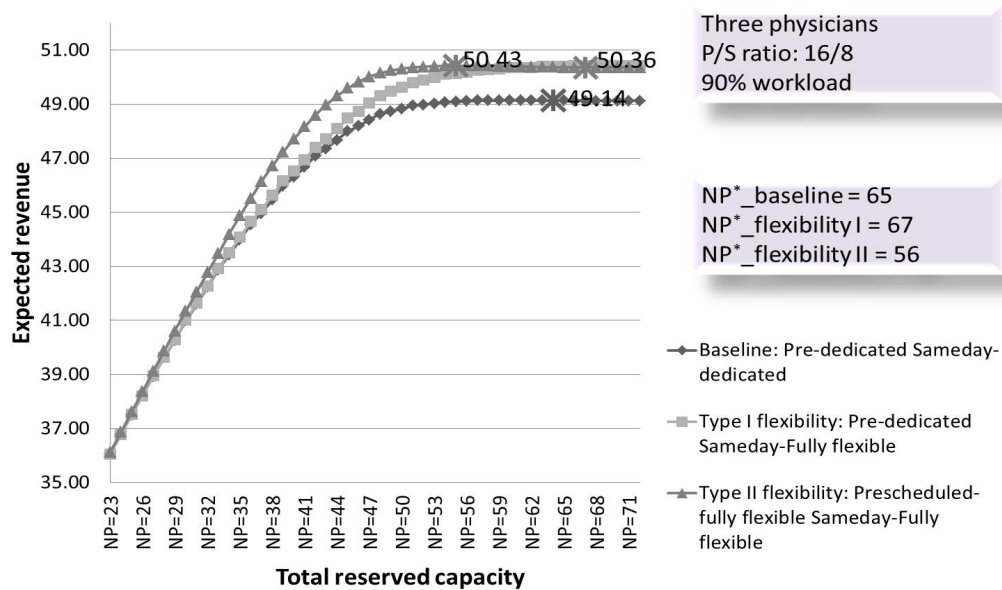
Consistently with the analysis in Section 2.5, the prescheduled pooled framework works better than a real prescheduled fully flexible configuration as long as same-day patients are fully flexibly shared. However, the difference between these two configurations is always marginal (0.0% – 0.3%). In addition, under some particular scenarios (not all the scenarios), as long as same-day patients are fully flexibly shared, the prescheduled fully flexible configuration work slightly worse than the prescheduled dedicated configuration while the prescheduled pooled model works slightly better than the prescheduled dedicated configuration. This is reasonable due to the diversion cost and the natural arrival sequence of two demand streams.

Generally, when the inherent physician flexibility is used to serve prescheduled patients as well as same-day patients, continuity in care for the chronic patients suffers while minimal additional benefits in access are observed. Furthermore, the improvement obtained from the flexibility to serve prescheduled demands is not significant in increasing access even when same-day flexibility is not viable, in applications where the same-day demand has greater need for continuity than the prescheduled demand. This is the case, for example, of a maintenance and repair service for a great variety of industrial or residential equipment (e.g. furnaces), where prescheduled demand is for standard maintenance operations, which any technician could effectively complete, while same-day demand will require deeper knowledge of the equipment, spare part availability, and quick resolution, and thus greatly benefit from the continuity provided by a technician that is an expert on that particular piece of equipment. In this case, it is not so much the client-server relationship that matters, but the match between the particular expertise of the technician and the needs of the client.

#### **4.2.2 Evaluation of booking policies**

In our model, we suggest a booking limit for clinics to reserve capacity for prescheduled patients; however, in most primary care practices, there may not be any such

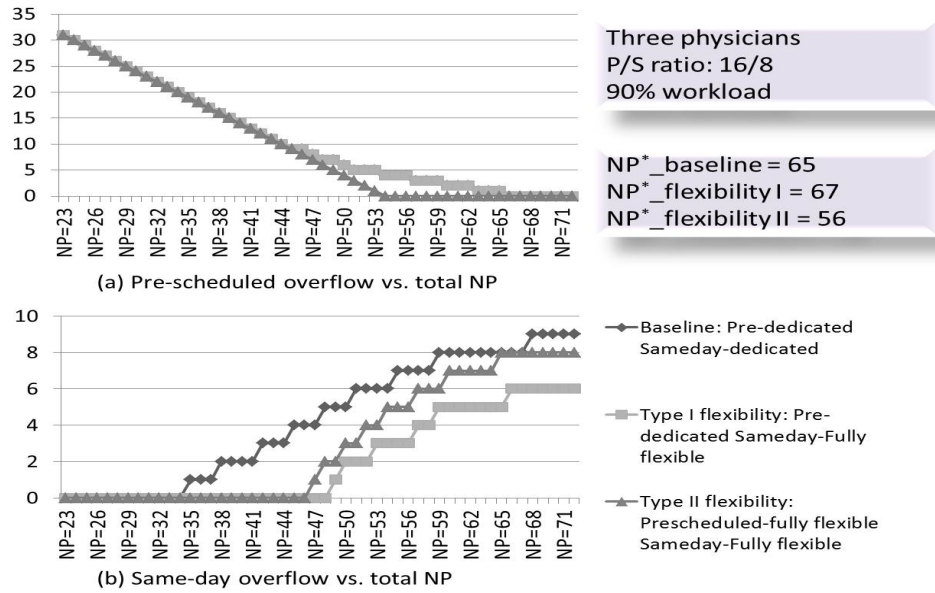
booking limit. We find that, when the workload of the various physicians in the practice is well balanced or less than 100%, the expected revenue of the practice is surprisingly insensitive to the booking limit. This is true as long as the booking limit is sufficiently high, suggesting that most practices could function appropriately without a booking limit, that is, simply accepting all prescheduled patient requests. Moving beyond the average performance metrics, however, we find that not setting a booking limit would result in a sizeable proportion of days where significant lack of access to same-day patients, or alternatively physician overtime to serve them, occurs.



**Figure 4.1.** Sensitivity analysis: expected revenue vs. booking limit in the entire system.

For example, Figure 4.1 shows how the expected revenue changes along with booking limit in the entire system, under a 90% workload case with more prescheduled patients, which is a typical workload in practice. Three different flexibility configurations are tested: baseline, type I flexibility and type III flexibility. All the optimal solutions are marked with \* in the figure. The figure shows that the expected revenue is not sensitive to the total booking limit beyond some point. With respect to the ex-

pected revenue, a booking limit does not seem to be necessary; however, inappropriate booking limit may result risk of having large same-day patient overflow.



**Figure 4.2.** Sensitivity analysis: patient overflow vs. booking limit in the entire system.

Still based on same samples from Figure 4.1, Figure 4.2 shows how the 95% percentile of overflow changes along with the total booking limit in the entire system. Note that in our model, we assume that prescheduled overflow is scheduled on another workday or refused while same-day overflow is served with overtime or refused. Figure 4.2 shows that an inappropriate booking limit could result risk(5%) of having large overflow (may be up to 9), resulting long overtime to accommodate the same-day patient overflow. A suitable booking limit should be considered to balance prescheduled patient requests and same-day patient requests.

We compare the performances of different booking policies under typical workloads (80%, 90%, 100%, 110%, 120%) and calculate the 75%, 85% and 95% percentile of overflow. Selected results (85%) are summarized in Table 4.3. ‘Baseline\*’ denotes the optimal booking limit under no flexibility configuration. ‘typeI\*’ denotes the optimal booking limit under type I flexibility configuration(prescheduled patients are

dedicated, and same-day patients are fully flexible). ‘typeIII\*’ denotes the optimal booking limit under type III flexibility configuration (prescheduled patients are pooled and same-day patients are fully flexible). ‘No limit dedicated’ denotes the booking limit is equal to the total capacity under dedicated system and ‘No limit flexible’ denotes the booking limit is equal to the total capacity under fully flexible system. ‘Pre-over’ denotes the prescheduled overflow and ‘Same-over’ denotes the same-day overflow.

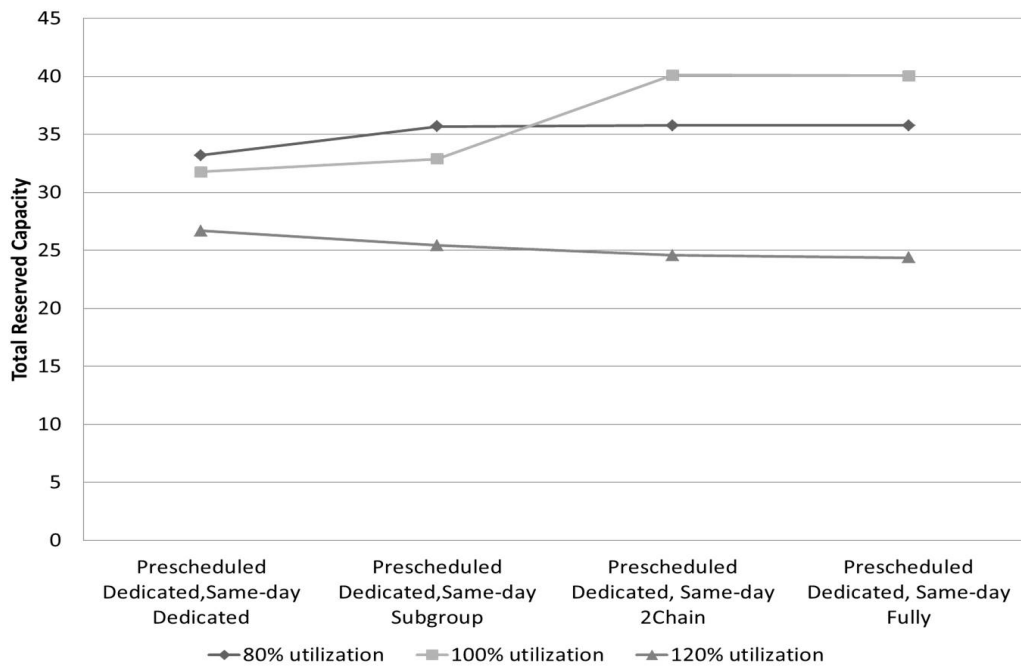
**Table 4.3.** Performances of different booking policies: expected revenue and 85% percentile of patient overflow

		Booking Policy	Revenue	Pre-over	Same-over
80% workload		Baseline*	49.95	0	2
		Type I*	51.60	0	0
		Type III*	51.60	0	0
		No limit dedicated	49.95	0	2
		No limit flexible	51.60	0	0
90% workload		Baseline*	52.65	4	8
		Type I*	57.15	0	3
		Type III*	57.05	0	3
		No limit dedicated	52.65	0	11
		No limit flexible	57.15	0	3
100% workload		Baseline*	56.25	3	7
		Type I*	57.7	0	7
		Type III*	57.7	2	5
		No limit dedicated	56.25	0	10
		No limit flexible	57.7	0	7
110% workload		Baseline*	57	9	11
		Type I*	57.9	10	9
		Type III*	57.65	10	9
		No limit dedicated	56.25	1	17
		No limit flexible	57.6	1	16
120% workload		Baseline*	57.45	15	14
		Type I*	58.15	17	11
		Type III*	57.9	18	11
		No limit dedicated	56.85	1	22
		No limit flexible	57.4	1	22

Although the corresponding expected revenues are similar to these different policies, the performances of patient overflow vary a lot. Meanwhile, policy ‘baseline’, policy ‘no limit dedicated’ and policy ‘no limit flexible’ tend to be more beneficial for prescheduled patients to get access, while same-day patient overflow could be very high. Policy ‘type I\*’ and Policy ‘type III\*’ have better performance to balance prescheduled demand and same-day demand. Decisions should be made based on clinics’ needs to reserve appropriate amount of capacity for prescheduled patients.

### 4.2.3 $N^p$ changes as a function of prescheduled flexibility or same-day flexibility

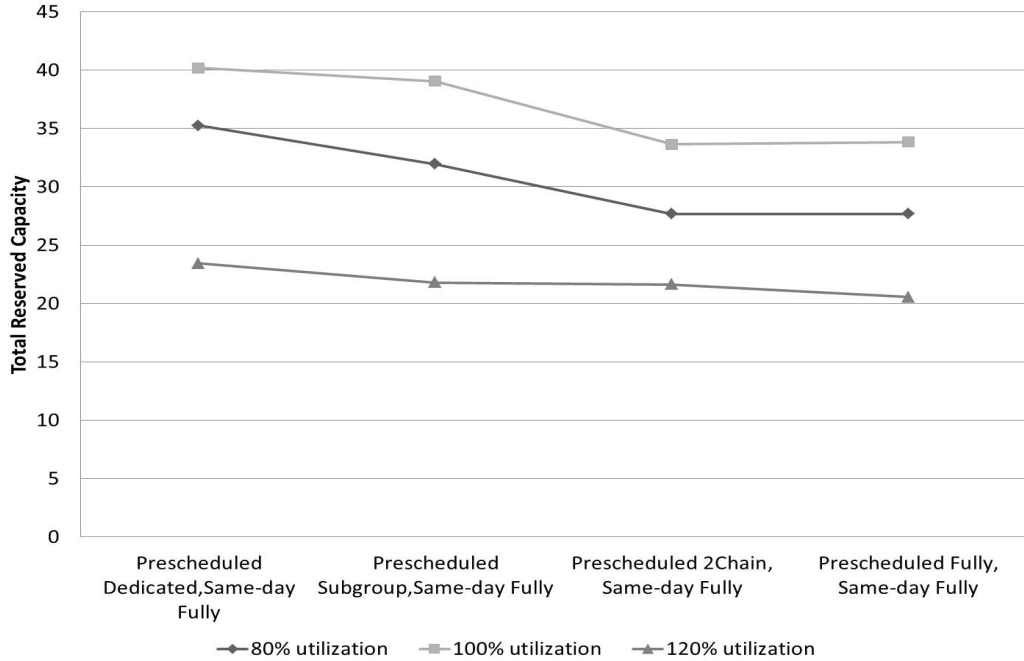
In Section 3.3, we shows how the optimal booking limit  $N^p$  changes as a function of prescheduled flexibility or same-day flexibility. In this section, we summarize results from computational study to confirm those analytical results.



**Figure 4.3.**  $N^p$  changes as a function of same-day flexibility when prescheduled patients are dedicated: under asym 6/12 8/16 10/20

Figure 4.3 shows how the solution of  $N^p$  changes as functions of same-day flexibility when prescheduled patients are dedicated. The results in Figure 4.3 are obtained by using Formulation I and four different flexibility configurations (dedicated, subgroup, 2chain, and full flexibility) are tested. This figure presents an asymmetric case (P/S ratios are 6/12, 8/16, 10/20). Which is same as presented in [10] and our proof in Section 3.3, when prescheduled patients are dedicated, the optimal  $N^p$  is a function of not only same-day flexibility, but also the system utilization. When the system is low-utilized, the optimal  $N^p$  will increase in same-day flexibility; on the other hand, when the system is high-utilized, the optimal  $N^p$  will decrease in same-day flexibility. Not surprisingly, higher flexibility results a higher revenue. When the utilization is 100%, the impact of same-day flexibility on revenue is highest.

From above observations, we find that the practice tends to set aside fewer slots for prescheduled appointments when the system is over-utilized, and reserve more when the system is under-utilized, if full flexibility to see same-day patients versus no flexibility is used. The reason is that when the system is under-utilized or even at 100% utilization, the total capacity in the system is generally sufficient to address the total demand. The full flexibility configuration tends to plan for more prescheduled appointments than the dedicated case since the same-day appointments can be flexibly shared and easily absorbed in the former but not in the latter. As the utilization of the system increases, there is a greater availability of same-day appointments which in our model produce greater revenue than prescheduled appointments. This prompts both the fully-flexible and dedicated configurations to decrease the number of prescheduled appointments so that more same-day appointments can be seen. The flexible configuration ends up offering even fewer prescheduled appointments than its dedicated counterpart, thus reserving more capacity for open access, since there is a higher probability of fully using the additional capacity when it is shared across all same-day appointments in the practice.



**Figure 4.4.**  $N^p$  changes as a function of prescheduled flexibility when same-day patients are fully flexible: under sym 8/16

By using Formulation I (Section 2.3) and Formulation II (Section 2.4), we also discuss how the solution of  $N^p$  changes as functions of prescheduled flexibility when same-day patients are fully flexible under a symmetric case (8/16). Figure 4.4 shows that the total  $N^p$  is always decreasing in prescheduled flexibility when same-day patients are fully flexible. Recall our observations in Section 4.2.1 that benefit from the additional prescheduled flexibility is always marginal, because the same-day flexibility is enough to adjust the demand uncertainty. However, though no significant benefit can be achieved in revenue, the prescheduled flexibility has positive impact on balancing two types of demand in the system. Because if there is no prescheduled flexibility in the system, the optimal  $N^p$  is always larger to achieve a better revenue; however, there would be a risk to have a large amount of same-day overflow due to demand variance over different workdays. Although a higher continuity could be provided to prescheduled patients, two demand streams (prescheduled patients and same-day patients) might be unbalanced. Meanwhile, if there is prescheduled flexibility in the

system, the optimal  $N^p$  is smaller, which can stably balance prescheduled demand and same-day demand. Decisions should be made based on clinics' particular needs.

#### 4.2.4 Negative correlation study

In Section 3.2.4, we always assume that demands are independent to prove diminishing return property for two special groups of flexibility configurations. In this section, however, we establish a small experiment to roughly study the impact of negative correlation on the performance of diminishing returns property. To keep the trials typical and not intricate, we focus on the computational study on single physician with two negatively correlated demand streams, which are realized successively. The total capacity of this physician is set to be 8 slots per day. We assume that the prescheduled and same-day demand both follow normal distribution and these two demand streams are negatively correlated. In another word, we have  $D^p \sim N(\mu^p, \sigma^p)$   $D^s \sim N(\mu^s, \sigma^s)$ , and  $\sigma_{D^p, D^s} < 0$  simultaneously.

In our study,  $\mu^p = \mu^s = 4$ , and  $\sigma^p = \sigma^s$  is any value from  $[2, 4, 6, 8]$ . Based on the settings, covariance of  $\sigma_{D^p, D^s}$  is calculated with given correlation  $\{-10\%, -20\%, \dots, -90\%, -100\%\}$ . We run the simulation for 100 times, and in each scenario, 1000 demand realizations are generated to estimate the corresponding returns when  $N^p$  is increased by one unit. The results under the high variance case are summarized in table 4.4. The simulation results show that, which is quite surprising, even under the worst case (high variance with  $-100\%$  correlation), the diminishing return property still holds consistently based on the stochastic settings.

**Table 4.4.** Does diminishing return property holds when two types of demand are negatively correlated?

$N^p$	correlation				
	-20%	-40%	-60%	-80%	-100%
$N^p=0$	0.578	0.588	0.598	0.613	0.631
$N^p=1$	0.472	0.488	0.506	0.528	0.554
$N^p=2$	0.345	0.367	0.393	0.423	0.459
$N^p=3$	0.220	0.249	0.274	0.312	0.354
$N^p=4$	0.114	0.144	0.176	0.210	0.249
$N^p=5$	0.045	0.068	0.095	0.126	0.161
$N^p=6$	0.007	0.025	0.042	0.068	0.093
$N^p=7$	-0.004	0.005	0.016	0.029	0.049
$N^p=8$	-0.007	-0.002	0.004	0.012	0.022
$N^p=9$	-0.004	-0.003	-0.0004	0.004	0.009
$N^p=10$	-0.002	-0.002	-0.000948	0.001	0.003
$N^p=11$	-0.001	-0.001	-0.0005	-5.4E-05	0.001
$N^p=12$	-0.0004	-0.0003	-0.0003	-4.5E-06	0.0003
...	...	...	...	...	...

Consistent with **Example 3.2.4**, we do observe some scenarios can produce the increasing returns when the absolute value of correlations are larger than 80%. However, the number of this type of scenarios is not high thus the diminishing return property still perfectly hold from the stochastic view.

### 4.3 Summary and conclusions

For multi-physician primary care practices, we apply the 3-stage stochastic model established in Chapter 2 to capture the scheduling of appointments for two subsequently realized demands, prescheduled and same-day.

Computationally, we investigate the value of different flexibility configurations and find that when practices flexibly serve same-day patients, any additional prescheduled flexibility has very marginal value. Restricting flexibility to serve prescheduled patients provides necessary continuity for prescheduled patients while still maintaining revenue. Furthermore, prescheduled flexibility is not effective in improving access

even when same-day patients are not flexibly shared. This could be of interest in service settings where continuity is more relevant in servicing same-day calls than in routine prescheduled maintenance calls.

When the inherent physician flexibility is used to serve prescheduled patients as well as same-day patients, continuity in care for the prescheduled patients with chronic conditions suffers while minimal additional benefits in access are observed. Furthermore, the flexibility to serve prescheduled demands is ineffective in increasing access even when same-day flexibility is not viable, in applications where the same-day demand has greater need for continuity than the prescheduled demand. This is the case, for example, of a maintenance and repair service for a great variety of industrial or residential equipment (e.g. furnaces), where prescheduled demand is for standard maintenance operations, which any technician could effectively complete, while same-day demand will require deeper knowledge of the equipment, spare part availability, and quick resolution, and thus greatly benefit from the continuity provided by a technician that is an expert on that particular piece of equipment. In this case, it is not so much the client-server relationship that matters, but the match between the particular expertise of the technician and the needs of the client.

We find that the practice tends to set aside fewer slots for prescheduled appointments when the system is over-utilized, and reserve more when the system is under-utilized, if full flexibility to see same-day patients versus no flexibility is used. The reason is that when the system is under-utilized or even at 100% utilization, the total capacity in the system is generally sufficient to address the total demand. The full flexibility configuration tends to plan for more prescheduled appointments than the dedicated case since the same-day appointments can be flexibly shared and easily absorbed in the former but not in the latter. As the utilization of the system increases, there is a greater availability of same-day appointments which in our model produce greater revenue than prescheduled appointments. This prompts both the fully-flexible

and dedicated configurations to decrease the number of prescheduled appointments so that more same-day appointments can be seen. The flexible configuration ends up offering even fewer prescheduled appointments than its dedicated counterpart, thus reserving more capacity for open access, since there is a higher probability of fully using the additional capacity when it is shared across all same-day appointments in the practice.

When the workload of the various physicians in the practice is well balanced, the expected revenue of the practice is surprisingly insensitive to the booking limit. This is true as long as the booking limit is sufficiently high, suggesting that most practices could function appropriately without a booking limit, that is, simply accepting all prescheduled patient requests. Moving beyond the average performance metrics, however, we find that not setting a booking limit would result in a sizeable proportion of days where significant lack of access to same-day patients, or alternatively physician overtime to serve them, occurs.

Several future research directions are possible. First, our research focuses on the aggregate level of considering the capacity allocation problem for one single workday. The results help us understand the impact of flexibility in a multi-physicians practice; however, our model does not capture the reality of dynamic allocation decisions as calls come in over the course of the day in the practice; decisions are made under partial knowledge of future demand. Second, we discuss prescheduled patient overflow in the paper, and assume the patient overflow will be assigned to another workday if prescheduled, result in overtime if same-day or simply lost. The impact of flexibility in the dynamic setting is worthy of further study.

Finally, the root cause of inadequate patient access in primary care practices is the shortage of primary care providers. Due to this reality, some clinics employ primary care practitioners (registered nurses or other medical assistants) to improve access for same-day patients. This additional provider typically focuses on flexibly serving

same-day patients across the practice. In such settings, given adequate capacity of this additional, flexible provider, little is gained by allowing physicians to flexibly share either prescheduled or same-day patients.

## CHAPTER 5

### IMPACT OF FLEXIBILITY AND CAPACITY ALLOCATION IN PRIMARY CARE PRACTICES WITH ADDITIONAL PROVIDER

#### 5.1 Introduction

In this chapter, we apply the framework and models in Chapter 2 and study the impact of flexibility and capacity allocation problem in a primary care practices with extra capacity due to an additional provider. Given a heavily utilized practice and the need to accommodate day-to-day variability, we consider the addition of  $Y$  flexible slots, perhaps through a nurse practitioner or a part-time provider, to serve same-day patients that cannot be seen by their own physicians. Observe that we denote this additional same-day provider by  $Y$  as well. The model for this configuration is an extension of model I in Section 2.3 by introducing an amount of capacity  $Y$  from additional provider. We assume that additional provider can only serve same-day patients while prescheduled patients are still only served by their own physicians in this chapter.

The results presented in this chapter cover the following topics:

- (i) Expected performances under primary care practices with additional provider in the system, for a range of clinic types that occur in practice.
- (ii) Impact of extra capacity from additional provider on the performances, such as expected revenue, expected overflow for two types of demand and the overflow probability.
- (iii) Impact of the extra capacity due to the additional provider on the optimal booking limit.

## 5.2 Computational results

We use the greedy heuristic described in the Section 3.4 to determine  $N^p$  values for the physicians for a variety of experimental cases. Our outcome measures are: 1) expected revenue; 2) expected satisfied prescheduled demand; 3) expected satisfied same-day demand; 4) expected number of non-PCP diversions (loss of continuity for same-day patients); and 5) probability that the per-physician overtime needed to satisfy same-day requests exceeds an hour.

Similar as in Chapter 4, note that the expected revenue is a weighted function of the expected number of prescheduled and same-day patients seen by the practice. ‘Revenue’ in our model is a surrogate for timely access. Each prescheduled appointment satisfied generates a revenue of 0.75 while each same-day appointment generates a revenue of 0.9. These are based on [12], where the no-show rates for prescheduled patients is 25% and for same-day patients is 10%.

A typical appointment in primary care takes about 20 minutes and a physician’s workday may be up to 8 hours. Therefore, in our experimental setting, each physician has 24 appointment slots in a day. In practice, this amount varies from physician to physician and from practice to practice. Our model can easily adjust for different capacities. The additional same-day provider, a nurse practitioner or physician assistant, works from  $Y = 0$  to  $Y = 24$  slots a day since we just explained this.

We still assume that the prescheduled and same-day demands are independent of each other and are Poisson distributed. Refer to Section 1.1 for more details and discussions. In our experimental results, we consider practices with asymmetry in the actual utilizations of individual physicians as well as  $P/S$  ratios. These asymmetries reflect situations where some senior physicians have greater number of patients than other physicians in the practice, or may have more patients with chronic conditions, with the result that their total prescheduled demand is higher in relation to their same-day demand.

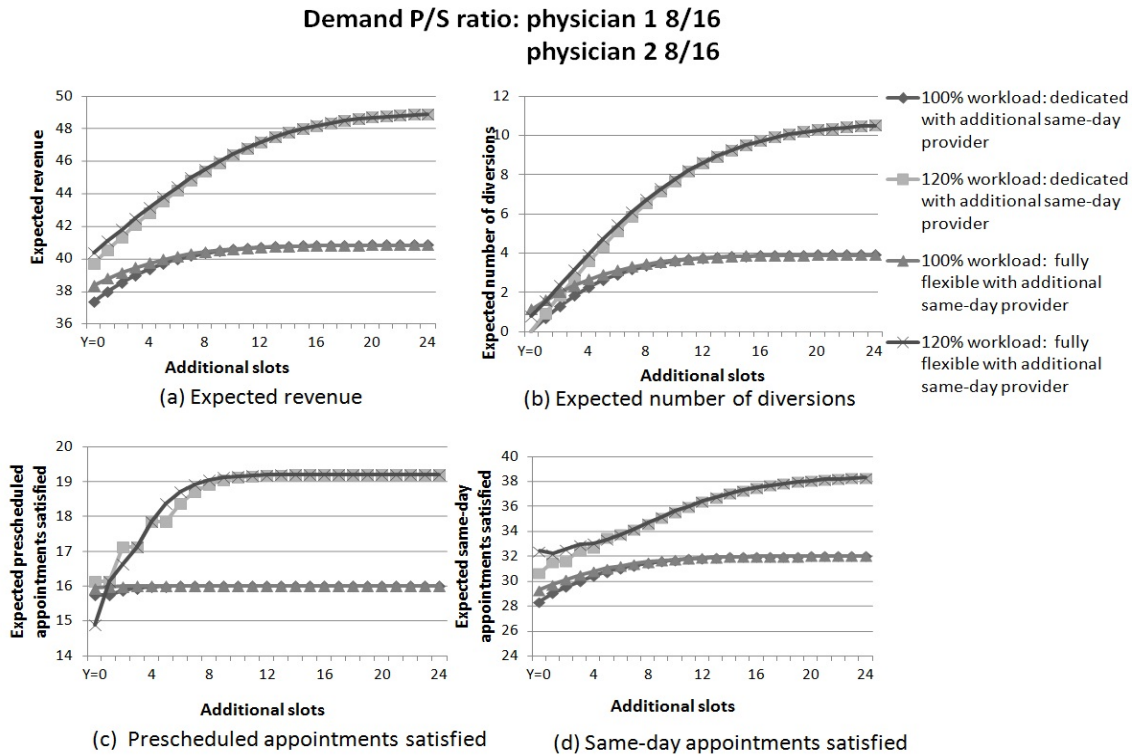
For asymmetric utilization in a 2-physician practice, we use  $P/S$  settings of 6/12 and 10/20 for Physician 1 and Physician 2, respectively. The total expected demand for Physician 1 is  $6 + 12 = 18$ , and for Physician 2 it is  $10 + 20 = 30$ . Since each physician works 24 slots a day, the utilizations of the two physicians are 75% and 125% respectively, while the overall clinic utilization remains 100%.

For asymmetry in  $P/S$  ratios, we consider a 2-physician practice with a  $P/S$  ratio of 8/16 for Physician 1 and 12/12 for Physician 2. Our experimental setup is summarized in the table below.

Physician capacity	24 slots per day
Number of physicians in practice	2,4
System Workload	100%,120%
Workload among physicians	Symmetric, Asymmetric
$P/S$ Ratios	8/16; 12/12; 6/12;/10/20
Y (Additional same-day provider slots)	0-24
'Revenue' of seeing one prescheduled patient	0.75
'Revenue' of seeing one same-day patient	0.9

**Table 5.1.** Numerical experiment setting

### 5.2.1 Expected performances: under symmetric case



**Figure 5.1.** Expected revenue, prescheduled appointments satisfied, same-day patients satisfied and diversions for fully flexible and dedicated configurations as a function of  $Y$

In Figure 5.1 (a), (b), (c), (d) show the expected revenue, expected prescheduled demand satisfied, expected same-day demand satisfied and the expected number of diversions for the 100% and 120% utilization cases. These results are for a 2-physician practice with a  $P/S$  setting of 8/16 for each physician: the physicians have identical workloads in this symmetric case. All performance measures are graphed as a function of  $Y$ . Recall that the additional same-day slots (denoted by  $Y$ ) represents a new provider (a physician assistant, a nurse practitioner or a newly hired physician) who will see same-day requests that the physicians in the practice are unable to satisfy.

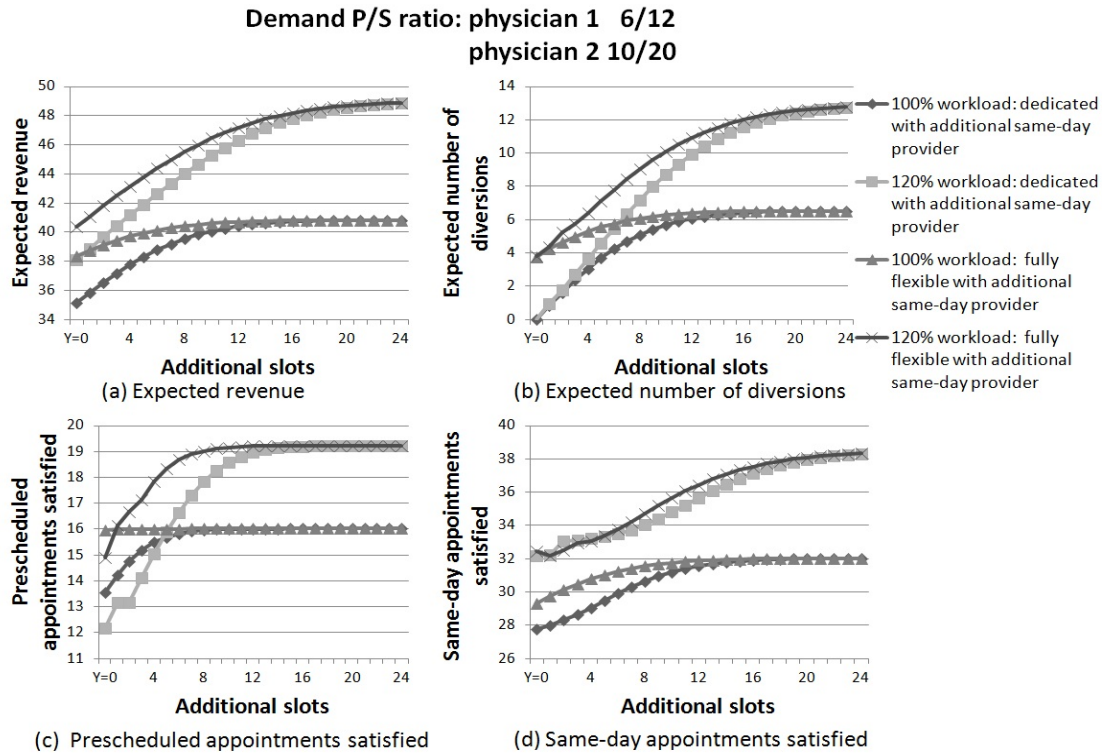
We see from Figure 5.1 (a) that the impact of flexibility is not particularly significant (around 2% improvement in revenue) when  $Y = 0$  for both the 100% and 120%

utilization cases. Also, as  $Y$  increases, the difference between the dedicated and the fully flexible systems decreases even further.

The expected prescheduled demand and expected same-day demand satisfied are also not that different under the two configurations (Figure 5.1 (b) and (c)). Although these two performance measures have non-smooth curves, they combine to give a smooth curve for the revenue function in Figure 5.1 (a). The non-smooth curves are because  $N^p$  values for the two physicians depend on both the flexibility configuration as well as the utilization of the system. The utilization of the system decreases with the increase in  $Y$ . This variation in optimal  $N^p$  values in turn impacts the number of prescheduled and same-day patients seen. For a detailed discussion on how utilization and the level of flexibility impact optimal  $N^p$  values, see [10].

Finally, we also notice that the expected number of diversions - which reflects losses in continuity for same-day patients - do not differ by much, although as anticipated, we see about one more diversion per day on average for the fully flexible case when compared to the dedicated case at  $Y = 0$ . This difference diminishes as  $Y$  increases but the total number of diversions in both configurations increases, especially in the 120% case, where the demand is higher than the available capacity.

### 5.2.2 Expected performances: under asymmetric Case



**Figure 5.2.** Expected revenue, prescheduled appointments satisfied, same-day patients satisfied and diversions for fully flexible and dedicated configurations as a function of  $Y$

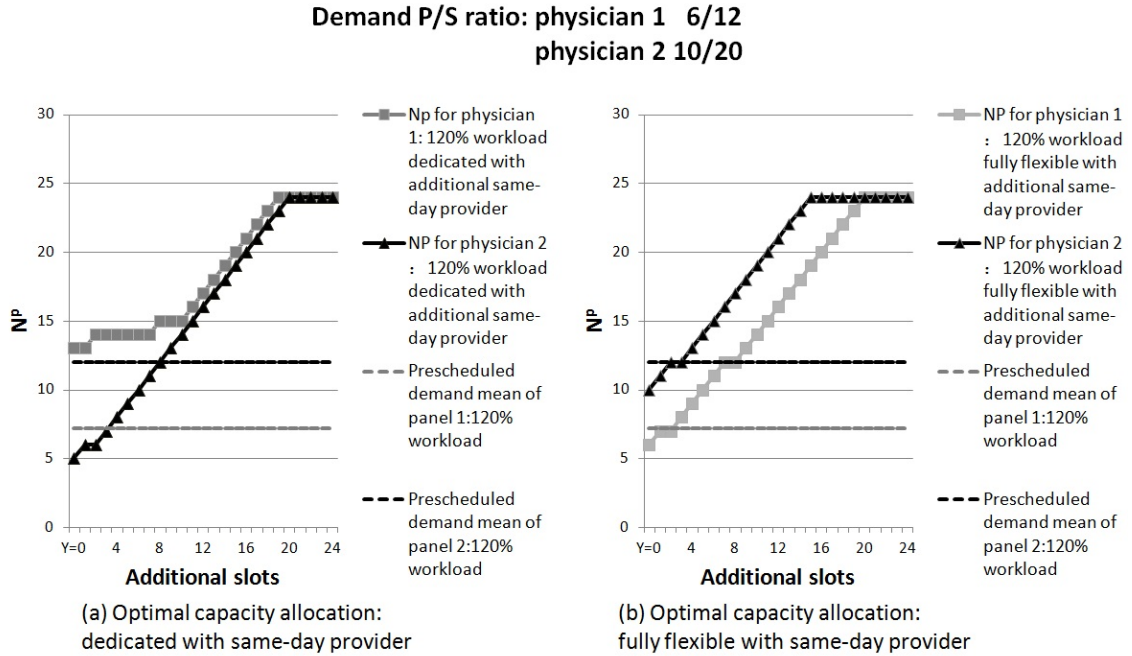
In the asymmetric case, one physician has a  $P/S$  setting of 6/12 while the other has as a  $P/S$  setting of 10/20. Here, the impact of flexibility is more significant, as shown in Figure 5.2. The optimal expected revenue is higher by nearly 5% in when  $Y = 0$  for both the 100% and 120% cases. Flexibility is useful in the asymmetric case not only to absorb the variability in demand but also to smooth the workload imbalance between the two physicians. As in the symmetric case, we note that the non-smooth curves for the satisfied prescheduled and same-day demand in Figure 5.2 (b) and (c) together combine to give us a smooth curve for the optimal expected revenue, shown in Figure 5.2 (a). Figure 5.2 (a) also shows that the additional same-day provider would have to work for  $Y = 3$  slots (about an hour) each day if the dedicated case is to match expected revenue of the fully-flexible case at  $Y = 0$ .

Next, we note that while flexibility applies only to same-day requests, it nevertheless allows a practice to see more prescheduled patients. As discussed before, prescheduled follow-ups for patients with chronic conditions require typically greater continuity than same-day requests for acute conditions. If each physician sees more prescheduled patients each day, then the wait for a prescheduled appointment will decrease. This means improved access for patients with non-urgent chronic conditions - a large percent of the United States population - for whom continuity is vital. In the  $Y = 0$  case under 120% utilization, the fully flexible 2-physician practice is able to see nearly 3 more prescheduled patients a day than the dedicated case.

Thus, for the many practices in the US that struggle to provide timely appointments to non-urgent requests, same-day flexibility will be beneficial, especially when the physicians have uneven workloads as often happens in practice. Figure 5.2 (c) shows that the fully flexible practice succeeds simultaneously in satisfying more same-day demand.

Finally, Figure 5.2 (d) gives the expected number of same-day diversions. When  $Y = 0$ , the fully flexible case uses 4 additional same-day diversions to non-PCP physicians on average to satisfy same-day demand. That is out of a total of 32 (38.4) same-day requests that the practice expects to see in the 100% (120%) utilization case, about 4 (4) of these same-day patients experience a loss in continuity on average. With the increase in  $Y$  the expected number of diversions increases in both configurations, but the difference between the two diminishes as the dedicated system begins to utilize the additional provider to satisfy same-day demand.

### 5.2.3 Impact of flexibility and additional provider on the optimal booking limit



**Figure 5.3.** 120% workload: comparison of optimal capacity allocation between dedicated with additional same-day provider and fully flexible with additional same-day provider system

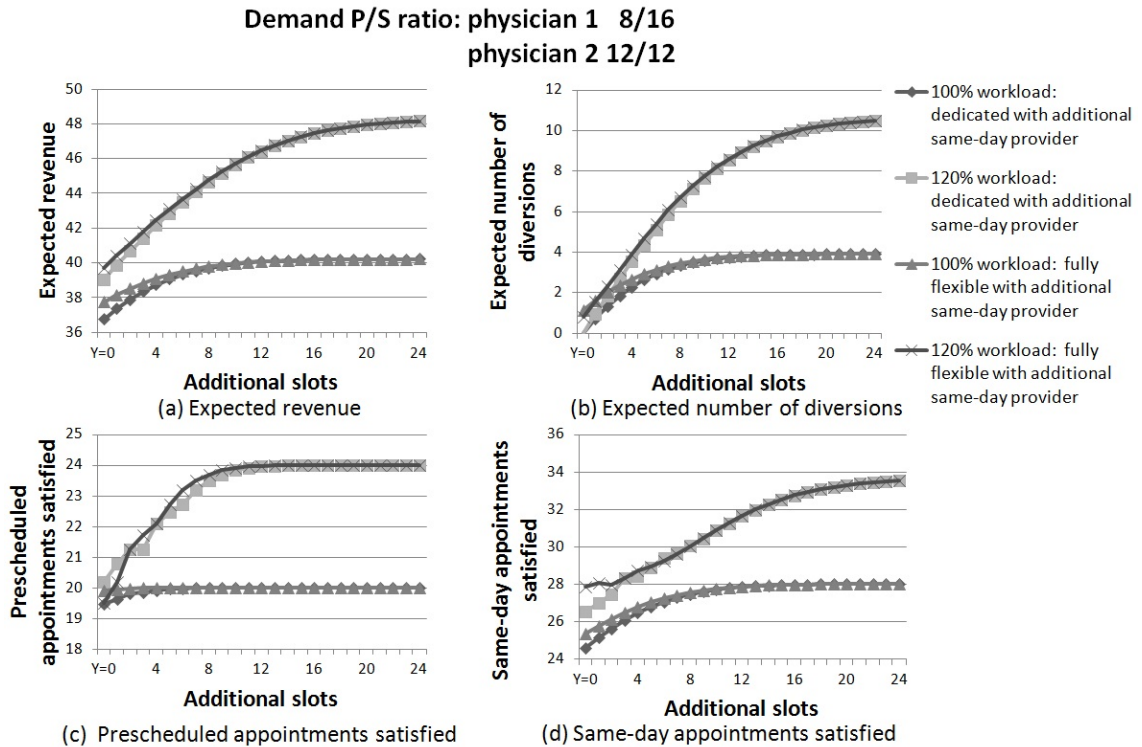
Why does flexibility produce a higher impact on prescheduled demand satisfied? To understand this, it is instructive to look at the optimal  $N^p$  values in the asymmetric case as a function of  $Y$  in both configurations. Figure 5.3 shows the optimal  $N^p$  values in the 120% utilization case. In this case, the average prescheduled and same-day demand for Physician 1 is  $[7.2, 14.4]$  (shown in Figure 5.3 as well) and Physician 2 is  $[12, 24]$  respectively. Each physician has a capacity of 24, so the total capacity of the clinic is 48. The total expected demand (prescheduled + same-day) is 57.6, which leads to an overall clinic utilization of 120%. While Physician 2 (150%) and the clinic as a whole (120%) are over-utilized, notice that Physician 1 is relatively under-utilized (total demand of 21.6 and utilization of 90%).

Consider the  $Y = 0$  case at the beginning point in Figure 5.3. We notice that the dedicated case paradoxically has a lower  $N^p$  value for physician 2 ( $N^p = 5$ ) compared to physician 1 ( $N^p = 13$ ), even though it is the second physician that has a higher prescheduled demand. This is because physician 2, whose utilization is high, has to make space in her schedule for higher revenue same-day appointments. As a result physician 2 compromises on the number of prescheduled patients she is able to see. In the long term, this means that the wait for a non-urgent appointment with this physician will be very high. This is in fact the prevailing situation in many primary care clinics in underserved areas, where the sheer proportion of urgent requests forces non-urgent appointments to be pushed months into the future. In contrast, physician 1, who is underutilized, can afford to have a high  $N^p$  value since any leftover prescheduled slots can be shifted on the day of the appointment to meet same-day demand.

In the fully flexible case same-day patients are shared between the two physicians, hence the optimal  $N^p$  values (6 and 10) are more in proportion with the physicians' respective prescheduled demand averages. Physician 2's same-day patients are seen by physician 1, and this flexibility allows Physician 2 to increase the number of prescheduled appointments she is able to see. While Physician 1's  $N^p$  value of 6 seems much lower than 13 in the dedicated case, this does not translate proportionally to fewer prescheduled patients seen, since physician 1's utilization is relatively low to begin with.

As  $Y$  increases, the dedicated case is able to address the imbalance between the physicians somewhat: notice that the  $N^p$  value for physician 2 increases quickly, while the  $N^p$  value for Physician 1 remains mostly steady. In the fully flexible case, the increase in  $Y$  is used by both physician 1 and physician 2 to increase their  $N^p$  values. This is because same-day flexibility had allowed them to set their  $N^p$  values in proportion to their respective prescheduled demands to begin with.

## 5.2.4 Expected performances: under asymmetric P/S values



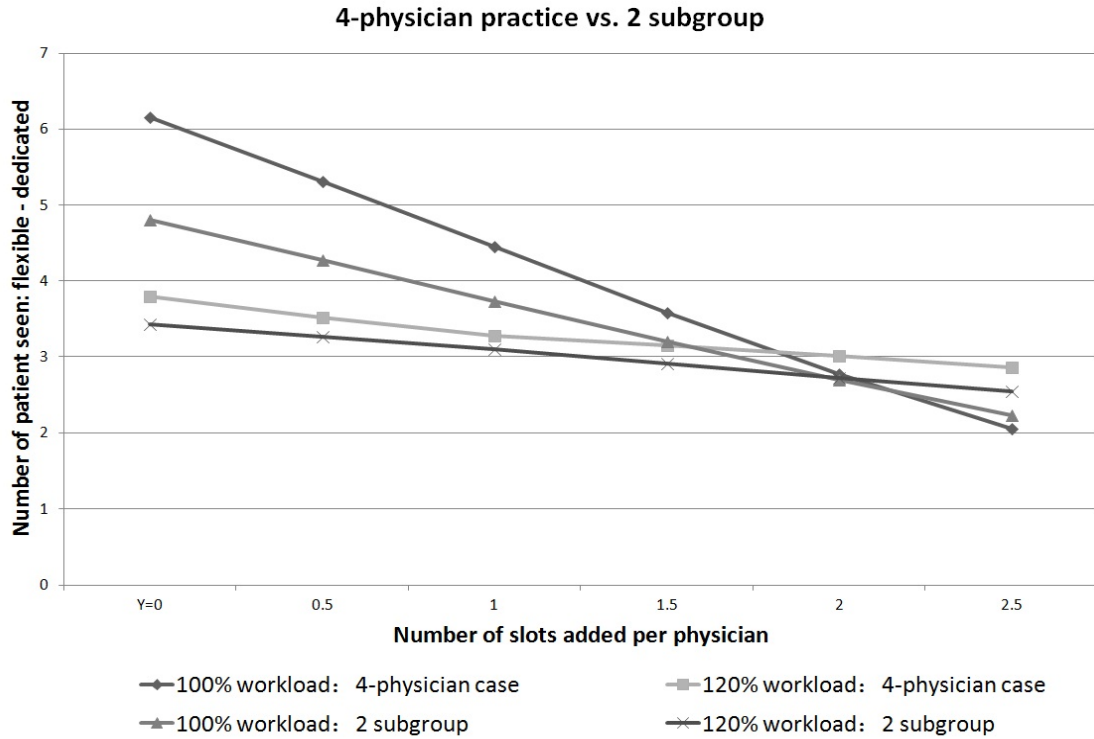
**Figure 5.4.** Expected revenue, prescheduled appointments satisfied, same-day patients satisfied and diversions for fully flexible and dedicated configurations as a function of  $Y$

So far we have seen  $P/S$  values that are symmetric across physicians. In other words, the ratio of prescheduled to same-day demand for each of the two physicians has been identical. In practice, physician panels differ in their case-mix. Some physicians have more patients with chronic conditions than others [8] and [47]. Therefore some physicians will have higher prescheduled or same-day demands than others. To reflect this reality, we test the fully flexible and dedicated cases when physician 1 has a  $P/S$  setting of 8/16 and physician 2 has a  $P/S$  setting of 12/12.

Figure 5.4 shows the expected revenue, expected prescheduled and same-day patients seen, and expected number of diversions under the two configurations at 120% and 100% utilization. Note that while the physicians differ in their  $P/S$  values, their utilizations are identical. The figure reveals that - as in symmetric case - flexibility

does not have a strong impact on expected revenue or on diversions. The expected prescheduled and same-day patients seen do differ but this is because of the differences in  $P/S$  values.

### 5.2.5 Expected performances: under asymmetric 4 physician practices



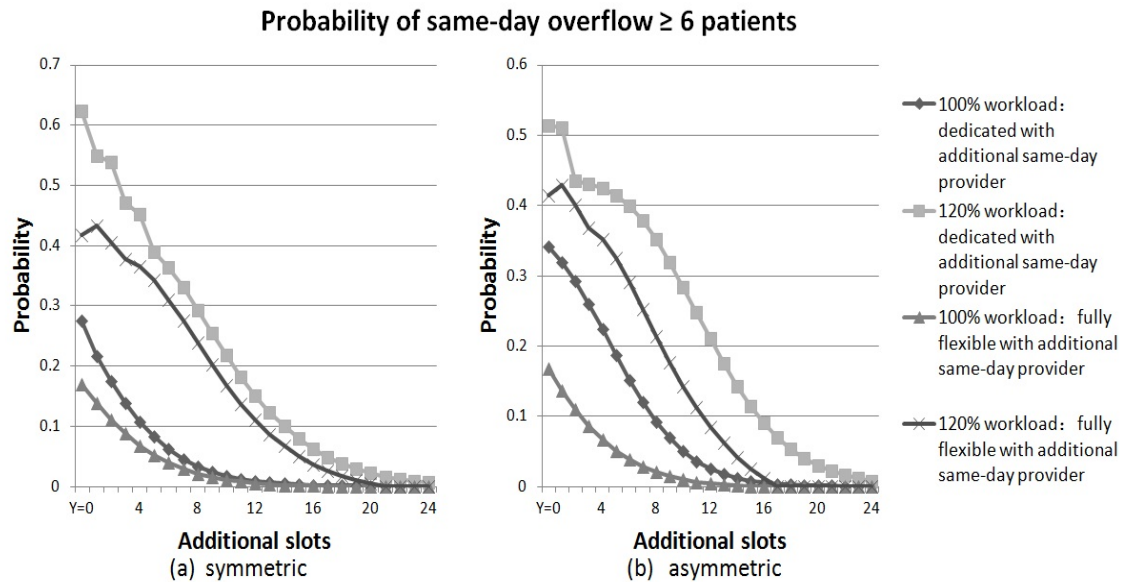
**Figure 5.5.** Comparison of impact of flexibility in 4 physicians’ practice and 2 physicians’ practice

We have seen already in the 2-physician case that flexibility has the highest impact when the physicians have imbalances in their total workloads and hence unequal utilizations. Now we study a 4-physician asymmetric practice with  $P/S$  settings of 6/12, 7/14, 9/18 and 10/20 (two of the physicians are under-utilized and two are over-utilized). We plot the difference in expected revenue between the fully flexible and dedicated cases as a function of  $Y$  for two different utilizations, 100% and 120%. This difference is plotted for both the 2 and 4 physician practices to evaluate the impact of practice size.

In Figure 5.5, we see that the impact of flexibility is higher at 100% utilization for both the 2 and 4 physician asymmetric practices. This is very much in line with the findings in the flexibility literature [34]. Furthermore, at  $Y = 0$ , the gains due to flexibility in the 4-physician practice are approximately twice the gains in the 2-physician practice. The difference in gains decreases as  $Y$  increases. In the 120% case, the benefit of flexibility is lower in both 2-physician and 4-physician practices. Because of the high utilization the difference between the 2 and 4 physician cases does not decline as much as in the 100% case with the increase in  $Y$ . This analysis confirms the impact of flexibility is highest at 100% utilization, for larger physician practices, and when the physicians have unequal workloads.

It is also necessary to point out that the number of diversions for same-day patients - which measures the loss in continuity for these patients - is also higher when full flexibility is used in a 4-physician practice. Diversions in a 4-physician practice, where a same-day patient can end up seeing up 3 other unfamiliar physicians (in addition to the additional same-day provider), are likely to have a greater impact on continuity of care compared to a 2-physician practice, where a same-day patient sees one of two providers.

### 5.2.6 Impact of extra capacity from additional provider on the probability of overtime



**Figure 5.6.** Impact of additional provider on the probability of overtime

We looked at the probability that the total missed same-day appointments exceeds 6 patients in a 2-physician practice. Recall that each unit of capacity in our model is a 20 minute slot. Therefore if the total missed same-day demand equals or exceeds 6 patients, it implies that the practice as a whole spends at least 2 hours of overtime (one hour per physician). Another way of thinking of this is that the overtime per physician will be an hour or more.

An important measure for practices is the probability of a given amount of overtime for the physicians in the practice. In many practices, same-day requests are not refused but squeezed into the schedule, often after regular working hours. This is true for the 3-provider family medicine practice we have collaborated with for this paper. If a request is triaged to be truly urgent, the practice makes sure that the patient is seen (though not always by the patient's own physician).

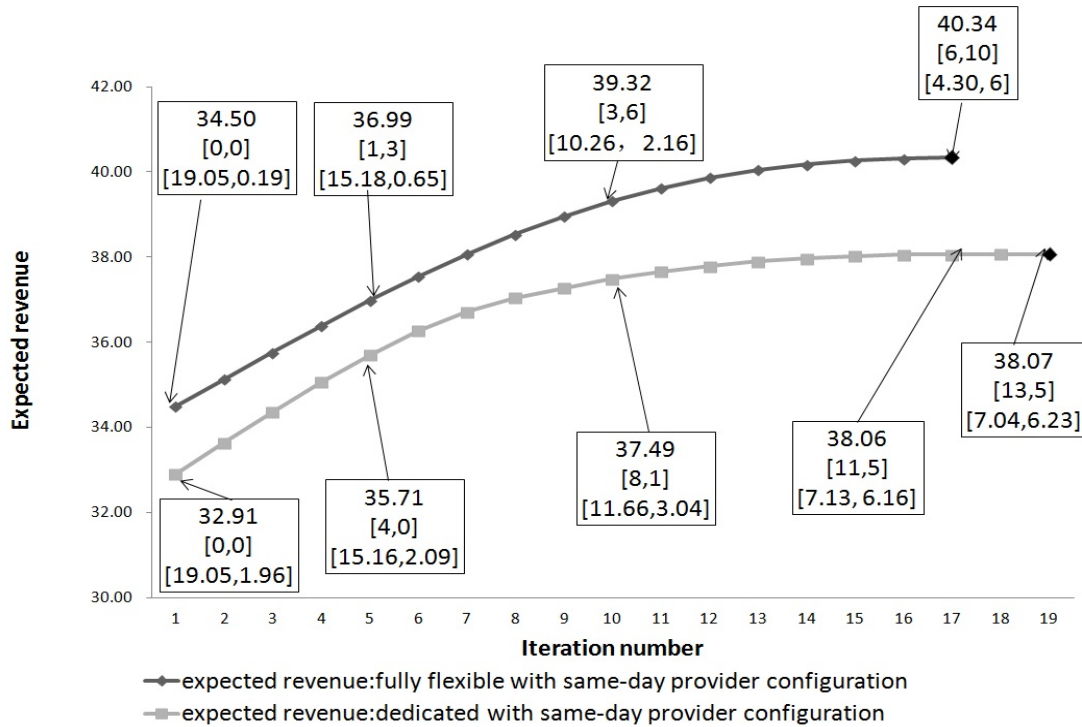
We have seen so far that same-day flexibility has a significant impact on the expected number of patients seen only when physicians have unequal utilizations.

But does flexibility help in reducing the probability of overtime as we expect it to? And if so, how significant is the impact of flexibility? Figure 5.6 shows the probability of missed same-day demand exceeding 6 patients in the fully flexible and dedicated 2-physician practice for 120% and 100% utilization, under symmetric and asymmetric cases. We clearly notice that the flexibility has a significant impact on reducing the probability of practice overtime exceeding 2 hours total (one hour per physician). For example, at  $Y = 0$ , in the 120% symmetric case, the overtime probability is 60%. For every 60 of 100 workdays, a physician in the practice will have to spend at least one hour of overtime, which seems unsustainable. Same-day flexibility reduces this probability significantly down to 40%, but this still seems high. A number of studies in the clinical literature have suggested that high workloads combined with low reimbursement rates are causing burnout among primary care physicians, which in turn has contributed to the nationwide shortage of primary care providers in the US. The overtime results of our model at 120% only confirm this reality.

Even at 100% utilization, the probability that the practice overtime exceeds 2 hours is nearly 30% in the dedicated symmetric case and nearly 35% in the dedicated asymmetric case, when  $Y = 0$ . Same-day flexibility brings this down in both symmetric and asymmetric cases to 18%. Flexibility helps to adjust the capacity allocation due to demand variability, however, the limited flexibility among physicians is not enough to appropriately satisfy appointment requests. More additional slots are needed in the system to reduce the risk of overlong overtime. As we observe, when  $Y$  increases, the probability of overtime decreases in both systems. If an asymmetric practice needs to manage the risk of overtime is larger than 2 hours total to be smaller than 10%, at least 8 more slots needs to be employed without flexibility among physicians. By adding same-day flexibility, about 5 slots can be saved to manage the risk.

In summary, while same-day flexibility does not have a significant impact on the expected revenue in the symmetric case, it does significantly reduce the probability of overtime in both symmetric and asymmetric cases. This makes the use of same-day flexibility worthwhile. We note that the other significant advantage of flexibility is that it allows clinics with uneven physician utilizations to see more prescheduled patients.

### 5.2.7 Sensitivity analysis: flexibility vs. booking limit



**Figure 5.7.** Performance of dedicated and fully-flexible systems as a function of the iteration number in our algorithm for a 2-physician practice, with  $y = 0$ , 120% workload and demand asymmetry

We have seen Figure 5.7 in section 3.4 once to discuss the searching path of the greedy heuristic. Now we discuss the sensitivity analysis based on flexibility vs. booking limit, still based on the searching path figure.

Note that the values shown in each frame contain three rows. First row gives the expected revenue at the current iteration; second row gives the  $N^p$  values of the two physicians at the current iteration; third row provides the expected number of missed prescheduled patients and the expected of missed same-day patients for the practice at the current iteration.

We see, as anticipated, that the expected revenue of the fully flexible case remains always above the dedicated case. We also see the property of diminishing returns. The benefits are higher in the early iterations. In the later iterations, an increase in the  $N^p$  values does not produce a significant change. For a practice, therefore, there exists some flexibility in the choice of  $N^p$  values - within a certain range the expected revenue does not change substantially. We can also use this property to stop the search earlier, by mandating that an increment in an  $N^p$  value (and therefore an additional iteration) is not necessary if the difference in revenue is less than some satisfactorily small (predetermined) value.

The expected missed (or unsatisfied) prescheduled and same-day demands are additional measures that provide useful perspective for a practice. In the early part of the search, as the  $N^p$  values are small, the number of missed prescheduled appointments is high. These appointments will have to be scheduled on other workdays. There is some flexibility around the specific day on which a prescheduled appointment is booked. But in general a high value of missed prescheduled demand implies more delays in obtaining a non-urgent appointment or scheduling follow-up appointments for patients with chronic conditions.

A high value of missed same-day appointments implies either that 1) the requesting patients visit an emergency room for their care, thereby increasing healthcare costs; and 2) the practice spends a significant amount of overtime. For example, the expected value of missed same-day appointments for the fully flexible case in the last iteration of the algorithm - when the optimal  $N^p$  values are reached - is 6. This

means that approximately 2 hours of overtime for the two physicians on average - or (approximately) an hour of overtime for each physician.

We will analyze the reasons for the significant difference in the optimal  $N^p$  values between the fully flexible and dedicated case for the asymmetric case again in the summary section.

### 5.3 Summary and conclusions

We compare dedicated with additional same-day provider configuration and fully flexible with additional same-day provider configuration under the following cases: 1) All physicians have identical utilizations (symmetric case); 2) Physicians differ in utilizations (asymmetric case); and 3) Physicians have identical utilizations but each physicians prescheduled and same-day distributions differ from that of the other physicians. The demands in our model are Poisson distributed. For the configurations where a same-day provider is introduced, we consider system utilizations of 100% and 120% in the absence of the additional resource. These cases are motivated based on our interactions with small private primary care practices as well as larger academic practices. For each of these cases, we also study the benefits of incrementally adding capacity to the new provider who sees same-day patients. Our conclusions can be summarized as follows:

While the loss of continuity has to be minimized for all appointments, we show that it can be appropriately sacrificed for urgent appointments needing immediate attention by introducing partial flexibility. More specifically, we have found that appropriately limiting the number of physicians a patient sees (hence promoting patient-physician continuity in urgent visits as well) can yield virtually the same timely access benefits as a system in which the patient is seen by any of the doctors in the practice.

The full flexibility configuration tends to plan for more prescheduled appointments than the dedicated case since the same-day appointments can be flexibly shared and

easily absorbed in the former but not in the latter. As the utilization of the system increases, there is a greater availability of same-day appointments which in our model produce greater revenue than prescheduled appointments. This prompts both the fully-flexible and dedicated configurations to decrease the number of prescheduled appointments so that more same-day appointments can be seen. The flexible configuration ends up offering even fewer prescheduled appointments than its dedicated counterpart, thus reserving more capacity for open access, since there is a higher probability of fully using the additional capacity when it is shared across all same-day appointments in the practice.

We find that the fully flexible configuration performs significantly better in the asymmetric case, i.e. when some physicians have higher demand in relation to others. In this case, flexibility is not only used to hedge against the variability in arriving same-day patient demands, but also to balance expected demand and available supply of each of the physicians. In the flexible system, the busier physician reserves more slots to satisfy prescheduled patient demands, while the lower utilized physician picks up the extra same-day appointment burden. Thus while flexibility implies a loss of continuity for same-day patients (who need it less anyway), it improves a physician's ability to provide more prescheduled appointments. These additional appointments can then be used for non-urgent but important follow-ups for patients with chronic conditions who are in greater need for continuity.

In order to provide adequate capacity to their increasing demand, primary care practices can choose to either (i) maintain continuity by restricting each physician to seeing patients of its own panel, but adding substantial capacity in the form of an additional provider (a nurse practitioner or physician assistant) to absorb the excess same-day demand; or (ii) allow full flexibility for physicians to see same-day patients and add minimal, if any, additional capacity. Not unsurprisingly, as more

capacity is added to the additional provider, the differences in performance between the dedicated and fully-flexible system decrease.

Generally, we observe that the fully flexible configuration help to achieve 2.70% more in revenue under 100% for symmetric practice. If adding flexibility for prescheduled patients, about 1% more revenue can be produced. For asymmetric case, the impact of same-day flexibility on revenue is 9.12% and impact of same-day and prescheduled flexibility is observed to be 10.30%. Compare to context in manufacturing, our report of impact of flexibility is much smaller. The reason is that, in primary care practices, we are usually facing to demand streams with small ratio of demand variance and demand mean, for example, poisson distribution, which is frequently used in health care research. For the small primary care practice, which has a smaller demand mean but larger variance, a much more impact of flexibility is expected to be seen.

## CHAPTER 6

### SIMULATION: APPOINTMENT SCHEDULING PROBLEM IN PRIMARY CARE PRACTICES UNDER DYNAMIC ARRIVALS

#### 6.1 Introduction

Thus far in the dissertation, we have focused on the impact of flexibility in primary care practices. Results in Chapter 4 and Chapter 5 are based on the framework described in Chapter 2, which are stochastic models at an aggregate level. In those models, in order to analyze different flexibility configurations, we assume that demands are realized and fulfilled instantly to avoid computational intractability. Based on this key assumption, we study the capacity allocation problem under two successively realized demand streams - prescheduled appointment requests and same-day appointment requests, and we also study the impact of flexibility in primary care practices.

However, in reality, appointment requests arrive over time. We not only need to know the capacity allocation for prescheduled patients vs. same-day patients, but also need to study what time slot a patient request should be scheduled to see a doctor. As mentioned before, a stochastic optimization model that considers such realistic issues would be both difficult to formulate and computationally intractable. We therefore create a simulation model under dynamic arrivals to study the described appointment scheduling problem. In this simulation model, we capture some typical realistic issues in primary care practices, such as patients' preference for time of the day, patients' willing to be diverted to another physician, dynamic and non-homogeneous same-day, etc. We also allow flexibility in the physician's capacity-sharing behavior in the

simulation model. Because we want to validate the threshold policy proposed in the aggregate model and we are also interested in the impact of flexibility in primary care practices under dynamic arrivals for both prescheduled and same-day requests.

In this chapter we first describe the details (assumptions, data, etc.) of the simulation model. Next, we describe results of the computational experiments. Our results revolve around the following research questions:

(i) As mentioned earlier, we capture prescheduled patients' preference for time of the day. Some prescheduled patients may prefer an early morning hour, while others may prefer an appointment during an afternoon hour and so on. If this is the case, then what is the impact of number of choices we allow patients to have on a practice's performance measures?

(ii) Some clinics may reserve parts of the workday for prescheduled patients, and leave the remaining slots for same-day patients. For example, some clinics usually allocate prescheduled appointment requests early in the morning and late in the afternoon and use the remaining slots for urgent same-day appointment requests. How much impact does the location of such blocked slots have? For example, what is the difference between the policy that a practice blocks the morning for prescheduled patients versus the policy that the practice blocks afternoon for prescheduled patients?

(iii) In Chapter 4 and Chapter 5, we propose a threshold policy for the capacity allocation problem. We are now interested in validating and testing the robustness of this threshold policy in a dynamic environment with patient time-of-day preferences.

(iv) We are also interested in the impact of flexibility in primary care practices under dynamic arrivals. In particular, are the findings from Chapter 4 robust under dynamic arrivals and patient preferences?

## 6.2 Simulation model and assumptions

We establish a simulation model to capture the appointment scheduling problem for a single workday. In our model, both prescheduled and same-day appointment requests arrive dynamically. We assume that the physician works 9:00am - 5:00pm each day (usually Monday - Friday for primary care practices). In reality, some physicians work longer than 8 hours and others may work part time. However, a full time physician's work time is typically 8 hours per workday.

As in the aggregate model, we still assume that each appointment slot is 20 minutes in length, so that each physician provides 24 slots in one single workday and that appointments are always scheduled in 20-minute increments, at 9:00am, 9:20am, 9:40am, etc. Note that this study does not deal with stochasticity in appointment durations; the focus is only on the appointment call in and slot allocation process.

Also note that, we still keep the capacity reservation policy in the simulation because we want to evaluate the proposed threshold policy for the booking limit under dynamic arrivals. For example, if we run simulation for a single physician practice and  $N^p = 10$ , then we accept prescheduled patient requests until the number of prescheduled appointments is equal to 10. In previous study based on aggregate model, we found that, with respect to expected revenue, actually reserving optimal  $N^p$  policy performs almost same as the no threshold policy. Recall from the aggregate case that the benefit of reserving an optimal quantity of slots,  $N^p$ , is to reduce the risk of long overtime to see same-day patients.

Our simulation model is capable of evaluating either single or multiple physician practices. In our study, we focus on a 3 physicians practice to test different flexibility configurations. We assume that different flexibility configurations could hold for both prescheduled patients and same-day patients, which means physicians could see patients from other panels if there is flexibility allowed to do so. We use similar

revenue coefficients as shown in Chapter 4 and Chapter 5 to make the results from dynamic model and the results from aggregate model comparable.

### 6.2.1 Modeling the appointment call-in process

Based on prescheduled demand mean and same-day demand mean, we calculate the average number of appointment requests per time unit (p-value for the Bernoulli trial), and then use Bernoulli trial to randomly determine whether a patient request arrives in that time unit or not.

Prescheduled and same-day requests arrive in different time-scales and we explain this below. Recall that we are simulating the appointment allocation process for a single workday. Prescheduled requests for a particular slot on this workday can arrive up to 3 months prior. This translates to approximately 63 workdays. If a practice's phone lines are open for 10 hours a day (8:00am to 6:00pm, as is often the case), then there are 630 hours in which these requests can come in. We assume that in one hour at most one prescheduled request arrives. If the mean prescheduled demand is 16 then the probability that a prescheduled request arrives in any hour is  $16/630 = 0.0253$ . Thus  $p=0.0253$  in the random Bernoulli trial for each hour in the simulation.

Same-day calls arrive only after the particular workday being studied begins - 8:00am in our model - and the practice stops satisfying same-day requests at 5:00pm since no more appointments are available beyond 5 pm. In total this is a 9-hour range. We assume at most one same-day call arrives each minute. We can generate a Bernoulli trial for each minute similar to the one described above. However, unlike the arrival of prescheduled requests, which we assume to be uniform due to lack of empirical data, we do have some data on frequency of calls by hour of day. Thus the p value of the Bernoulli trials for same-day requests in a particular hour can be raised or decreased to reflect this non-homogeneity. Figure 6.1 shows the relevant

time-scales of our simulation model for a two physician practice. Figure 6.2 ([46]) shows frequency of calls over a workday.

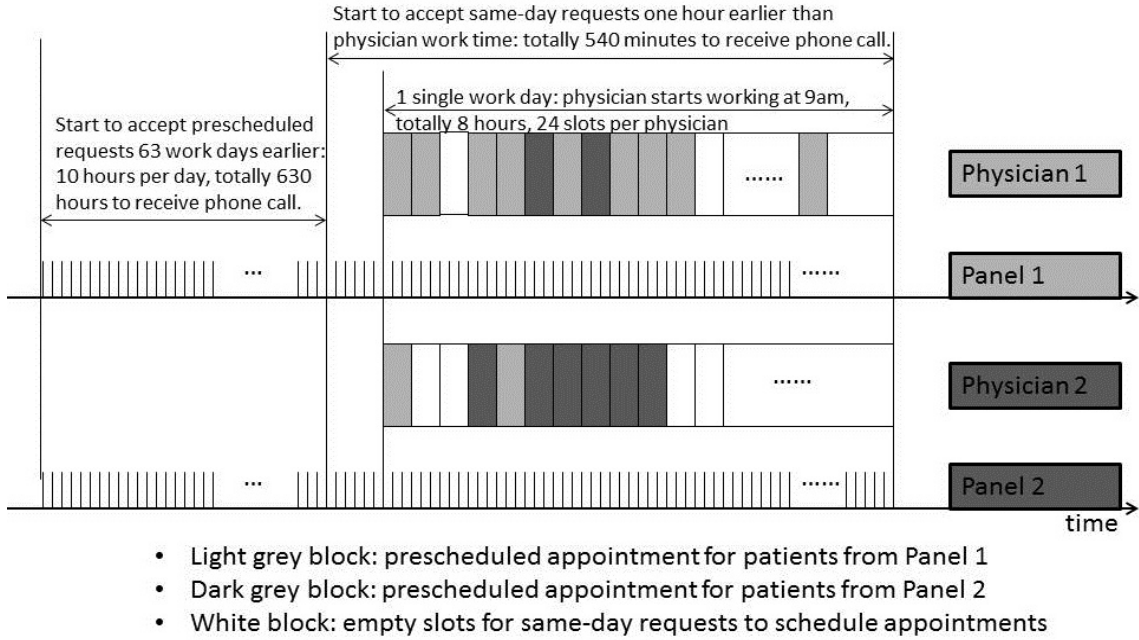


Figure 6.1. The simulation time-scales: example of two physicians practice

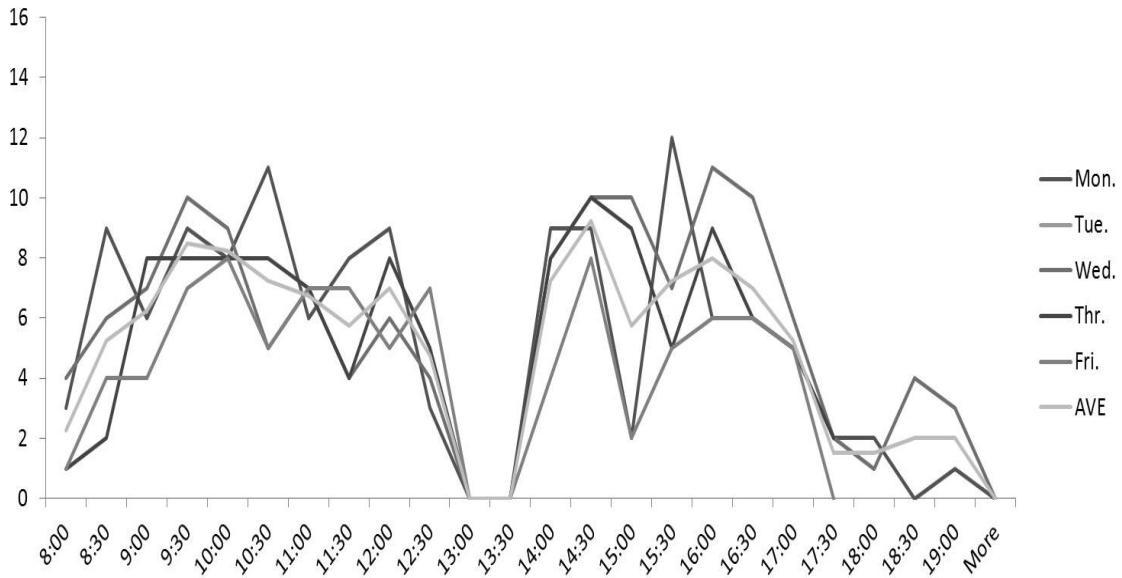


Figure 6.2. Phone call frequency over a workday

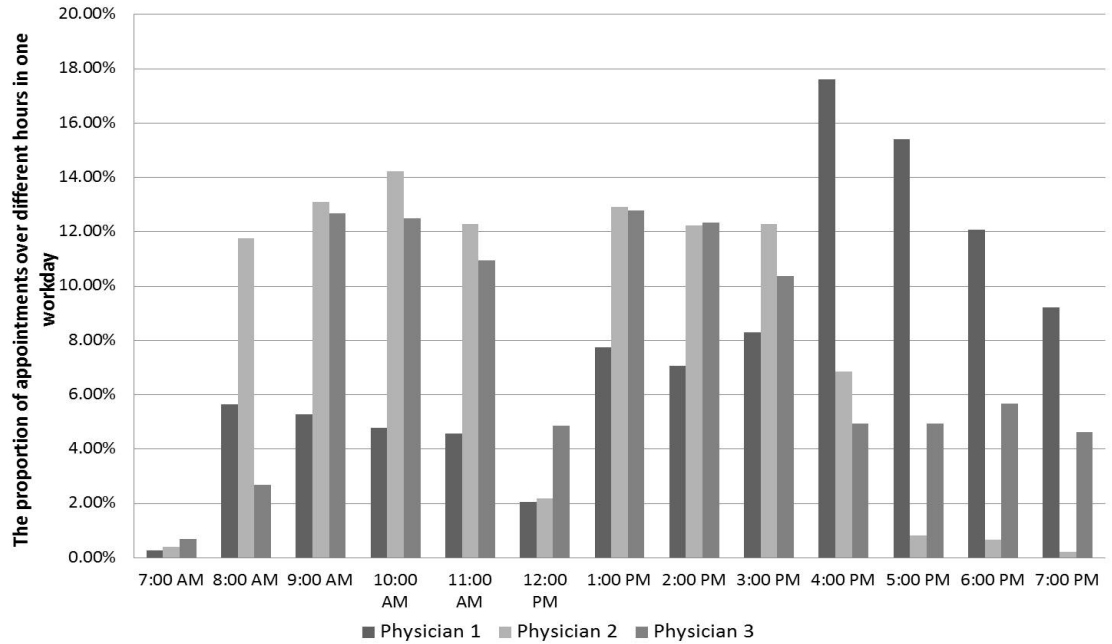
## 6.2.2 Modeling patient preferences

In our simulation model, we capture the important issue of patient preferences. Both prescheduled patients and same-day patients have preferences to scheduled appointments due to constraints on availability and also due to their need to see physician as quickly as possible. The final slot chosen is usually a complicated decision based on information exchanged between the patient and the scheduler.

Till now, few papers have studied the impact of patient's preferences. In primary care practices, patient's preference usually is impacted by two issues: (i) the time of the day the appointment is preferred, and (ii) if the patient insists on seeing his/her own physician. We call issue (i) as time-of-day preference and issue (ii) as physician preference.

### 6.2.2.1 Modeling time-of-day preferences

It is difficult to accurately measure the patients' time-of-day preferences. [53] studied appointment processes of a large health system and obtained historical appointment data concerning 37 primary care clinics that operate in urban, suburban, and rural areas. In their study, they used the *realized* appointment times to reflect patients' time-of-day preferences. They believed that, although the realized appointment times could not truly reflect patients' time-of-day preferences, it is very likely that clinics tried to respond to patients' needs. In our simulation model, similarly, we assume that the observed data of appointment times could reflect the likelihood that the patient would accept a slot at a particular time. We use the observed data on appointment frequencies by the hour in [53], see Figure 6.3, to generate an individual patient's time-of-day preferences.



**Figure 6.3.** Patients' time-of-day preferences

Patient preferences (for prescheduled requests) in our simulation model work in the following way. A patient calling in provides three choices for hour of day in decreasing order of preference; the three choices are randomly generated based on the observed distribution in Figure 6.3. The scheduler/clinic tries to provide a slot to the patient, depending on availability, in the same preference order. For example, suppose a patient's preferences are 10:00am - 11:00am, 4:00 pm - 5:00pm, and 9:00am - 10:00am, then the clinic would attempt to schedule this patient in an available slot in the 10:00am - 11:00am range first. If not successful, then the clinic would attempt to schedule this patient in an available slot 4:00 pm - 5:00 pm. If not successful again, then the clinic would attempt to finally search slots between 9:00am - 10:00am. If all attempts fail, this patient cannot get an appointment for this workday and has to be considered as prescheduled overflow.

For a same-day appointment request, we simply assume this patient could accept/prefer the earliest available slot after the phone call to schedule an appointment. This makes sense since same-day requests typically arise out of urgent needs.

### 6.2.2.2 Modeling physician preferences

We know that, some patients, especially prescheduled patients are more likely to insist on seeing their own physicians as continuity is more critical for these patients. In our simulation, we adapt the observed data for loyalty class shown in [53]. In their study, Wang and Gupta look at all patients with more than three visits, and then count the proportion of Patient-Physician matched visits among all his/her historical visits. For example, if a patient’s proportion of Patient-Physician matched visits is 0.24, then this patient’s loyalty class is 0.2 (there are 10 loyalty classes in increments of 0.1).

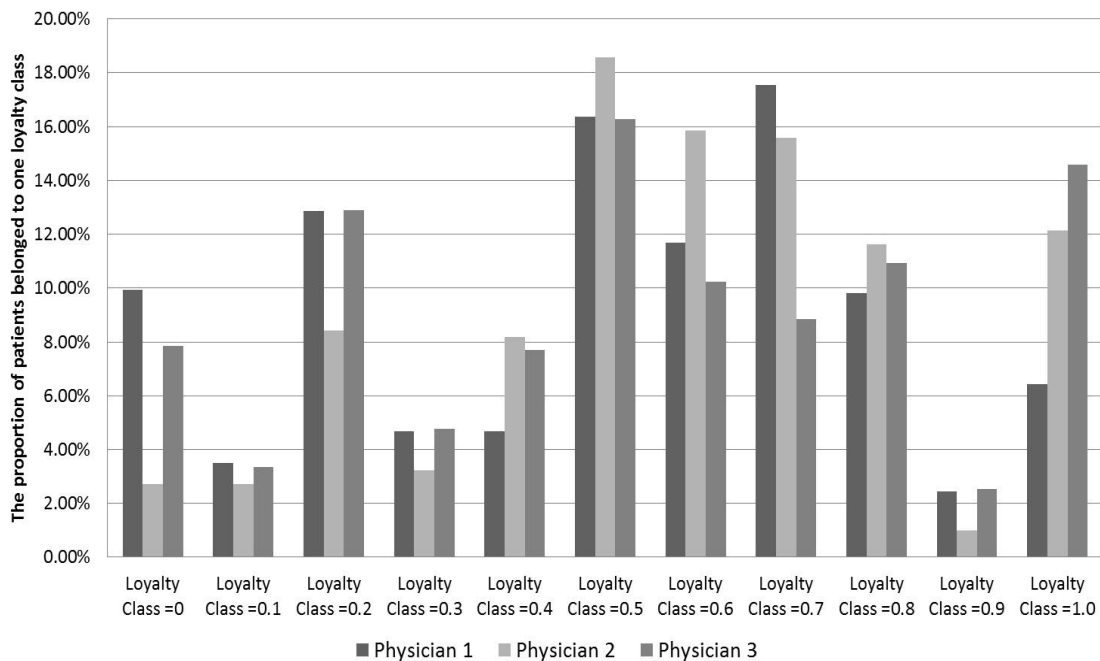


Figure 6.4. Patients’ physician preferences

In Figure 6.4, the x axis shows the loyalty class, the lowest is 0, and the highest is 1.0, with 10% increments. The y axis shows the proportion of a panel’s patients

that belonged to a particular loyalty class. For the data set shown in [53], patients in panel 1 have lowest level of loyalty while patients in panel 2 have highest level of loyalty.

Based on this reference, we use the loyalty proportion to generate patient's physician preference in our simulation. After we generate a prescheduled patient request, we also generate a random value based on the loyalty distribution in Figure 6.4 to decide this prescheduled patient's loyalty. Then another random value could decide if this patient will accept to be diverted to another physician (only for this particular request).

### 6.2.3 Modeling allocation process

Instead of looking for an optimal policy to schedule appointments, we propose some basic rules in our simulation model to allocate the appointment requests.

For prescheduled patients:

(1) Start with this patient's first time-of-day preference to see if her own physician has available slot (less than the given  $N^p$ ) during that hour. If yes, schedule this patient request with the earliest available slot during that hour. If not, go to the next time-of-day preference to search until find one available slot for this patient. If none of her time-of-day preferences matches, go to step (2).

(2) Track if this patient would accept being diverted to another physician. If not, then this prescheduled patient request is not satisfied. If yes, go to step (3).

(3) Start with this patient's first time-of-day preference to see if other physicians have available slot during that hour. If only one of them has availability, schedule this patient request with that physician's earliest available slot during that hour. If both of them have availability, schedule this patient request to the lowest utilized physician's earliest available slot during that hour. If neither have availability, go to the next time-of-day preference to search until an available slot is found. If still

none of her time-of-day preferences matches with availability, report this prescheduled patient request could not be satisfied.

For same-day patients, we propose two different scheduling policies: (1) Simply schedule the same-day request with first available slot after the phone call. If multiply physician have available slots at that time, schedule the request with the lower-utilized physician. This scheduling policy is reasonable in practice due to the urgent needs of same-day requests.

(2) Search if this patient’s own physician has available slot. If yes, schedule this patient request with the earliest available slot from her own physician. If not, search if there are available slots for other physicians. Schedule this patient request with the earliest available slot from other physicians. If two or more of them have available slot at the same time to be earliest, schedule this patient to the lowest-utilized physician. If none of the other physicians has available slot, report that this same-day patient request could not be satisfied.

Note that, if one prescheduled appointment request is not satisfied, it is possible for this patient to obtain a slot on another workday or it is also possible for this request to be “lost”. If a same-day appointment request is not satisfied, the patient is likely to be satisfied with physician overtime or refused. In our study, instead of assigning particular cost to these refusals, we simply output interesting measures to compare different policies or strategies. For example, number of satisfied requests, overflow values for prescheduled and same-day patients, revenue, etc.

Based on the simulation model we described in this section, we run computational experiments to study the appointment scheduling problem in primary care practices under two demand streams. We summarize our observations in section 6.3.

## 6.3 Computational results

### 6.3.1 Impact of prescheduled patient time-of-day flexibility

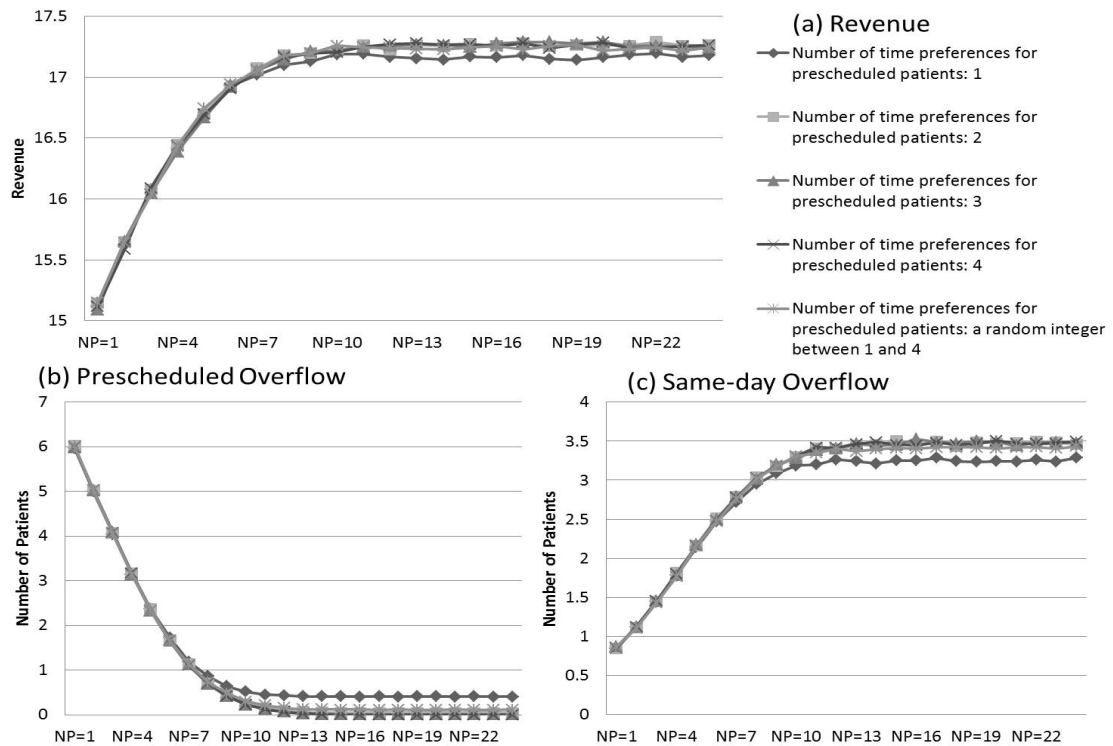
Recall from Section 6.2 that a prescheduled patient lists the hours she would to schedule the appointment in decreasing order of preferences. The first choice she prefers the most, the second choice is her second preference and so on. The farther down the list the patient has to go, the more the patient compromises with her preferences, and therefore the more flexible the patient has to be. In this section, we quantify the impact of patient time-of-day flexibility (actually this is a measure to show how flexible the patient could be with different slots of a day) on a practice's performance. We vary this flexibility by varying the number of time-of-day preferences the patient will allow. If there is only hour of the day the patient can schedule an appointment, such a patient has least flexibility. If the patient is willing to suggest another hour of the day to schedule the appointment, the patient is more flexible.

To study the impact of the number of time-of-day preferences, we test single physician practice, and focus on expected revenue, prescheduled overflow and same-day overflow. Simulations are run under 4 different demand ratios (4/20, 8/16, 16/8, and 20/4) and 3 different workloads (80%, 100%, 120%). Simulations are based on 20000 replications (see appendix B.1 for details about the discussion of how many replications are sufficient for our simulation model). In each setting, the value of  $N^p$  is increased from 1 to 24.

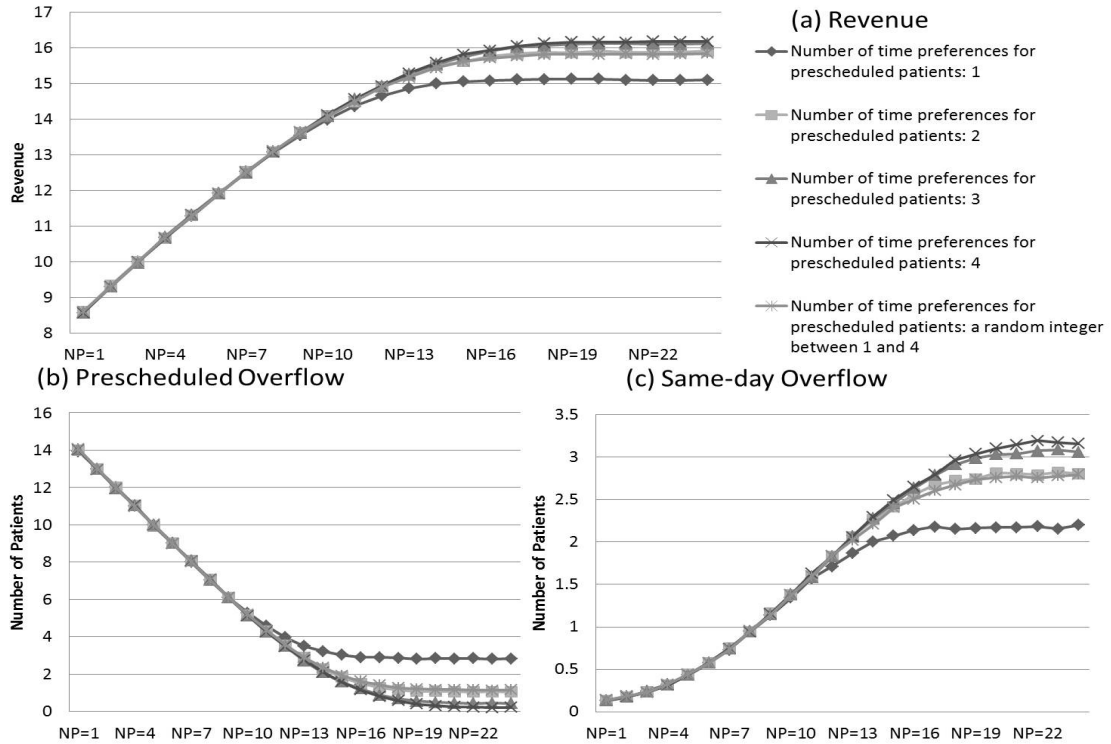
Surprisingly, we find that the practice performs quite well even if a patient has only 2 time-of-day choices. Thus the scheduler requires the patient to have only one additional time-of-day choice in addition to her first, most preferred choice. Asking the patient for further time-of-day choices does not contribute significantly to the overall practice performance. For example, Figure 6.5 and Figure 6.6 show the results of single physician under demand ratio 8/16, 100% workload and the results of single

physician under demand ratio 16/8, 100% workload. The results under other demand ratios and other workloads follow the same trend.

Note that, Figure 6.5 and Figure 6.6 are based on the observed data shown in [53], to generate prescheduled patients' time-of-day preferences. We also run simulations based on (1) uniformly distributed time-of-day preference (i.e. all hours of the day are equally preferred); and (2) observed appointment time in a small 3 provider practice in Massachusetts. Interestingly, we observe similar results: a little patient flexibility (2 choices for hour of the day for prescheduled patients) performs almost the same as requiring the patient to have more than 2 choices.



**Figure 6.5.** The impact of number of time preferences allowed for prescheduled patients on revenue, prescheduled overflow, and same-day overflow: single physician under demand ratio 8/16, 100% workload



**Figure 6.6.** The impact of number time preferences allowed for prescheduled patients on revenue, prescheduled overflow, and same-day overflow: single physician under demand ratio 16/8, 100% workload

### 6.3.2 Impact of guiding prescheduled patient appointment times

A practice may prefer to schedule prescheduled patients in certain predetermined blocks of the day. For example, [9] showed that the earlier in the day prescheduled appointments are scheduled, the better the practice’s ability of satisfy same-day appointments during the 8-hour workday. However, [9] does not consider the fact that prescheduled patients have time-of-day preferences. For example, a clinic may block early morning (9:00am - 11:00am) and late morning (11:00am - 1:00pm) for prescheduled patients. However, a prescheduled patient calling in may only want an afternoon appointment.

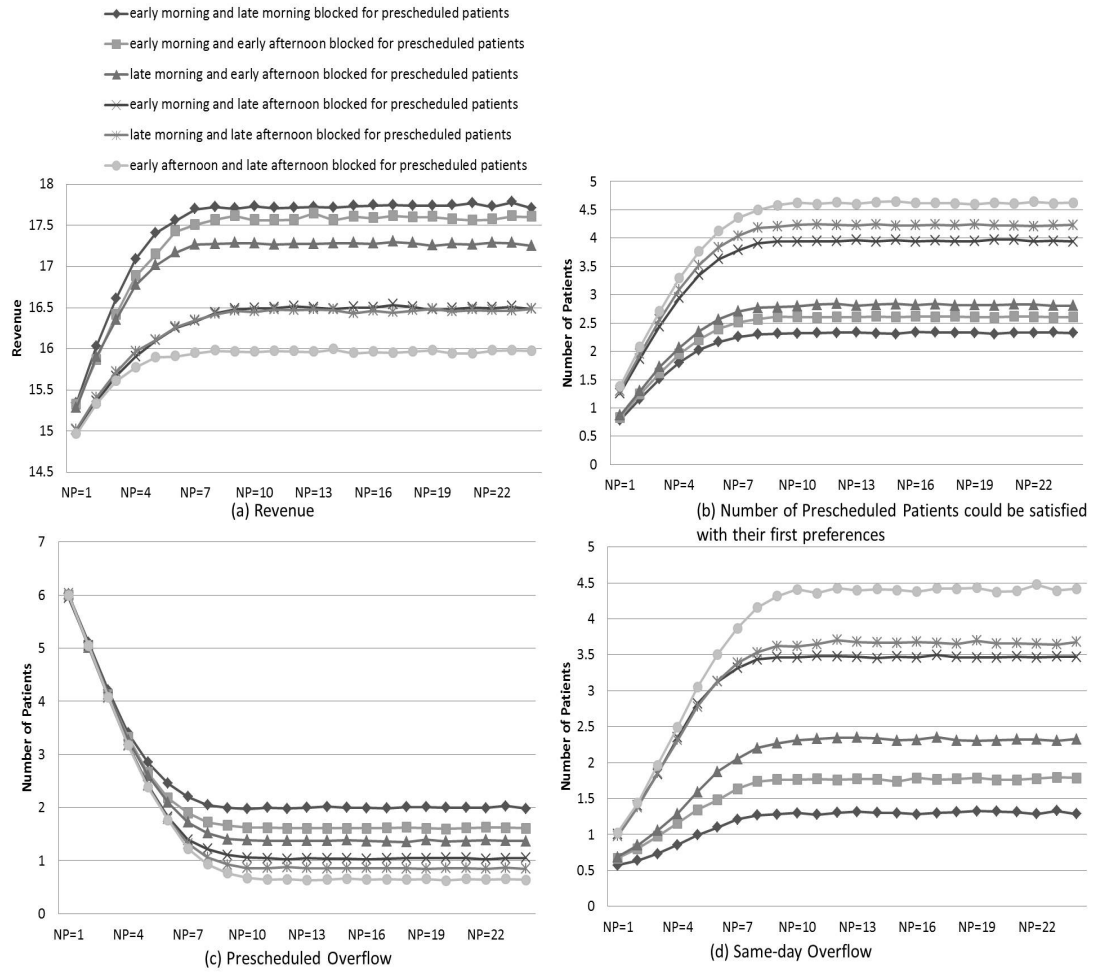
We therefore test cases where a practice blocks certain hours of the day for prescheduled appointments, and then simulate two situations: 1) The practice strictly adheres to these predetermined blocks for prescheduled patients; and 2) The practice

books prescheduled appointments outside of these blocks when necessary (i.e. relaxes its block constraints).

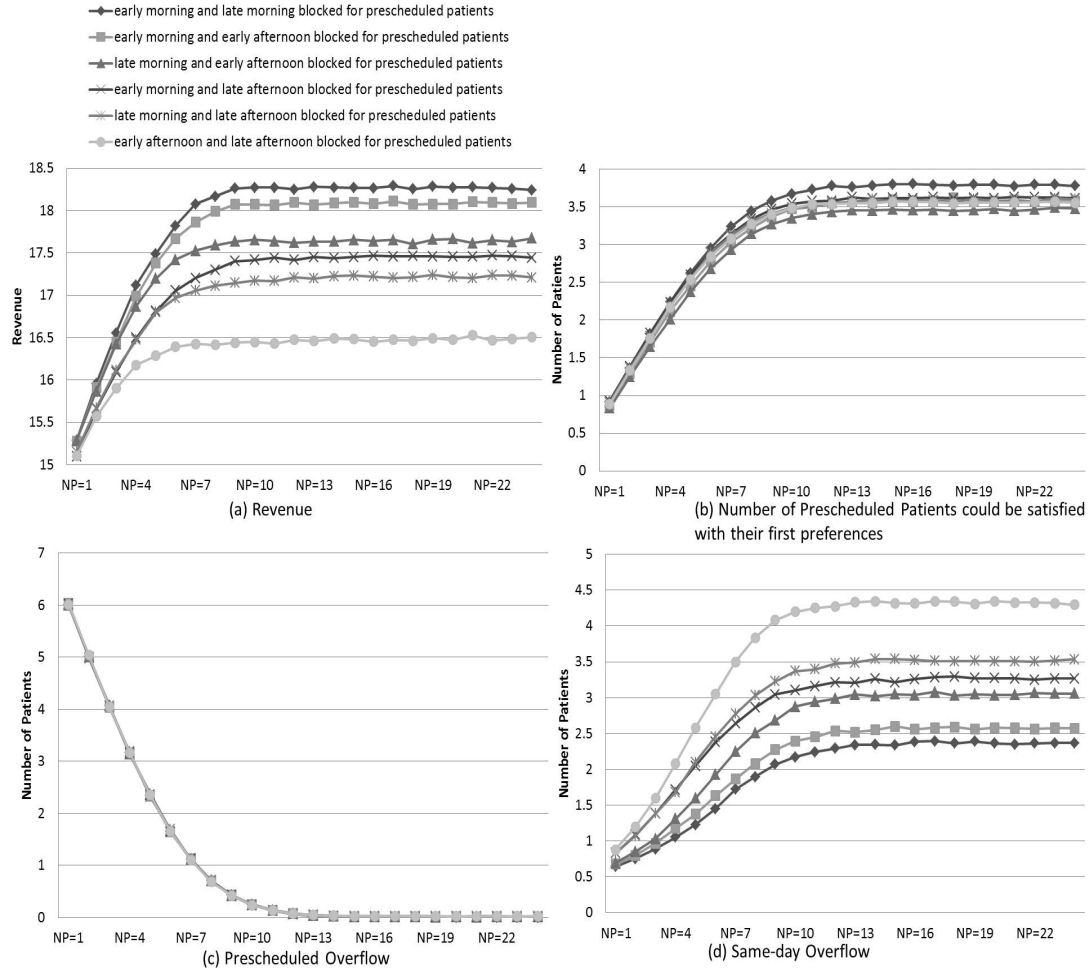
In the first or strict scenario, the practice checks if any of the 3 time-of-day choices for a prescheduled appointment (a) fall within the predetermined block, and (b) slot availability exists. If so, the practice schedules the appointment; otherwise the practice refuses the prescheduled appointment. The unfulfilled patient request is counted as a prescheduled overflow.

In the second, more relaxed scenario, if there is no overlap between the 3 time-of-day choices and the the predetermined blocks, the practice agrees to satisfy the preferences in decreasing order, depending on slot availability, outside of the blocks.

Based on these two different rules, we run simulations under 2 demand ratios (8/16 and 16/8), different workloads (80%, 100% and 120%) and different blocked time over one workday for a single physician. Note that, we still use 20,000 replications to test these scenarios. The performances are similar under different demand ratios and different workload. Here we present Figure 6.7 for the case with demand ratio 8/16 under 100% workload with refusals (i.e strict case) and Figure 6.8 for the case with demand ratio 8/16 under 100% workload without refusals (i.e. relaxed case).



**Figure 6.7.** Expected revenue, prescheduled overflow and same-day overflow: single physician with demand ratio 8/16 under 100% workload; clinic guides the patients to schedule appointments with refusals



**Figure 6.8.** Expected revenue, prescheduled overflow and same-day overflow: single physician with demand ratio 8/16 under 100% workload; clinic guides the patients to schedule appointments without refusals

The legends of Figure 6.7 and Figure 6.8 show six different time-block possibilities. In fact, regardless of how strict or relaxed the practice is about adhering to these blocks, the rank of these blocks with regard to expected revenue are same. We order these six policies from best to worst in the legends of Figure 6.7 and Figure 6.8.

Based on the same-day call frequency shown in Figure 6.2, a better policy is always the policy could leave more available slots for same-day patients in the afternoon. This is why the early morning and late morning blocks policy performs best and the early afternoon and late afternoon blocks policy performs the worst. In fact, we find that,

the fourth, fifth, and sixth policy all block late afternoon for prescheduled patients. In other words, although a large proportion of prescheduled patients prefer to be served with a late-afternoon slot, clinics should avoid guiding prescheduled patients to those hours, in order to leave sufficient slots to serve dynamically arriving same-day patients.

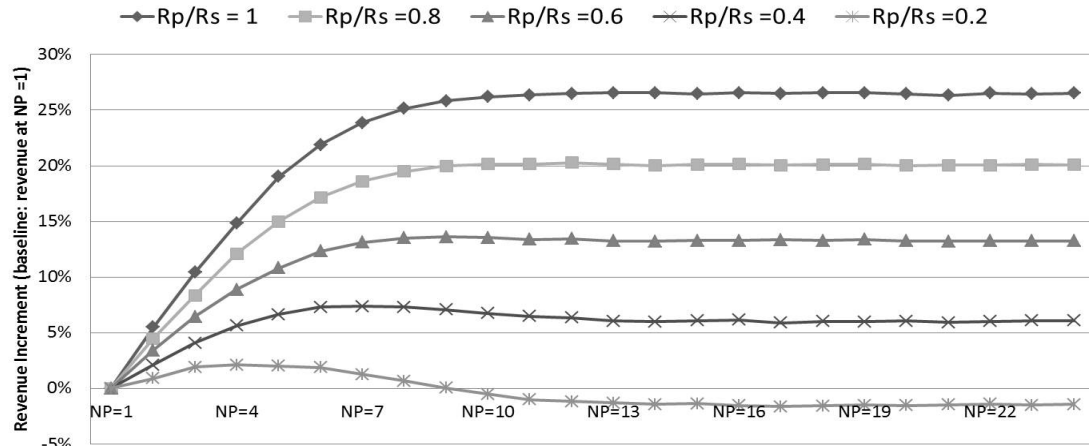
However, note that, there are three major differences between the performances depending on whether the practice is strict or relaxed in its adherence to the blocks. First, the revenues for the latter are generally larger than the former (around 0.7 higher) because the former has the risk to lose prescheduled patients whose time preferences do not overlap with the blocks. Secondly, the difference between the fourth policy (early morning and late afternoon blocks) and the fifth policy (late morning and late afternoon blocks) are much smaller for the rule without refusals. Finally, the prescheduled overflows are almost identical for each policy when the practice is relaxed in its adherence to the blocks. This is because prescheduled patients only get refused when there are not enough available slots (this is related to the  $N^p$  constraint), and not depending on the blocks chosen by the practice.

### **6.3.3 Evaluation of a threshold policy for the appointment scheduling problem under dynamic arrivals**

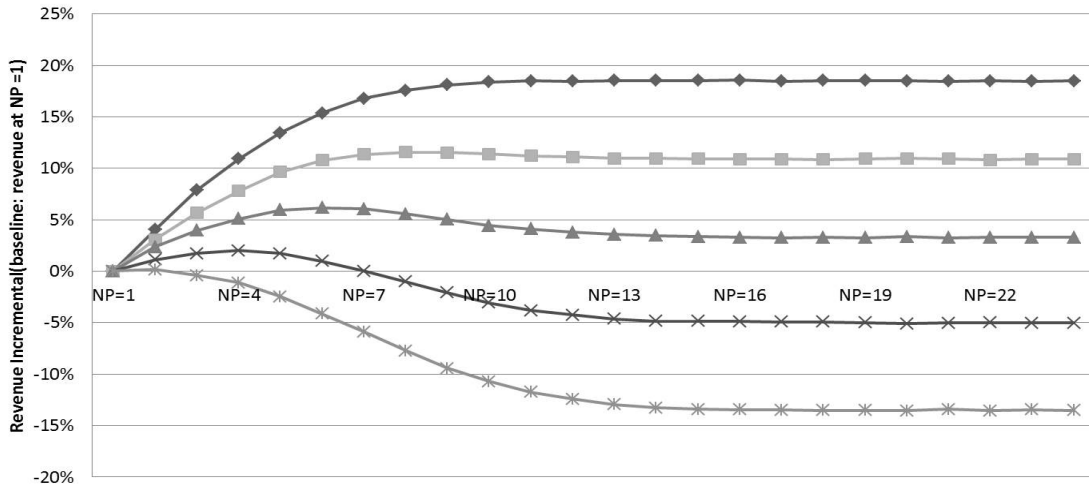
#### **6.3.3.1 Impact of different ratios of $R^p$ and $R^s$**

As discussed in Chapter 4, with respect to the expected revenue, reserving an optimal quantity of  $N^p$  is not essential for a primary care practice, because expected revenue is not sensitive to  $N^p$  beyond some point. We speculate that this is because of the small difference between  $R^p$  and  $R^s$  (currently 0.75 and 0.9). We now report on computational experiments for single physician practice, under different demand ratios (4/20, 8/16, 16/8, and 20/4), under different workloads (80%, 100%, 120%), and under different ratios of  $R^p$  and  $R^s$  (1, 0.8, 0.6, 0.4, and 0.2). Note that all the

simulations are based on 20,000 replications. The results are very similar so we only present results for 8/16, 100% and 120% (Figure 6.9) and results for 16/8, 100% and 120% (Figure 6.10) to discuss.

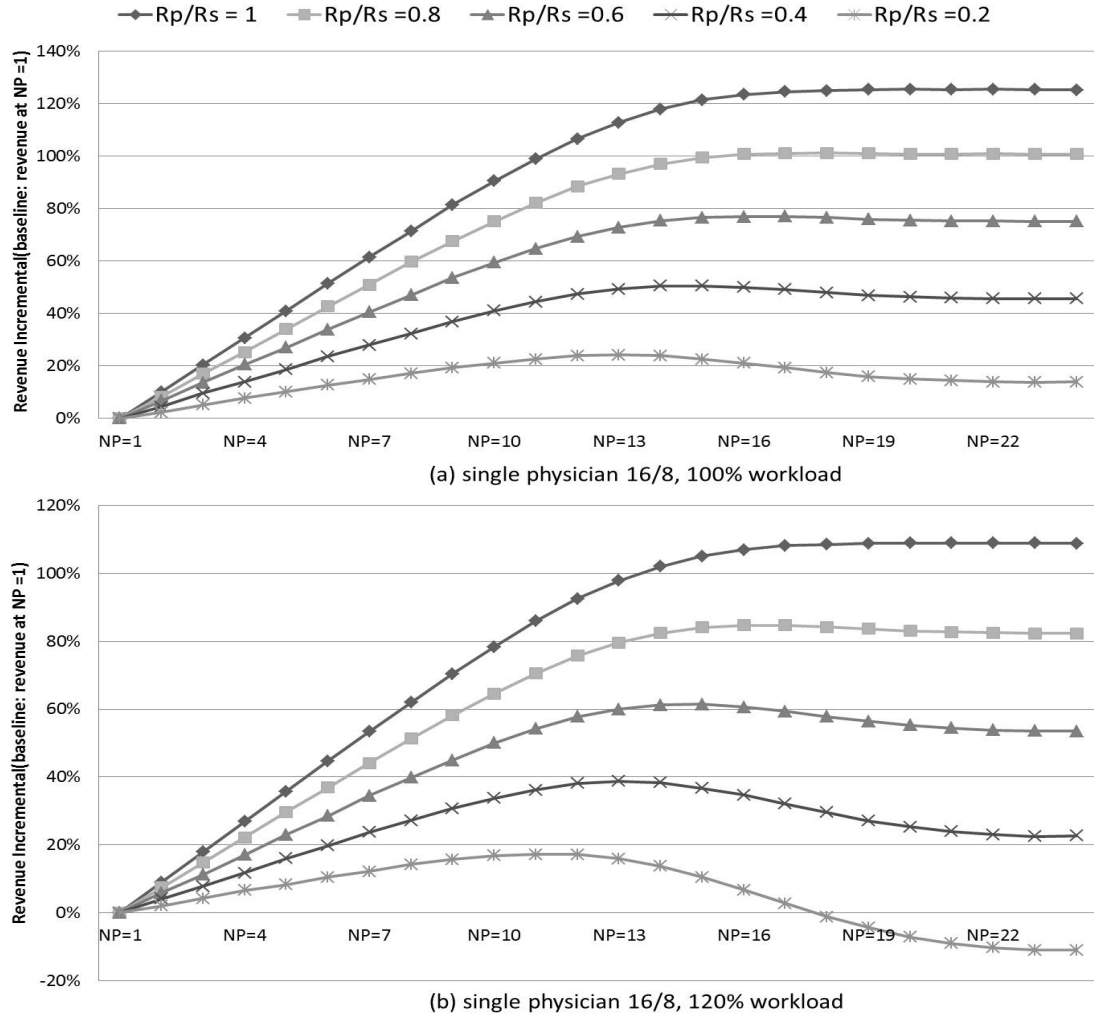


(a) single physician 8/16, 100% workload



(b) single physician 8/16, 120% workload

**Figure 6.9.** Expected revenue increment vs.  $N^P$  under different ratios of  $\frac{R^P}{R^S}$  : single physician under demand ratio 8/16, 100% and 120% workload



**Figure 6.10.** Expected revenue increment vs.  $N^P$  under different ratios of  $\frac{R^P}{R^S}$  : single physician under demand ratio 16/8, 100% and 120% workload

It turns out, as we speculated, that a small difference between  $R^p$  and  $R^s$  leads to insensitivity in revenue outcomes. From Figure 6.9 and Figure 6.10, we observe that the revenue becomes more sensitive to  $N^p$  when the ratio of  $R^p$  and  $R^s$  is smaller. Furthermore, for a given  $R^p$  and  $R^s$  ratio, the revenue is more sensitive to  $N^p$  when there is more prescheduled demand in the practice.

In summary, only when the ratio of  $R^p$  and  $R^s$  is small (for example, less than 0.6), or when the system is over-utilized, is it essential to reserve an optimal value of

$N^p$  for prescheduled demands; in all other cases, a booking limit is unnecessary with respect to expected revenue.

### 6.3.3.2 Validation of reserving optimal $N^p$ policy: single physician practices and multiple physician practices

The previous section evaluates the threshold policy with respect to the expected revenue. Although reserving optimal value of  $N^p$  does not affect revenue in the settings we are interested in, it does help to reduce the risk of long overtime due to same-day overflow. Does this finding still hold when patient requests arrive dynamically and when patients have preferences? The simulation model considers dynamical arrivals and other realistic issues like patients' preferences while the aggregate model does not. We present results for both single and multiple physician practices to answer this question.

To compare the aggregate models (proposed in chapter 2) and the simulation model (proposed in section 6.2), we present results from aggregate model and results from simulations, for single physician practice. In both aggregate and dynamic cases, we use the same values for  $R^p$  and  $R^s$  ( $R^p = 0.75$  and  $R^s = 0.9$ ). Four different demand ratios (4/20, 8/16, 16/8, and 20/4) and 3 different workloads (80%, 100%, and 120%) are tested. Results are summarized in Table 6.1 and Table 6.2.

**Table 6.1.** Comparisons of aggregate model and simulation model: single physician practice

		4/20		8/16		16/8		20/4	
		NP*	Revenue	NP*	Revenue	NP*	Revenue	NP*	Revenue
80%	Aggre	11	16.37	14	15.92	20	14.88	23	14.44
	Simul	10	15.53	14	14.86	21	13.75	22	13.31
100%	Aggre	8	19.25	12	18.66	19	17.51	22	16.97
	Simul	11	18.13	13	17.24	20	15.92	22	15.36
120%	Aggre	5	20.49	9	19.81	17	18.64	21	18.04
	Simul	6	19.49	11	18.51	19	17.21	23	16.54

Table 6.1 compares the optimal  $N^p$  and optimal revenue obtained from aggregate model and simulation model, under different scenarios. Note that, in Table 6.1, ‘Aggre’ denotes ‘aggregate model’, ‘Simul’ denotes ‘simulation model’, and ‘NP\*’ denotes ‘optimal  $N^p$ ’.

Generally, aggregate model could provide good guideline for clinic to reserve capacity for prescheduled patients due to following observations: (i) the optimal booking limit  $N^p$  reported by these two models are very similar, and (ii) with consideration of dynamic arrivals, the expected revenue based on the booking limit from aggregate model and the revenue under the actual optimal booking limit from simulations are quite close.

With the same input values of demand ratio and workload, the revenue obtained from the aggregate model is usually larger than the revenue obtained from the simulation model (difference range: 0.8-1.6). This observation is reasonable, because dynamic arrivals always carry an implied risk of losing same-day patients if the same-day patients arrive too late during a workday. The difference between the aggregate models and the dynamic models is always largest when the workload is perfectly balanced. Compared to a perfectly balanced system, when the system is low-utilized, generally there are many free slots in both models, so the difference is smaller. When the system is over-utilized, there are more same-day requests in relation to slots; the risk of slots going empty and idle is smaller, resulting a smaller difference between aggregate model and dynamic model .

Table 6.2 compares the expected revenue, prescheduled overflow, same-day overflow, and 95% percentile of same-day overflow under both optimal threshold policy and no threshold policy under dynamic arrivals and patient preferences. Note that, in Table 6.2, ‘Pre-Over’ denotes ‘prescheduled overflow’, ‘Same-Over’ denotes ‘same-day overflow’, ‘95% perc. ‘Same-Over’ denotes ‘95% percentile of same-day overflow’.

**Table 6.2.** Comparisons of reserving optimal  $N^p$  policy and no threshold policy: single physician practice

		P/S: 4/20				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	10	15.53	0	1.225	5
	No threshold	24	15.53	0	1.225	5
100%	Optimal NP	11	18.13	0.002	3.069	8
	No threshold	24	18.13	0.001	3.07	8
120%	Optimal NP	6	19.49	0.191	5.858	13
	No threshold	24	19.48	0.002	6.02	13
		P/S:8/16				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	14	14.86	0.019	1.408	5
	No threshold	24	14.86	0.019	1.408	5
100%	Optimal NP	13	17.24	0.054	3.28	9
	No threshold	24	17.24	0.04	3.292	9
120%	Optimal NP	11	18.51	0.381	6.027	13
	No threshold	24	18.51	0.101	6.256	13
		P/S:16/8				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	21	13.75	0.419	1.389	4
	No threshold	24	13.75	0.419	1.389	4
100%	Optimal NP	20	15.92	1.084	2.708	7
	No threshold	24	15.91	1.074	2.717	7
120%	Optimal NP	19	17.21	2.181	4.536	9
	No threshold	24	17.21	2.015	4.673	10
		P/S:20/4				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	22	13.31	0.995	0.863	3
	No threshold	24	13.31	0.995	0.863	3
100%	Optimal NP	22	15.36	2.408	1.56	4
	No threshold	24	15.37	2.382	1.572	4
120%	Optimal NP	23	16.54	4.345	2.629	7
	No threshold	24	16.54	4.345	2.629	7

As in section 4.2.2, with respect to expected revenue, the difference between the optimal threshold policy and no threshold policy is not significant, while reserving an optimal amount of slots could help to reduce the risk of long overtime (see 95% percentile of same-day overflow). However, the impact of reserving optimal  $N^p$  on reducing the risk of long overtime becomes even smaller under dynamic arrivals. The reason is that, in our simulation, many same-day patients call early afternoon and late afternoon (see Figure 6.2). Therefore not all same-day patients can be accommodated, unlike the aggregate case.

We also run simulations for multiple physician practices. We focus on a dedicated configuration for both prescheduled same-day patients to test 6 different demand inputs (symmetric 4/20, symmetric 8/16, symmetric 16/8, symmetric 20/4, asymmetric 6/12, 8/16, 10/20, and asymmetric 12/6, 16/8, 20/10) and 3 different workloads (80%, 100%, and 120%). Unlike previous results, all the simulations are based on 1000 replications with common random numbers (see appendix B.2 for details). The results under symmetric demands are summarized in Table 6.3 and the results under asymmetric demands are summarized in Table 6.4.

Table 6.3 and Table 6.4 compares the performances (expected revenue, prescheduled overflow, same-day overflow, and 95% percentile of same-day overflow) for both optimal threshold policy and no threshold policy under symmetric demands and asymmetric demands. The observations are same as the observations for single physician practice. That is, with respect to expected revenue, the difference between optimal threshold policy and no threshold policy is not significant, while the impact of reserving optimal  $N^p$  on reducing the risk of long overtime is also not significant. And this observation holds for both symmetric demands and asymmetric demands.

**Table 6.3.** Comparisons of reserving optimal  $N^p$  policy and no threshold policy: multiple physicians practice under symmetric demands

		sym 4/20				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	[10 10 10]	46.71	0.001	3.865	10
	No threshold	[24 24 24]	46.71	0	3.866	10
100%	Optimal NP	[8 8 8]	54.17	0.039	9.408	19
	No threshold	[24 24 24]	54.16	0.001	9.442	19
120%	Optimal NP	[5 5 5]	58.48	1.19	17.497	29
	No threshold	[24 24 24]	58.43	0.007	18.545	31
		sym 8/16				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	[13 13 13]	44.55	0.039	4.423	10
	No threshold	[24 24 24]	44.55	0.033	4.428	10
100%	Optimal NP	[16 16 16]	51.63	0.157	10.107	19
	No threshold	[24 24 24]	51.62	0.153	10.111	19
120%	Optimal NP	[11 11 11]	55.61	1.272	18.031	29
	No threshold	[24 24 24]	55.58	0.311	18.864	31
		sym 16/8				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	[21 21 21]	41.29	1.203	4.054	9
	No threshold	[24 24 24]	41.29	1.203	4.054	9
100%	Optimal NP	[21 21 21]	47.48	3.088	8.427	15
	No threshold	[24 24 24]	47.47	3.078	8.436	15
120%	Optimal NP	[21 21 21]	51.42	6.202	14.051	22
	No threshold	[24 24 24]	51.41	6.127	14.118	22
		sym 20/4				
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	[23 23 23]	39.79	3.077	2.49	6
	No threshold	[24 24 24]	39.79	3.077	2.49	6
100%	Optimal NP	[22 22 22]	45.84	7.187	4.682	9
	No threshold	[24 24 24]	45.84	7.163	4.705	9
120%	Optimal NP	[23 23 23]	49.76	13.153	7.863	14
	No threshold	[24 24 24]	49.76	13.153	7.863	14

**Table 6.4.** Comparisons of reserving optimal  $N^p$  policy and no threshold policy: multiple physicians practice under asymmetric demands

asym 10/20						
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	[13 13 12]	44.55	0.075	4.41	10
	No threshold	[24 24 24]	44.54	0.052	4.285	10
100%	Optimal NP	[13 12 14]	51.51	0.36	10.031	19
	No threshold	[24 24 24]	51.50	0.18	10.491	20
120%	Optimal NP	[12 10 13]	55.62	1.689	18.163	30
	No threshold	[24 24 24]	55.61	0.404	19.305	32
asym 20/10						
		NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	Optimal NP	[20 18 19]	41.91	1.462	3.205	7
	No threshold	[24 24 24]	42.88	1.404	3.194	8
100%	Optimal NP	[21 18 21]	48.41	3.514	6.951	13
	No threshold	[24 24 24]	48.40	3.583	7.209	14
120%	Optimal NP	[19 17 19]	52.37	8.178	11.394	19
	No threshold	[24 24 24]	52.33	6.929	12.619	22

### 6.3.4 Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences

In section 4.2, we test the impact of flexibility based on the aggregate capacity allocation model. We also find that when same-day patients are already fully flexible, adding additional flexibility to serve same-day patients doesn't obtain noticeable revenue improvements in that environment. Now, in the simulation, we are interested in the impact of same-day full flexibility and also the impact of additional prescheduled flexibility while same-day patients are already fully flexible.

We run simulations for multiple physicians practice under 3 different flexibility configurations to compare the impact of flexibility under dynamic arrivals. These 3 configurations are (i) prescheduled patients and same-day patients are both dedicated, (ii) prescheduled patients are dedicated and same-day patients are fully flexible, and

(iii) prescheduled patients are pooled to share a given amount of  $N^p$  for the entire practice and same-day patients are fully flexible. Note that, as discussed in section 2.5, a prescheduled pooled model always work slightly better than a prescheduled fully flexible model as long as same-day patients are fully flexibly shared. Also note that, the difference of performances between configuration (i) and configuration (ii) shows the impact of same-day flexibility and the difference between configuration (ii) and configuration (iii) shows the impact of the additional prescheduled flexibility while same-day patients are already fully flexible.

Six different demand inputs (symmetric 4/20, symmetric 8/16, symmetric 16/8, symmetric 20/4, asymmetric 6/12, 8/16, 10/20, and asymmetric 12/6, 16/8, 20/10 ) combined with 3 different workloads (80%, 100%, and 120%) are tested. All the simulations are based on 1000 replications with common random numbers.

The results under symmetric demands are summarized in Table 6.5 and Table 6.6. The results under asymmetric demands are summarized in Table 6.7. Note that, in Table 6.5, Table 6.6 and Table 6.7, ‘P-D S-D’ denotes configuration (i), ‘P-D S-F’ denotes configuration (ii), and ‘P-P S-D’ denotes configuration (iii). Other abbreviations like ‘Pre-Over’, ‘Same-Over’, etc., are same as explained in the previous section.

First, from Table 6.5 - Table 6.7, same-day flexibility is beneficial in both (i) increasing expected revenue and (ii) reducing the expected same-day overflow. Under symmetric demands, the improvement of revenue gained from same-day flexibility is in the 0.5% - 2.8% range and the expected same-day overflow can be reduced by 0.9-1.4 units due to same-day flexibility. Under asymmetric demands, the improvement of revenue gained from same-day flexibility is in the 3.6% - 5.6% range and the expected same-day overflow can be reduced by 2.5-3.9 units due to same-day flexibility. The impact of same-day flexibility is smaller than that obtained from the aggregate model, for the same reasons discussed earlier. An available slot that can be used in the

aggregate model when same-day requests are flexibly shared might go idle in the dynamic case.

Second, we know that in the aggregate model, same-day flexibility is most beneficial under balanced workload. However, in the dynamic case, although this observation still holds under most scenarios, we do observe inconsistency under two different scenarios. For example, when input demand is symmetric 4/20, under 80% workload, 100% workload, and 120% workload, same-day flexibility gains 2.2%, 1.5%, and 0.5% respectively. And when the input demand is symmetric 8/16, under 80% workload, 100% workload, and 120% workload, same-day flexibility gains 2.4%, 1.9%, and 1.1% respectively.

Finally, consistent with the findings in Section 4.2, when same-day patients are already fully flexible, the impact of additional prescheduled flexibility is unnoticeable. In other words, same-day flexibility is sufficient to balance demands with unused slots to achieve a higher revenue. In a dynamic environment prescheduled flexibility is not essential as long as same-day patients are flexibly shared.

**Table 6.5.** Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences: symmetric demands

sym 4/20						
		optimal NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	P-D S-D	[10 10 10]	46.71	0.001	3.865	10
	P-D S-F	[10 10 10]	47.78	0.001	2.561	8
	P-P S-F	20	47.78	0.001	2.561	8
100%	P-D S-D	[8 8 8]	54.17	0.039	9.408	19
	P-D S-F	[11 11 11]	55.02	0.001	8.31	18
	P-P S-F	20	55.02	0.011	8.301	18
120%	P-D S-D	[5 5 5]	58.48	1.19	17.497	29
	P-D S-F	[5 5 5]	58.82	1.19	16.914	29
	P-P S-F	12	58.83	2.125	16.116	28
sym 8/16						
		optimal NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	P-D S-D	[13 13 13]	44.55	0.039	4.423	10
	P-D S-F	[15 15 15]	45.70	0.033	3.029	9
	P-P S-F	32	45.70	0.033	3.029	9
100%	P-D S-D	[16 16 16]	51.63	0.157	10.107	19
	P-D S-F	[19 19 19]	52.69	0.153	8.743	18
	P-P S-F	37	52.69	0.153	8.743	18
120%	P-D S-D	[11 11 11]	55.61	1.272	18.031	29
	P-D S-F	[17 17 17]	56.30	0.317	17.839	30
	P-P S-F	33	56.30	0.638	17.569	29
sym 16/8						
		optimal NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	P-D S-D	[21 21 21]	41.29	1.203	4.054	9
	P-D S-F	[21 21 21]	42.45	1.203	2.667	7
	P-P S-F	50	42.45	1.204	2.666	7
100%	P-D S-D	[21 21 21]	47.48	3.088	8.427	15
	P-D S-F	[23 23 23]	48.87	3.078	6.741	14
	P-P S-F	57	48.87	3.078	6.741	14
120%	P-D S-D	[21 21 21]	51.42	6.202	14.051	22
	P-D S-F	[23 23 23]	52.73	6.127	12.481	21
	P-P S-F	60	52.73	6.154	12.458	21

**Table 6.6.** Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences: symmetric demands(continuit with Table 6.5)

sym 20/4						
		optimal NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	P-D S-D	[23 23 23]	39.79	3.077	2.49	6
	P-D S-F	[23 23 23]	40.63	3.077	1.49	4
	P-P S-F	60	40.63	3.077	1.49	4
100%	P-D S-D	[22 22 22]	45.84	7.187	4.682	9
	P-D S-F	[24 24 24]	46.99	7.163	3.336	8
	P-P S-F	63	46.99	7.164	3.335	8
120%	P-D S-D	[23 23 23]	49.76	13.153	7.863	14
	P-D S-F	[23 23 23]	51.00	13.153	6.375	12
	P-P S-F	68	51.00	13.155	6.373	12

**Table 6.7.** Impact of flexibility in primary care practices under dynamic arrivals and patients' preferences: asymmetric demands

asym 10/20						
		optimal NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	P-D S-D	[13 13 12]	44.55	0.075	4.41	10
	P-D S-F	[16 13 14]	46.93	0.048	1.593	6
	P-P S-F	30	46.93	0.049	1.592	6
100%	P-D S-D	[13 12 14]	51.51	0.36	10.031	19
	P-D S-F	[19 15 15]	54.40	0.181	6.652	17
	P-P S-F	38	54.39	0.174	6.658	17
120%	P-D S-D	[12 10 13]	55.62	1.689	18.163	30
	P-D S-F	[15 8 13]	57.60	1.961	15.335	28
	P-P S-F	30	57.59	1.268	15.924	28
asym 20/10						
		optimal NP	Revenue	Pre-Over	Same-Over	95% perc. Same-Over
80%	P-D S-D	[20 18 19]	41.91	1.462	3.205	7
	P-D S-F	[21 20 19]	44.24	1.441	0.478	3
	P-P S-F	54	44.24	1.443	0.476	3
100%	P-D S-D	[21 18 21]	48.41	3.514	6.951	13
	P-D S-F	[23 20 20]	51.73	3.418	3.068	10
	P-P S-F	59	51.73	3.42	3.066	10
120%	P-D S-D	[19 17 19]	52.37	8.178	11.394	19
	P-D S-F	[21 16 16]	54.88	8.302	8.187	17
	P-P S-F	54	54.89	7.212	9.078	18

Note that, results in Table 6.5 - Table 6.7 are all based on the scheduling policy that same-day requests are always assigned to the first available slot after phone call time. In other words, we didn't consider continuity of same-day requests due to their urgent needs in those runs. In fact, we also propose another scheduling policy to consider the continuity needs for same-day patients when same-day patients are flexible. The policy is, clinic will first try to find available slot for same-day patients from their own physicians. Only when their own physicians do not have matched availabilities, clinic will try to assign the requests with other physicians' availabilities.

In Table 6.8, the performances of these two policies under different cases are compared. ‘P-D S-F’ denotes ‘prescheduled is dedicated and same-day is fully flexible’ and ‘P-P S-F’ denotes ‘prescheduled is pooled and same-day is fully flexible’. These two configurations are tested under two different policies: Policy 1 and Policy 2. Here ‘Policy 1’ denotes ‘always assign prescheduled requests as early as possible’ and ‘Policy 2’ denotes ‘first assign with patient’s own physician then with other physicians’. Four different symmetric P/S ratios (4/20, 8/16, 16/8 and 20/4) are tested and each P/S ratio is repeated under 3 different workloads (80%, 100% and 120%). Revenues under Policy 1 and Policy 2 are presented and then we calculate the deduction of revenue in percentage. Observe that, with consideration of same-day patients’ continuity needs, the expected revenue is generally reduced by a small proportion (0% - 1.79%). In addition, regardless of policies, we observe that the benefit of additional prescheduled flexibility is always marginal.

**Table 6.8.** Comparisons of two same-day scheduling policies: continuity considered vs. continuity not considered for same-day patients

		P-D S-F			P-P S-F		
		Policy 1 revenue	Policy 2 revenue	deduction in %	Policy 1 revenue	Policy 2 revenue	deduction in %
sym 4/20	80%	47.78	46.93	-1.79%	47.78	46.93	-1.82%
	100%	55.02	55.02	0.00%	55.02	55.02	0.00%
	120%	58.82	58.65	-0.29%	58.83	58.63	-0.33%
sym 8/16	80%	45.70	44.94	-1.67%	45.70	44.94	-1.70%
	100%	52.69	52.44	-0.48%	52.69	52.44	-0.49%
	120%	56.30	56.20	-0.17%	56.30	56.15	-0.27%
sym 16/8	80%	42.45	41.75	-1.65%	42.45	41.75	-1.68%
	100%	48.87	48.33	-1.12%	48.87	48.32	-1.14%
	120%	52.73	52.33	-0.76%	52.73	52.31	-0.80%
sym 20/4	80%	40.63	40.18	-1.13%	40.63	40.18	-1.14%
	100%	46.99	46.45	-1.14%	46.99	46.45	-1.15%
	120%	51.00	50.48	-1.03%	51.00	50.48	-1.04%

## 6.4 Summary and conclusions

To study the capacity allocation problem under two successively realized demand streams under dynamic arrivals, we establish a simulation model in this chapter. In this simulation model, we capture some typical realistic issues in primary care practices, such as patients' preference over different time of the day, patients' willing to be diverted to another physician, dynamic and non-homogeneous same-day arrivals, etc. We also allow both prescheduled patients and same-day patients to share physicians' capacity based on designed flexibility configurations, to test the impact of flexibility under dynamic environment.

Prescheduled patients' time-of-day preferences are considered in the simulations by using clinic data. We study the impact of prescheduled patients' time-of-day preferences on the performances of the clinic through simulations under single physician practice. Interestingly, regardless of the distribution of prescheduled patients' time-of-day preferences, limited prescheduled patient time-of-day flexibility (2 hours-flexibility in our simulation framework) performs almost as well as more prescheduled patient time-of-day flexibility.

We also study the impact of guiding prescheduled patient appointment times over a workday because some clinics may reserve parts of the workday for prescheduled patients, and leave the remaining slots for same-day patients. Due to the large frequency of requests arrivals in the late afternoon and the risk of idle time before late afternoon, although the clinic data shows that prescheduled patients do prefer late-afternoon as their first choice to schedule appointments, clinic should be cautious to block late-afternoon for prescheduled patients, in order to leave sufficient later slots in a day for same-day patients.

In Chapter 4 and Chapter 5, we propose a threshold policy for the capacity allocation problem. Although the performances of a threshold policy and a no-threshold policy are almost identical due to the small difference between  $R^p$  and  $R^s$ , the thresh-

old policy has significant impact to reduce the risk of long overtime in the aggregate model. To validate and to test the robustness of this proposed threshold policy under a dynamic environment with patient time-of-day preferences, we run simulations under both single physician practice and multiple physicians practice. Different demand scenarios are tested to compare the results. Not surprising, the impact of an optimal threshold policy performs almost identical with no threshold policy, with respect to revenue. However, the impact of optimal threshold policy on reducing the risk of long overtime is less significant than this impact in the aggregate model. In fact, for most primary care practices (not too busy), due to the small difference between  $R^p$  and  $R^s$ , a threshold policy is not essential for clinics to be implemented. Simply accepting and scheduling all the patients requests due to clinic availability could obtain desirable revenue for clinics.

Finally, we also study the impact of flexibility in primary care practices under dynamic arrivals. The same-day flexibility gains more significantly under asymmetric demands (3.5% - 5.6%) than under symmetric demands (0.5% - 2.8%). Meanwhile, this gain under dynamic cases is less than the gain from same-day flexibility under aggregate cases, because a planned slot under aggregate cases is possible to be idle due to late same-day arrivals under dynamic cases. In addition, consistently with the findings in Chapter 4, when same-day patients are already fully flexibly served, the additional prescheduled flexibility has unnoticeable gain.

## CHAPTER 7

### DISCUSSION

#### 7.1 Applications in other contexts

In this dissertation, we focus on the study of physicians' capacity allocation problem in primary care practices. However, the resource allocation problems among two successively customer classes - non-urgent demands (scheduled in advance) vs. urgent demands (arrive at short notice) - can be commonly observed in many domains. The models proposed in Chapter 2 and the analytical results presented in Chapter 3 can be well applied to any such resource allocation problem in a production/service system, with the following characteristics: (i) two successively realized demand streams, one stream arrives and has to be scheduled in advance and the other stream arrives randomly and with urgent needs (ii) the total capacity to fill up demands is fixed. We now provide a few examples of such applications.

For example, we could commonly observe maintenance and repair service for a great variety of industrial or residential equipment (e.g. furnaces) under multiple demand classes. Our model could be well extended to this situation; however, as opposed to the greater continuity needs for prescheduled patients in primary care practices, the benefit from the continuity provided by a technician is particularly important for urgent demand. Because under these situations, the prescheduled demand is usually for standard maintenance operations, which any technician could effectively complete. And the urgent demand will require deeper knowledge of the equipment, spare part availability, and quick resolution, to provide quicker and more efficient service.

Striking a balance between non-urgent and urgent demand occurs routinely in other healthcare settings. For example, limited operating room and surgeon capacity in hospitals needs to be allocated to balance elective surgeries demand while simultaneously accommodating emergency surgeries.

If we consider a delivery company, the manager needs to decide how to allocate two-day delivery in order to leave sufficient capacity (i.e. the postman's time) to fill up the urgent needs of same-day delivery.

Another typical example is the air charter problem, which receives much attention recently. As opposed to an aircraft seats selling problem, air charter companies focus on the operations of renting an entire aircraft. The air charter service is involved in air ambulance, individual private aircraft itineraries, and some ad hoc air transportation. Due to the different costs of different aircrafts and considering urgent demand vs. the prescheduled calendar, air charter companies need to allocate their limited aircraft resources (helicopters and business jets).

## **7.2 Implications for primary care practices**

In this dissertation, we focus on the physicians' capacity allocation problem to study the impact of flexibility (allowing patients from different panels to share capacity from different physicians) in primary care practices. Generally, when the inherent physician flexibility is used to serve prescheduled patients as well as same-day patients, continuity in care for the prescheduled chronic patients suffers while minimal additional benefits in access are observed. Furthermore, the improvement obtained from the flexibility to serve prescheduled demands is not significant in increasing access even when same-day flexibility is not viable, in applications where the same-day demand has greater need for continuity than the prescheduled demand. Based on these findings, we recommend clinics to allow same-day flexibility to increase the access for patients; however, prescheduled flexibility is not essential in primary care

practices due to the greater continuity needs of prescheduled patients, as long as same-day patients are already flexibly shared. If clinics do prefer to involve prescheduled flexibility in practices, it is better introduced in the form of letting patients from different panels share a common booking limit while prescheduled patients are always served by their own physicians as long as the actual demand does not exceed the corresponding physician's total capacity. This design is particularly beneficial in primary care practices to maintain the continuity for prescheduled patients (for whom continuity is much more critical) while improving access of patients as a regular full flexibility configuration.

In our model, we suggest a booking limit for clinics to reserve capacity for prescheduled patients; however, in most primary care practices, there may not be any such booking limit especially in small primary care clinics. We evaluate the necessity of the booking limit policy through quantitative computational experiments. Generally, in a typical primary care practice, in which a prescheduled and same-day patient produce similar (or not quite different) amount of revenue, the expected revenue of the practice is surprisingly insensitive to the booking limit as long as the booking limit is sufficiently high. When the difference between revenues produced by a prescheduled patient and a same-day patient gets larger, the expected revenue of the practice becomes more and more sensitive to the booking limit. This finding seems to suggest that most practices could function appropriately without a booking limit, that is, simply accepting all prescheduled patient requests. However, we do suggest the booking limit policy for primary care practices due to two reasons: (i) reserving an optimal amount of slots for prescheduled patients could effectively reduce the risk of long over time of physicians even when a prescheduled patient and same-day patient produce similar revenue; (ii) In our study, the revenue coefficients for prescheduled and same-day patients are decided by show rate, resulting the expected revenue to be an estimate of number of patients seen, which is not a real profit measure. The

difference between profits produced by seeing one prescheduled patient and one same-day patient is definitely larger than what we used because of the higher cost of losing one same-day patients (same-day patients may end up visiting the emergency room, which results in huge expenditures in the health system). In that case, the booking limit policy should become more important to earn higher profit a clinic that is affiliated with a larger integrated health system. Based on this, we conclude that the booking limit policy is still essential for primary care practices. That is, once the booking limit for a particular physician is reached, any prescheduled request for this physician in future should be diverted to another physician or refused.

We also study the impact of a newly hired additional provider in primary care practices, where the additional provider is limited to serve same-day patients. Based on the results, adding a same-day provider and restricting it to only serve same-day patients is a good strategy for clinic to increase the access for patients, greater capacity for this additional provider could help to increase the access for prescheduled patients under asymmetric practices even though this provider can only serve same-day patient.

In Chapter 6, we establish a simulation model under dynamic arrivals to study the capacity allocation problem. To learn more insights for primary care clinics, we capture some realistic issues in practices, such as patients' preferences for time of the day, patients' willingness to be diverted to another physician, dynamic and non-homogeneous same-day arrivals, etc. Based on the results, we have two suggestions for primary care clinics. First, a better policy to achieve higher revenue for the primary care clinic is to always leave more slots for same-day patients in the afternoon even though some prescheduled patients do prefer late afternoon slots. Another suggestion is, although the aggregate model does not consider dynamic arrivals of patient requests, it does provide quite good guideline for practice. We suggest clinic to simply run the aggregate model to decide booking limit instead of running a complicated

simulation. However, note that, in terms of profit (“dollars earned” not “number of patients seen”), the gap between the estimated solution from aggregate model and a “true” optimal solution may become larger, and this needs further study.

### 7.3 Future study

For future study, there are multiple directions to extend our research: First, in Chapter 3, we computationally show that the greedy algorithm yields optimal solution for the capacity allocation problem under a range of demand scenarios, the variations of which could cover most primary care practices. We therefore propose the greedy algorithm to be an efficient heuristic for the capacity allocation problem in this dissertation. For future research, a rigorous analysis to demonstrate why the greedy algorithm yields optimal solution for this capacity allocation problem (or at least yields optimal solution for most practical scenarios) will be an interesting topic.

Second, in Chapter 5, we apply the framework in Chapter 2 to study the impact of additional provider in primary care practices. However, same-day patients dynamically arrive over a workday and some same-day patients may not agree to be served with the additional provider. Meanwhile, it is also highly possible for a patient to stay longer with the additional provider, compared to a regular visit with his/her own physician or any other physicians. Generally, an additional provider may not be well utilized even when the practice is busy. This topic needs further validation through empirical data.

Finally, in this dissertation, we always focus on a single workday model, both for the aggregate model in Chapter 2 - Chapter 5 and for the dynamic model in Chapter 6. However, a more realistic model to reflect appointment scheduling problem in primary care practices should be a multiple workday model, in which more factors needs to be considered. For example, if a prescheduled patient cannot be scheduled on a particular workday, usually she should be scheduled to another day. However,

will the patient's time preference match the available date? If it is not possible to schedule an appointment for this patient in a long time, will the patient be lost to the clinic? As a result, other issues like patient delay cost, patient dropout cost, etc., need to be appropriately captured and estimated based on clinical data.

## APPENDIX A

### EXPECTED REVENUE BASED ON ANALYTICAL METHOD

#### A.1 Prescheduled patients and same-day patients are both dedicated

For dedicated system under two demands streams, given demand realization  $[D_i^p, D_i^s]$ , the overflow for prescheduled demand is:

$$V^p(A) = \sum_{i=1}^m \max(0, D_i^p - N_i^p) \quad (\text{A.1})$$

while the overflow for same-day demand is:

$$V^s(A) = \sum_{i=1}^m \max(0, D_i^s - (C_i - \min[N_i^p, D_i^p])) \quad (\text{A.2})$$

Further, the expected overflow for prescheduled demand is:

$$E(V^p(A)) = E\left(\sum_{i=1}^m \max(0, D_i^p - N_i^p)\right) = \sum_{i=1}^m \sum_{o=N_i^p}^{\infty} (o - N_i^p) \cdot P(D_i^p = o) \quad (\text{A.3})$$

Note that the expected overflow for prescheduled demand will always remain same for all prescheduled dedicated configurations.

The expected overflow for same-day demand is given by

$$E(V^s(A)) = \sum_{i=1}^m \sum_{j=0}^{N_i^p} P(x_i^p = j) \cdot \sum_{o=C_i-j}^{\infty} (o - (C_i - j)) \cdot P(D_i^s = o) \quad (\text{A.4})$$

where  $x_i^p (i = 1, 2, \dots, m)$  denote the second stage decision variables, which can give us the number of slots fulfilled with prescheduled demand. As  $x_i^p = \min[N_i^p, D_i^p]$ , we

can easily obtain the probability distribution for  $x_i^p$ ,

$$P(x_i^p = a) = \begin{cases} P(D_i^p = a), a < N_i^p; \\ 1 - F_i(a - 1), a = N_i^p; \end{cases} \quad (\text{A.5})$$

From

$$ER(\cdot) = R^s \cdot \left[ \sum_{i=1}^m \lambda_i^s - E(V^p(A)) \right] + R^p \cdot \left[ \sum_{i=1}^m \lambda_i^p - E(V^s(A)) \right], \quad (\text{A.6})$$

we could compute the total expected revenue for this dedicated configuration based on  $E(V^p(A))$  and  $E(V^s(A))$ .

## A.2 Prescheduled patients are dedicated while same-day patients are fully flexibly shared

Given demand realization  $[D_i^p, D_i^s]$ , the prescheduled overflow is same as Equation A.1, while the same-day overflow is:

$$V^s(A) = \max\left\{0, \sum_{i=1}^m D_i^s - \sum_{i=1}^m (C_i - x_i^p)\right\}$$

Then the expected prescheduled overflow is equation A.3 and the expected same-day overflow is given as follows:

$$E(V^s(A)) = \sum_{j_m=0}^{N_m^p} \dots \sum_{j_1=0}^{N_1^p} \sum_{s=\sum_{i=1}^m C_i - \sum_{i=1}^m j_i}^{\infty} \left( s - \sum_{i=1}^m C_i + \sum_{i=1}^m j_i \right) \cdot P(D^s = s) \cdot P(x_1^p = j_1) \cdot \dots \cdot P(x_m^p = j_m) \quad (\text{A.7})$$

Here,  $D^s = D_1^s + D_2^s + \dots + D_m^s$  follows poisson distribution, and the distribution rate is given by  $\lambda^s = \sum_{i=1}^m \lambda_i^s$ .

From equation A.6, we compute the total expected revenue for prescheduled dedicated and same-day fully flexible configuration.

Note that, it is straightforward to calculate expected revenue for a subgroup configuration, because it is a combination of dedicated and fully flexibility configuration.

### **A.3 Prescheduled patients and same-day patients are both dedicated with their own physicians while one additional provider is added to serve all same-day patients**

#### **A.3.1 Analysis of dedicated with overflow system when $m=2$**

First, we name such configurations as ‘dedicated with overflow system’. As the dedicated with overflow system is more complicated than dedicated configuration, we start with  $m = 2$ .

For dedicated with overflow system, given demand realization  $[D_i^p, D_i^s]$ , the same-day overflow is very complex. To understand easily, we begin with the case of two physicians under same-day demand only. The overflow under demand realization  $[D_1, D_2]$  is:

$$V(A) = \max\{0, D_1 - C_1 - Y, D_2 - C_2 - Y, D_1 + D_2 - C_1 - C_2 - Y\}$$

where  $Y$  is the extra capacity assigned for the additional physician.

We want to calculate the expected value of  $V(A)$ . As 0 makes no effect to the value, we always ignore the term 0. Analyze the terms, we find that:

$$(1)V(A) = D_1 - C_1 - Y \iff \begin{cases} D_1 \geq C_1 + Y; \\ D_2 \leq C_2; \end{cases}$$

$$(2)V(A) = D_2 - C_2 - Y \iff \begin{cases} D_2 \geq C_2 + Y; \\ D_1 \leq C_1; \end{cases}$$

$$(3)V(A) = D_1 + D_2 - C_1 - C_2 - Y \iff \begin{cases} D_1 + D_2 \geq C_1 + C_2 + Y; \\ D_1 \geq C_1; \\ D_2 \geq C_2; \end{cases}$$

Note that, some conditions resulted from double counting cases should be substracted from above:

$$(a)D_1 - C_1 - Y = D_1 + D_2 - C_1 - C_2 - Y \text{ is maximum}$$

$$(b)D_2 - C_2 - Y = D_1 + D_2 - C_1 - C_2 - Y \text{ is maximum}$$

Then the conditions should be changed to:

$$(1)V(A) = D_1 - C_1 - Y \iff \begin{cases} D_1 \geq C_1 + Y; \\ D_2 < C_2; \end{cases}$$

$$(2)V(A) = D_2 - C_2 - Y \iff \begin{cases} D_2 \geq C_2 + Y; \\ D_1 < C_1; \end{cases}$$

$$(3)V(A) = D_1 + D_2 - C_1 - C_2 - Y \iff \begin{cases} D_1 + D_2 \geq C_1 + C_2 + Y; \\ D_1 \geq C_1; \\ D_2 \geq C_2; \end{cases}$$

From above analysis, we can get the expected overflow for dedicated with overflow system under demand realization  $[D_1, D_2]$ :

$$\begin{aligned} E[V(A)] &= P(D_2 < C_2) \cdot \left[ \sum_{i=C_1+Y}^{\infty} (i - (C_1 + Y)) \cdot P(D_1 = i) \right] \\ &\quad + P(D_1 < C_1) \cdot \left[ \sum_{j=C_2+Y}^{\infty} (j - (C_2 + Y)) \cdot P(D_2 = j) \right] \\ &\quad + \sum_{i=C_1}^{\infty} \sum_{j=\max(C_2, C_1+C_2+Y-i)}^{\infty} (i + j - (C_1 + C_2 + Y)) \cdot P(D_1 = i) \cdot P(D_2 = j) \end{aligned}$$

In summary, if the system is under two types of demands, the overflow for prescheduled demand is:

$$V^p(A) = \max(0, D_1^p - N_1^p) + \max(0, D_2^p - N_2^p)$$

while the overflow for same-day demand is:

$$V^s(A) = \max\{0, D_1^s - (C_1 - x_1^p + Y), D_2^s - (C_2 - x_2^p + Y), D_1^s + D_2^s - (C_1 + C_2 - x_1^p - x_2^p + Y)\}$$

The expected overflow for prescheduled demand is similarly with dedicated system and the expected overflow for same-day demand is provided by

$$\begin{aligned} E(V^s(A)) &= E[E(V^s(A)|x_1^p, x_2^p)] \\ &= \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right] \cdot \left[ \sum_{a=0}^{N_1^p} \sum_{i=C_1-a+Y}^{\infty} (i - (C_1 - a + Y)) \cdot P(D_1^s = i) \right] \\ &+ \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right] \cdot \left[ \sum_{b=0}^{N_2^p} \sum_{j=C_2-b+Y}^{\infty} (j - (C_2 - b + Y)) \cdot P(D_2^s = j) \right] \\ &+ \left[ \sum_{a=0}^{N_1^p} \sum_{b=0}^{N_2^p} \sum_{i=C_1-a}^{\infty} \sum_{j=\max(C_2-b, C_1+C_2-a-b+Y-i)}^{\infty} (i + j - C_1 - C_2 + a + b - Y) \right. \\ &\quad \left. \cdot P(D_1^s = i) \cdot P(D_2^s = j) \cdot P(x_1^p = a) \cdot P(x_2^p = b) \right] \end{aligned}$$

### A.3.2 Analysis of Dedicated with overflow system when m=4

By Jordan and Graves' results, following show us the overflow for same-day demand under demand realization  $[D_i]$  when  $m = 4$ :

$$\begin{aligned}
V(A) = \max\{ & 0, D_1 - (C_1 + Y), D_2 - (C_2 + Y), D_3 - (C_3 + Y), D_4 - (C_4 + Y), \\
& D_1 + D_2 - (C_1 + C_2 + Y), D_1 + D_3 - (C_1 + C_3 + Y), D_1 + D_4 - (C_1 + C_4 + Y), \\
& D_2 + D_3 - (C_2 + C_3 + Y), D_2 + D_4 - (C_2 + C_4 + Y), D_3 + D_4 - (C_3 + C_4 + Y), \\
& D_1 + D_2 + D_3 - (C_1 + C_2 + C_3 + Y), D_1 + D_2 + D_4 - (C_1 + C_2 + C_4 + Y), \\
& D_1 + D_3 + D_4 - (C_1 + C_3 + C_4 + Y), D_2 + D_3 + D_4 - (C_2 + C_3 + C_4 + Y), \\
& D_1 + D_2 + D_3 + D_4 - (C_1 + C_2 + C_3 + C_4 + Y)\}.
\end{aligned} \tag{A.8}$$

Analyze the terms in a similar way as  $m = 2$ , we have a similar conditions structure but contained 16 cases when  $m = 4$ (number of terms increase exponentially, only workable for small practices). Based on those disjoint conditions, we get the expected overflow for dedicated with overflow system under demand realization  $[D_1, D_2, D_3, D_4]$  as follow:

$$\begin{aligned}
E[V(A)] = & P(D_2 < C_2) \cdot P(D_3 < C_3) \cdot P(D_4 < C_4) \cdot \left[ \sum_{i=C_1+Y}^{\infty} (i - (C_1 + Y)) \cdot P(D_1 = i) \right] \\
& + P(D_1 < C_1) \cdot P(D_3 < C_3) \cdot P(D_4 < C_4) \cdot \left[ \sum_{j=C_2+Y}^{\infty} (j - (C_2 + Y)) \cdot P(D_2 = j) \right] \\
& + P(D_1 < C_1) \cdot P(D_2 < C_2) \cdot P(D_4 < C_4) \cdot \left[ \sum_{k=C_3+Y}^{\infty} (k - (C_3 + Y)) \cdot P(D_3 = k) \right] \\
& + P(D_1 < C_1) \cdot P(D_2 < C_2) \cdot P(D_3 < C_3) \cdot \left[ \sum_{l=C_4+Y}^{\infty} (l - (C_4 + Y)) \cdot P(D_4 = l) \right] \\
& + P(D_3 < C_3) \cdot P(D_4 < C_4) \cdot \left[ \sum_{i=C_1+Y}^{\infty} \sum_{j=\max(C_2, C_1+C_2+Y-i)}^{\infty} (i + j - (C_1 + C_2 + Y)) \cdot P(D_1 = i) \cdot P(D_2 = j) \right] \\
& + P(D_2 < C_2) \cdot P(D_4 < C_4) \cdot \left[ \sum_{i=C_1+Y}^{\infty} \sum_{k=\max(C_3, C_1+C_3+Y-i)}^{\infty} (i + k - (C_1 + C_3 + Y)) \cdot P(D_1 = i) \cdot P(D_3 = k) \right] \\
& + P(D_2 < C_2) \cdot P(D_3 < C_3) \cdot \left[ \sum_{i=C_1+Y}^{\infty} \sum_{l=\max(C_4, C_1+C_4+Y-i)}^{\infty} (i + l - (C_1 + C_4 + Y)) \cdot P(D_1 = i) \cdot P(D_4 = l) \right] \\
& + P(D_1 < C_1) \cdot P(D_4 < C_4) \cdot \left[ \sum_{j=C_2+Y}^{\infty} \sum_{k=\max(C_3, C_2+C_3+Y-j)}^{\infty} (j + k - (C_2 + C_3 + Y)) \cdot P(D_2 = j) \cdot P(D_3 = k) \right] \\
& + P(D_1 < C_1) \cdot P(D_3 < C_3) \cdot \left[ \sum_{j=C_2+Y}^{\infty} \sum_{l=\max(C_4, C_2+C_4+Y-j)}^{\infty} (j + l - (C_2 + C_4 + Y)) \cdot P(D_2 = j) \cdot P(D_4 = l) \right] \\
& + P(D_1 < C_1) \cdot P(D_2 < C_2) \cdot \left[ \sum_{k=C_3+Y}^{\infty} \sum_{l=\max(C_4, C_3+C_4+Y-k)}^{\infty} (k + l - (C_3 + C_4 + Y)) \cdot P(D_3 = k) \cdot P(D_4 = l) \right] \\
& + P(D_4 < C_4) \cdot \left[ \sum_{i=C_1+Y}^{\infty} \sum_{j=C_2+Y}^{\infty} \sum_{k=\max(C_3, C_1+C_2+C_3+Y-i-j)}^{\infty} (i + j + k - (C_1 + C_2 + C_3 + Y)) \cdot P(D_1 = i) \cdot P(D_2 = j) \cdot P(D_3 = k) \right] \\
& + P(D_3 < C_3) \cdot \left[ \sum_{i=C_1+Y}^{\infty} \sum_{j=C_2+Y}^{\infty} \sum_{l=\max(C_4, C_1+C_2+C_4+Y-i-j)}^{\infty} (i + j + l - (C_1 + C_2 + C_4 + Y)) \cdot P(D_1 = i) \cdot P(D_2 = j) \cdot P(D_4 = l) \right] \\
& + P(D_2 < C_2) \cdot \left[ \sum_{i=C_1+Y}^{\infty} \sum_{k=C_3+Y}^{\infty} \sum_{l=\max(C_4, C_1+C_3+C_4+Y-i-k)}^{\infty} (i + k + l - (C_1 + C_3 + C_4 + Y)) \cdot P(D_1 = i) \cdot P(D_3 = k) \cdot P(D_4 = l) \right] \\
& + P(D_1 < C_1) \cdot \left[ \sum_{j=C_2+Y}^{\infty} \sum_{k=C_3+Y}^{\infty} \sum_{l=\max(C_4, C_2+C_3+C_4+Y-j-k)}^{\infty} (j + k + l - (C_2 + C_3 + C_4 + Y)) \cdot P(D_2 = j) \cdot P(D_3 = k) \cdot P(D_4 = l) \right] \\
& + \left[ \sum_{i=C_1+Y}^{\infty} \sum_{j=C_2+Y}^{\infty} \sum_{k=C_3+Y}^{\infty} \sum_{l=\max(C_4, C_1+C_2+C_3+C_4+Y-i-j-k)}^{\infty} (i + j + k + l - (C_1 + C_2 + C_3 + C_4 + Y)) \cdot P(D_1 = i) \cdot P(D_2 = j) \cdot P(D_3 = k) \cdot P(D_4 = l) \right]
\end{aligned} \tag{A.9}$$

Further, suppose we are facing to two demand streams, then following show us the overflow for same-day demand under demand realization  $[D_i^p, D_i^s]$ ,

$$\begin{aligned}
V^s(A) = \max\{ & 0, D_1^s - (C_1 - x_1^p + Y), D_2^s - (C_2 - x_2^p + Y), D_3^s - (C_3 - x_3^p + Y), \\
& D_4^s - (C_4 - x_4^p + Y), D_1^s + D_2^s - (C_1 + C_2 - x_1^p - x_2^p + Y), D_1^s + D_3^s - (C_1 + C_3 - x_1^p - x_3^p + Y), \\
& D_1^s + D_4^s - (C_1 + C_4 - x_1^p - x_4^p + Y), D_2^s + D_3^s - (C_2 + C_3 - x_2^p - x_3^p + Y), \\
& D_2^s + D_4^s - (C_2 + C_4 - x_2^p - x_4^p + Y), D_3^s + D_4^s - (C_3 + C_4 - x_3^p - x_4^p + Y), \\
& D_1^s + D_2^s + D_3^s - (C_1 + C_2 + C_3 - x_1^p - x_2^p - x_3^p + Y), \\
& D_1^s + D_2^s + D_4^s - (C_1 + C_2 + C_4 - x_1^p - x_2^p - x_4^p + Y), \\
& D_1^s + D_3^s + D_4^s - (C_1 + C_3 + C_4 - x_1^p - x_3^p - x_4^p + Y), \\
& D_2^s + D_3^s + D_4^s - (C_2 + C_3 + C_4 - x_2^p - x_3^p - x_4^p + Y), \\
& D_1^s + D_2^s + D_3^s + D_4^s - (C_1 + C_2 + C_3 + C_4 - x_1^p - x_2^p - x_3^p - x_4^p + Y)\}.
\end{aligned}$$

The expected overflow for same-day demand is provided by,

$$E(V^s(A)) = \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right] \cdot \left[ \sum_{c=0}^{N_3^p} P(D_3^s < C_3 - c) \cdot P(x_3^p = c) \right] \\ \cdot \left[ \sum_{d=0}^{N_4^p} P(D_4^s < C_4 - d) \cdot P(x_4^p = d) \right] \cdot \left[ \sum_{a=0}^{N_1^p} \sum_{i=C_1-a+Y}^{\infty} (i - (C_1 - a + Y)) \cdot P(D_1^s = i) \cdot P(x_1^p = a) \right] \quad (\text{A.10})$$

$$+ \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right] \cdot \left[ \sum_{c=0}^{N_3^p} P(D_3^s < C_3 - c) \cdot P(x_3^p = c) \right] \\ \cdot \left[ \sum_{d=0}^{N_4^p} P(D_4^s < C_4 - d) \cdot P(x_4^p = d) \right] \cdot \left[ \sum_{b=0}^{N_2^p} \sum_{j=C_2-b+Y}^{\infty} (j - (C_2 - b + Y)) \cdot P(D_2^s = j) \cdot P(x_2^p = b) \right] \quad (\text{A.11}) \\ + \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right] \cdot \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right]$$

$$\cdot \left[ \sum_{d=0}^{N_4^p} P(D_4^s < C_4 - d) \cdot P(x_4^p = d) \right] \cdot \left[ \sum_{c=0}^{N_3^p} \sum_{k=C_3-c+Y}^{\infty} (k - (C_3 - c + Y)) \cdot P(D_3^s = k) \cdot P(x_3^p = c) \right] \quad (\text{A.12}) \\ + \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right] \cdot \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right]$$

$$\cdot \left[ \sum_{c=0}^{N_3^p} P(D_3^s < C_3 - c) \cdot P(x_3^p = c) \right] \cdot \left[ \sum_{d=0}^{N_4^p} \sum_{l=C_4-d+Y}^{\infty} (l - (C_4 - d + Y)) \cdot P(D_4^s = l) \cdot P(x_4^p = d) \right] \quad (\text{A.13}) \\ + \left[ \sum_{c=0}^{N_3^p} P(D_3^s < C_3 - c) \cdot P(x_3^p = c) \right] \cdot \left[ \sum_{d=0}^{N_4^p} P(D_4^s < C_4 - d) \cdot P(x_4^p = d) \right]$$

$$\cdot \left[ \sum_{b=0}^{N_2^p} \sum_{a=0}^{N_1^p} \sum_{i=C_1-a}^{\infty} \sum_{j=\max(C_2-b, C_1+C_2-a-b+Y-i)}^{\infty} (i + j - C_1 - C_2 + a + b - Y) \right. \\ \left. \cdot P(D_1^s = i) \cdot P(D_2^s = j) \cdot P(x_1^p = a) \cdot P(x_2^p = b) \right] \quad (\text{A.14}) \\ + \left[ \sum_{d=0}^{N_4^p} P(D_4^s < C_4 - d) \cdot P(x_4^p = d) \right] \cdot \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right]$$

$$\cdot \left[ \sum_{c=0}^{N_3^p} \sum_{a=0}^{N_1^p} \sum_{i=C_1-a}^{\infty} \sum_{k=\max(C_3-c, C_1+C_3-a-c+Y-i)}^{\infty} (i + k - C_1 - C_3 + a + c - Y) \right. \\ \left. \cdot P(D_1^s = i) \cdot P(D_3^s = k) \cdot P(x_1^p = a) \cdot P(x_3^p = c) \right] \quad (\text{A.15}) \\ + \left[ \sum_{c=0}^{N_3^p} P(D_3^s < C_3 - c) \cdot P(x_3^p = c) \right] \cdot \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right]$$

$$\cdot \left[ \sum_{d=0}^{N_4^p} \sum_{a=0}^{N_1^p} \sum_{i=C_1-a}^{\infty} \sum_{l=\max(C_4-d, C_1+C_4-a-d+Y-i)}^{\infty} (i + l - C_1 - C_4 + a + d - Y) \right. \\ \left. \cdot P(D_1^s = i) \cdot P(D_4^s = l) \cdot P(x_1^p = a) \cdot P(x_4^p = d) \right] \quad (\text{A.16}) \\ + \left[ \sum_{d=0}^{N_4^p} P(D_4^s < C_4 - d) \cdot P(x_4^p = d) \right] \cdot \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right]$$

$$\cdot \left[ \sum_{c=0}^{N_3^p} \sum_{b=0}^{N_2^p} \sum_{j=C_2-b}^{\infty} \sum_{k=\max(C_3-c, C_2+C_3-b-c+Y-j)}^{\infty} (j + k - C_2 - C_3 + b + c - Y) \right. \\ \left. \cdot P(D_2^s = j) \cdot P(D_3^s = k) \cdot P(x_2^p = b) \cdot P(x_3^p = c) \right] \quad (\text{A.17}) \\ + \left[ \sum_{c=0}^{N_3^p} P(D_3^s < C_3 - c) \cdot P(x_3^p = c) \right] \cdot \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right]$$

$$\cdot \left[ \sum_{d=0}^{N_4^p} \sum_{b=0}^{N_2^p} \sum_{j=C_2-b}^{\infty} \sum_{l=\max(C_4-d, C_2+C_4-b-d+Y-j)}^{\infty} (j + l - C_2 - C_4 + b + d - Y) \right. \\ \left. \cdot P(D_2^s = j) \cdot P(D_4^s = l) \cdot P(x_2^p = b) \cdot P(x_4^p = d) \right] \quad (\text{A.18})$$

$$\begin{aligned}
& + \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right] \cdot \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right] \\
& \cdot \left[ \sum_{d=0}^{N_4^p} \sum_{c=0}^{N_3^p} \sum_{k=C_3-c}^{\infty} \sum_{l=\max(C_4-d, C_3+C_4-c-d+Y-k)}^{\infty} (k+l-C_3-C_4+c+d-Y) \right. \\
& \quad \left. \cdot P(D_3^s = k) \cdot P(D_4^s = l) \cdot P(x_3^p = c) \cdot P(x_4^p = d) \right] \quad (\text{A.19}) \\
& + \left[ \sum_{d=0}^{N_4^p} P(D_4^s < C_4 - d) \cdot P(x_4^p = d) \right] \cdot \left[ \sum_{c=0}^{N_3^p} \sum_{b=0}^{N_2^p} \sum_{a=0}^{N_1^p} \sum_{i=C_1-a}^{\infty} \sum_{j=C_2-b}^{\infty} \sum_{k=\max(C_3-c, C_1+C_2+C_3-a-b-c+Y-i-j)}^{\infty} \right. \\
& \quad (i+j+k-C_1-C_2-C_3+a+b+c-Y) \\
& \quad \left. \cdot P(D_1^s = i) \cdot P(D_2^s = j) \cdot P(D_3^s = k) \cdot P(x_1^p = a) \cdot P(x_2^p = b) \cdot P(x_3^p = c) \right] \quad (\text{A.20}) \\
& + \left[ \sum_{c=0}^{N_3^p} P(D_3^s < C_3 - c) \cdot P(x_3^p = c) \right] \cdot \left[ \sum_{d=0}^{N_4^p} \sum_{b=0}^{N_2^p} \sum_{a=0}^{N_1^p} \sum_{i=C_1-a}^{\infty} \sum_{j=C_2-b}^{\infty} \sum_{l=\max(C_4-d, C_1+C_2+C_4-a-b-d+Y-i-j)}^{\infty} \right. \\
& \quad (i+j+l-C_1-C_2-C_4+a+b+d-Y) \\
& \quad \left. \cdot P(D_1^s = i) \cdot P(D_2^s = j) \cdot P(D_4^s = l) \cdot P(x_1^p = a) \cdot P(x_2^p = b) \cdot P(x_4^p = d) \right] \quad (\text{A.21}) \\
& + \left[ \sum_{b=0}^{N_2^p} P(D_2^s < C_2 - b) \cdot P(x_2^p = b) \right] \cdot \left[ \sum_{d=0}^{N_4^p} \sum_{c=0}^{N_3^p} \sum_{a=0}^{N_1^p} \sum_{i=C_1-a}^{\infty} \sum_{k=C_3-c}^{\infty} \sum_{l=\max(C_4-d, C_1+C_3+C_4-a-c-d+Y-i-k)}^{\infty} \right. \\
& \quad (i+k+l-C_1-C_3-C_4+a+c+d-Y) \\
& \quad \left. \cdot P(D_1^s = i) \cdot P(D_3^s = k) \cdot P(D_4^s = l) \cdot P(x_1^p = a) \cdot P(x_3^p = c) \cdot P(x_4^p = d) \right] \quad (\text{A.22}) \\
& + \left[ \sum_{a=0}^{N_1^p} P(D_1^s < C_1 - a) \cdot P(x_1^p = a) \right] \cdot \left[ \sum_{d=0}^{N_4^p} \sum_{c=0}^{N_3^p} \sum_{b=0}^{N_2^p} \sum_{j=C_2-b}^{\infty} \sum_{k=C_3-c}^{\infty} \sum_{l=\max(C_4-d, C_2+C_3+C_4-b-c-d+Y-j-k)}^{\infty} \right. \\
& \quad (j+k+l-C_2-C_3-C_4+b+c+d-Y) \\
& \quad \left. \cdot P(D_2^s = j) \cdot P(D_3^s = k) \cdot P(D_4^s = l) \cdot P(x_2^p = b) \cdot P(x_3^p = c) \cdot P(x_4^p = d) \right] \quad (\text{A.23}) \\
& + \left[ \sum_{d=0}^{N_4^p} \sum_{c=0}^{N_3^p} \sum_{b=0}^{N_2^p} \sum_{a=0}^{N_1^p} \sum_{i=C_1-a}^{\infty} \sum_{j=C_2-b}^{\infty} \sum_{k=C_3-c}^{\infty} \sum_{l=\max(C_4-d, C_1+C_2+C_3+C_4-a-b-c-d+Y-i-j-k)}^{\infty} \right. \\
& \quad (i+j+k+l-C_1-C_2-C_3-C_4+a+b+c+d-Y) \\
& \quad \left. \cdot P(D_1^s = i) \cdot P(D_2^s = j) \cdot P(D_3^s = k) \cdot P(D_4^s = l) \cdot P(x_1^p = a) \cdot P(x_2^p = b) \cdot P(x_3^p = c) \cdot P(x_4^p = d) \right] \quad (\text{A.24})
\end{aligned}$$

Then we can compute the total expected revenue for dedicated with overflow system through the above equation A.10 - A.24 immediately.

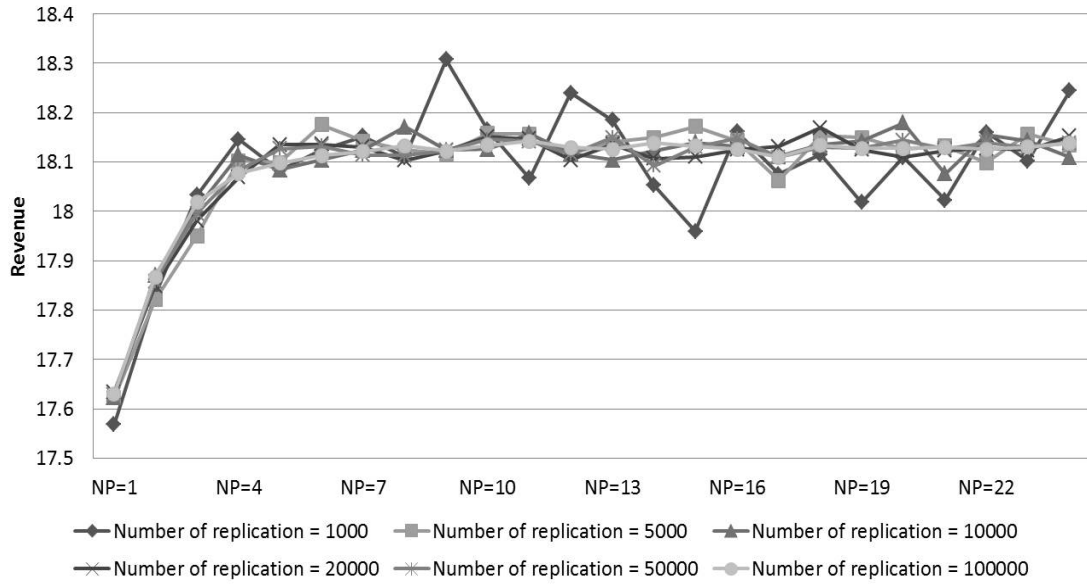
## APPENDIX B

### HOW TO REDUCE VARIANCE OF OUTPUT FOR OUR SIMULATIONS?

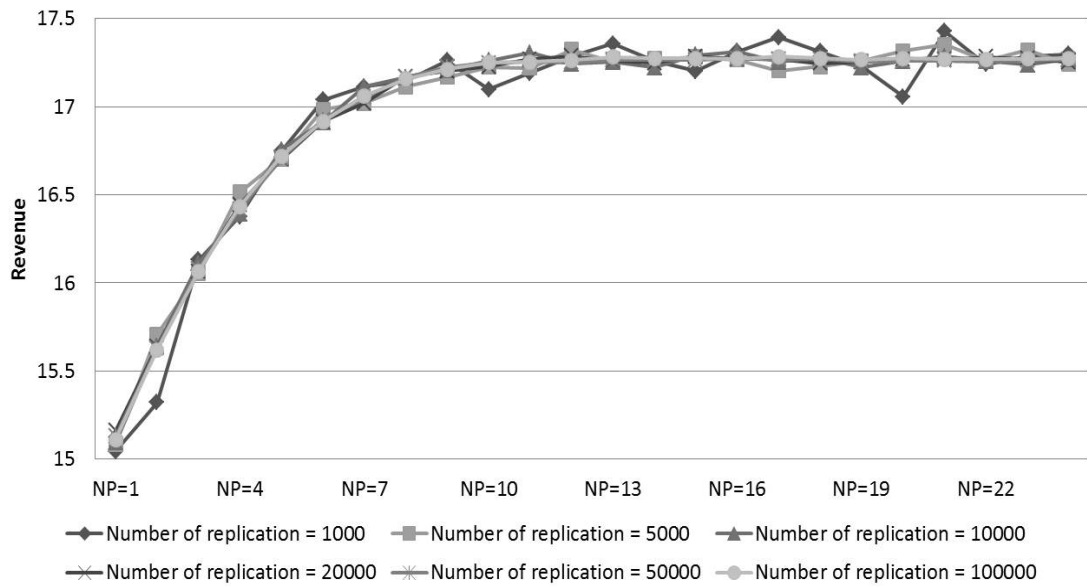
As we know, a simulation model is usually accompanied with simulation errors. In other words, simulations driven by random inputs will produce random output. There are several ways to reduce the variance of output. In this dissertation, we use two different methods to run simulation to reduce the variances. These two methods are described in following two sections.

#### **B.1 Simulations based on large-scaled random numbers**

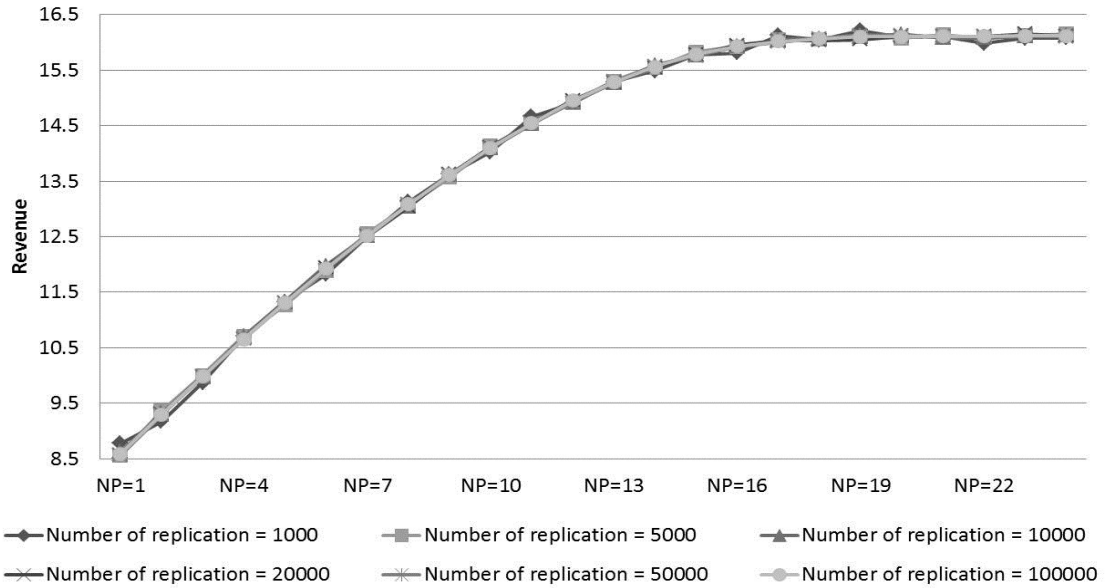
The most common method to reduce the variance of output is to run large-scaled simulations. Then question is, how many replications would be sufficient to reduce the variance of output at an acceptable level? We test single physician practices with P/S ratio (same as described in chapter 4, P/S ratio is the ratio of prescheduled demand mean and same-day demand mean) to be 4/20, 8/16, 16/8, and 20/4. We set the number of replication to be 1000, 5000, 10000, 20000, 50000, and 100000. The comparisons of expected revenue under different number of replications are summarized in Figure B.1 - Figure B.4.



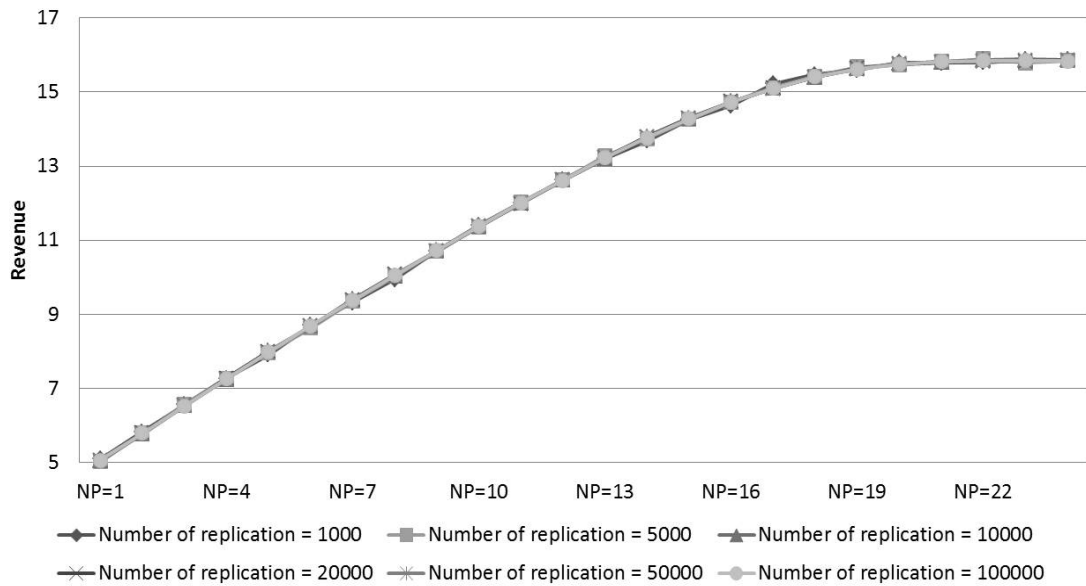
**Figure B.1.** Revenue vs. Number of replications : single physician 4/20, 100% workload



**Figure B.2.** Revenue vs. Number of replications : single physician 8/16, 100% workload



**Figure B.3.** Revenue vs. Number of replications : single physician 16/8, 100% workload



**Figure B.4.** Revenue vs. Number of replications : single physician 20/4, 100% workload

From the study based on aggregate model, we expect to observe a monotonously increasing then monotonously decreasing curve for a single practice when we increase  $N^p$  from 1 to 24, if there is no simulation errors. Due to the disturbance from

simulation errors, we might observe some up and down curves, when the number of replication is not sufficiently large.

From Figure B.1 - Figure B.4, we observe that, under any demand ratios, when we increase the number of replications, the variance of expected revenue becomes smaller, resulting a much smoother curve. However, the output is converged with different speed under different demand ratio. For example, when P/S ratio is 4/20, we still observe unstable curves under 10000 replications, while the results are perfectly converged under 10000 replications when P/S ratios are 16/8 and 20/4. This is because when we have very few prescheduled patients, the associated revenue change with increasing  $N^P$  by one unit is not large enough, comparing to the simulation error. In that case, the curve is not stable under the demand ratio 4/20.

Generally speaking, under demand ratios of 16/8 and 20/4, 10000 replications are sufficiently large while under demand ratios of 4/20 and 8/16, a suitable number of replications should be no less than 20000 to obtain stable simulation results. For our simulation, given a fixed combination of values for all input variables, 10000 replications take about 3 minutes for one single physician practice test and take about 10 minutes for multiple physician practice test.

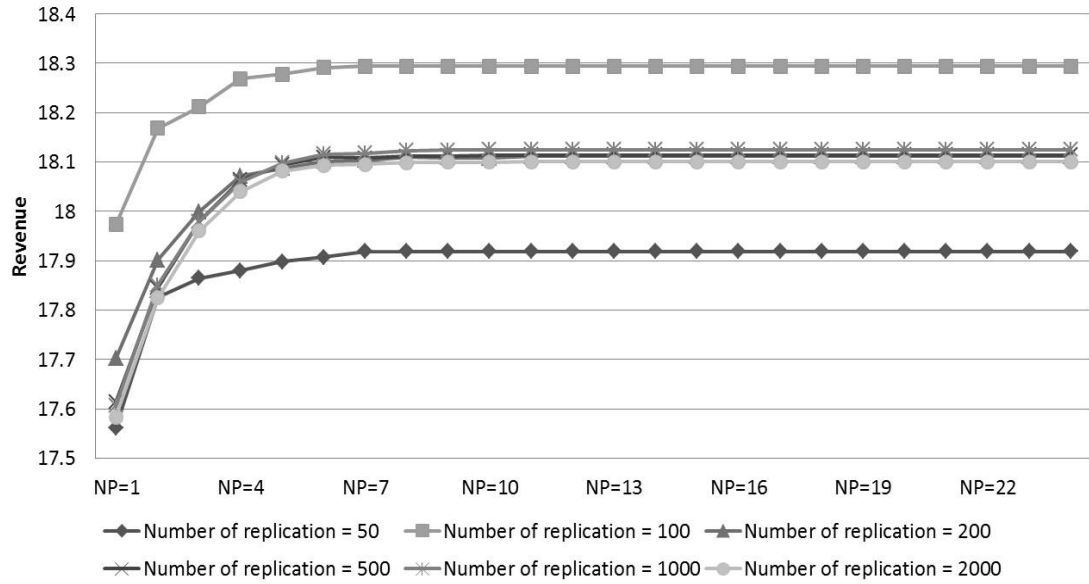
## **B.2 Simulations based on common random numbers**

Large-scaled simulations could definitely reduce the variance of output, however, simulations based on large number of replications usually require great amounts of computer time and storage, appropriate statistical analysis. If appropriate variance-reduction technique could be used, much less amount of replications would be sufficient to run simulations. In our dissertation, the second simulation method is common random numbers, probably the most useful and popular variance-reduction technique of all.

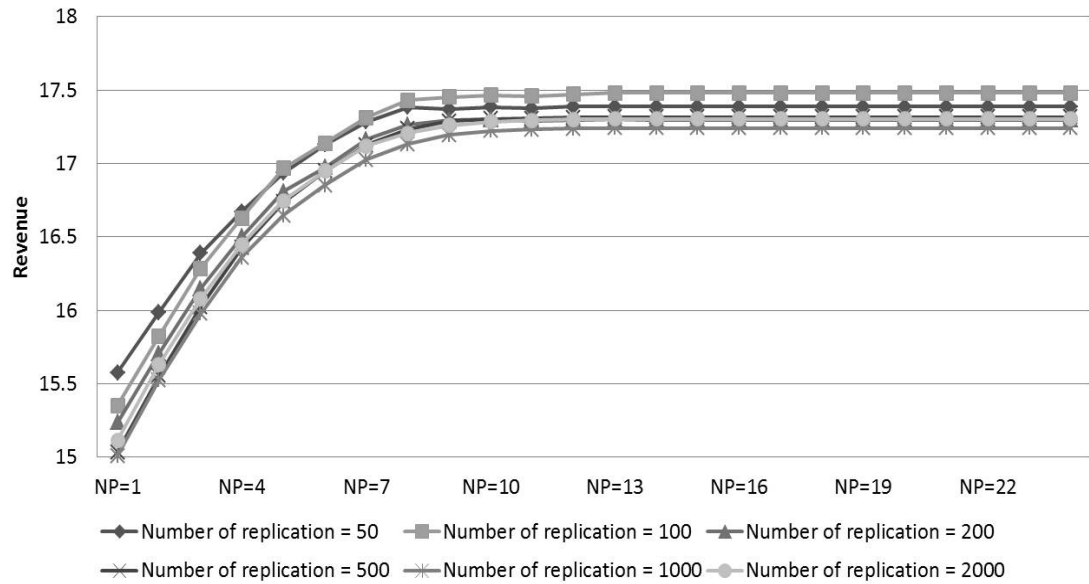
The basic idea of common random numbers is to compare the alternative configurations under similar experimental conditions. In that case, we know that any observed differences in performance are due to the configuration differences rather than to fluctuations of the experimental conditions - actually the variance of input.

Similar as shown in Appendix B.1, we also test single physician practices with P/S ratios to be 4/20, 8/16, 16/8, and 20/4, based on common random numbers. For each value of  $N^p$  under same demand ratio, we use same seeds to run simulations, that is, the amount of demands and the arrived time of each request are all fixed under same demand ratio. Keeping the generated random inputs same to run simulations could significantly reduce the variance of output, even with a small amount of replications. We test six different number of replications under this method - 50, 100, 200, 500, 1000, and 2000, and compare the expected revenue vs.  $N^p$  under different number of replications in .

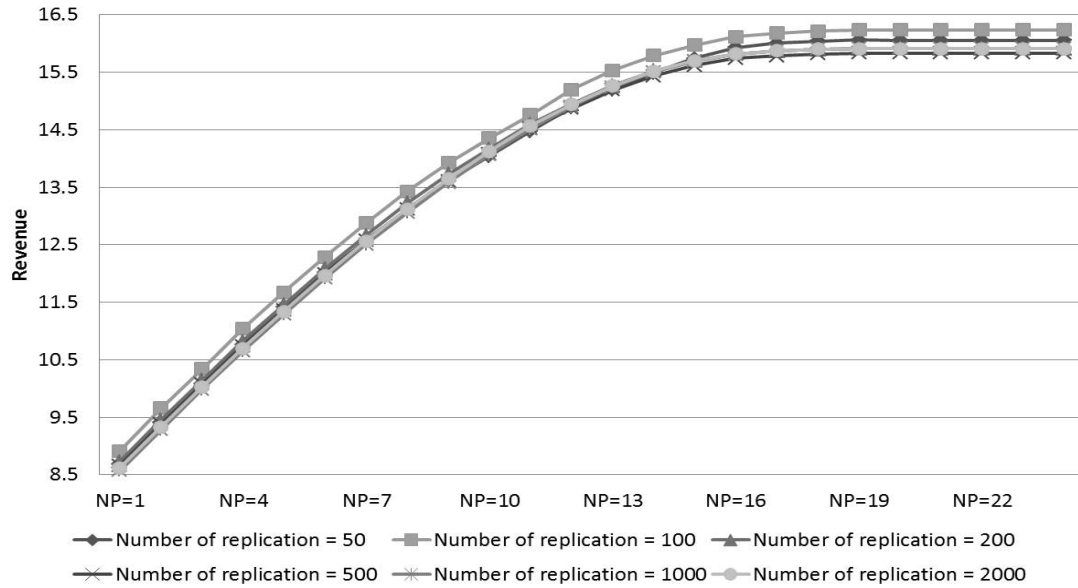
From Figure B.5 - Figure B.8, we observe that, under any demand ratios, the output of expected revenue are converged well for any given  $N^p$  when the number of replications is larger than 1000. Meanwhile, we could always observe a smooth curve even under a small amount of replications, like 100. In other words, the variance of output are controlled well even under a small amount of replications, when we run simulations under same seeds.



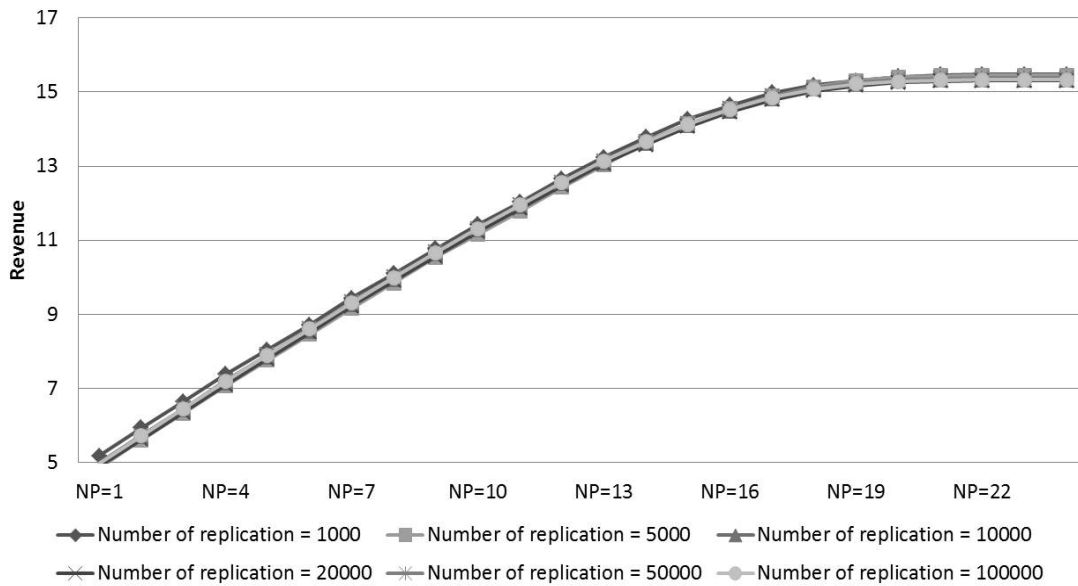
**Figure B.5.** Revenue vs. Number of replications : single physician 4/20, 100% workload under same seed



**Figure B.6.** Revenue vs. Number of replications : single physician 8/16, 100% workload under same seed



**Figure B.7.** Revenue vs. Number of replications : single physician 16/8, 100% workload under same seed



**Figure B.8.** Revenue vs. Number of replications : single physician 20/4, 100% workload under same seed

Generally speaking, if we run simulations under same seeds, 1000 replications are sufficient large to obtain a good estimate of the output of simulations, which could help to reduce computer time a lot.

## BIBLIOGRAPHY

- [1] Vermont department of health: 2002 physician survey: statistical report [internet]. *Montpelier (VT): Vermont Department of Health. Available from: <http://healthvermont.gov/pubs/physis/phys02bk.pdf>* (March 2005).
- [2] Texas medical association: Survey of texas physicians research findings. *Austin (TX): Texas Medical Association* (2006).
- [3] Vermont department of health: 2006 physician survey: statistical report [internet]. *Montpelier (VT): Vermont Department of Health, Available from: <http://healthvermont.gov/research/documents/phys06bk.PDF>* (2007).
- [4] Massachusetts medical society: 2008 physician workforce study. *Waltham (MA): Massachusetts Medical Society* (2008).
- [5] Aksin, OZ., and Karaesmen, F. Characterizing the performance of process flexibility structures. *Operations Research Letters* 35, 4 (July 2007), 477–484.
- [6] Angalakudati, M., Balwaniy, S., Calzada, J., Chatterjee, B., Perakisz, G., Raad, N., and Uichanco, J. Business analytics for flexible resource allocation under random emergencies. *Management Science* 60, 6 (June 2014), 1552–1573.
- [7] Ayvaz, N., and Huh, WT. Allocation of hospital capacity to multiple types of patients. *Journal of Revenue and Pricing Management* 9 (September 2010), 386–398.
- [8] Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., Wood, D., and Stahl, J. Improving clinical access and continuity using physician panel redesign. *Journal of General Internal Medicine* 25, 10 (2010), 1109–1115.
- [9] Balasubramanian, H., Biehl, S., Dai, L., and Muriel, A. Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. *Health Care Management Science* 17, 1 (July 2013), 31–48.
- [10] Balasubramanian, H., Muriel, A., and Wang, L. The impact of provider flexibility and capacity allocation on the performance of primary care practices. *Flexible Services and Manufacturing Journal* 10696, 011 (2011), 9112–5.
- [11] Bassamboo, A., Randhawa, RS., and Van Mieghem, JA. Optimal flexibility configurations in newsvendor networks: Going beyond chaining and pairing. *Management Science* 56, 8 (August 2010), 1285–1303.

- [12] Bennett, J., and Baxley, G. The effect of a carve-out advanced access scheduling system on no-show rates. *Family Medicine* 41, 1 (2009), 51–56.
- [13] Carr, S., and Duenyas, I. Optimal admission control and sequencing in a make-to-stock/make-to-order production system. *Operations Research* 48, 5 (September-October 2000), 709–720.
- [14] Chakraborty, S., Muthuraman, K., and Lawley, M. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions* 42, 5 (February 2010), 354–366.
- [15] Chou, M., Teo, C-P., and Zheng, H. Process flexibility revisited: The graph expander and its applications. *Operations Research* 59, 5 (September 2011), 1090–1105.
- [16] Chou, M.C., Chua, G.A., Teo, C-P., and Zheng, H. Design for process flexibility: Efficiency of the long chain and sparse structure. *Operations Research* 58, 1 (January 2010), 43–58.
- [17] Chou, W.C., Cooney, L.M. Jr., Van, N. P.H., Allore, H.G., and Gill, T.M. Access to primary care for medicare beneficiaries. *Journal of the American Geriatrics Society* 55, 5 (May 2007), 763–768.
- [18] Chua, G.A., Chou, M.C., and Teo, C.P. On range and response: Dimensions of process flexibility. *European Journal of Operational Research* 207, 2 (December 2010), 711–724.
- [19] Dobson, G., Hasija, S., and Pinker, J. Reserving capacity for urgent patients in primary care. *Production and operations management* 20, 3 (May-June 2011), 456–473.
- [20] Feldman, J., Liu, N., Topaloglu, H., and Ziya, S. Appointment scheduling under patient preference and no-show behavior. *Operations Research* 62, 4 (July-August 2014), 794–811.
- [21] Gerchak, Y., Gupta, D., and Henig, M. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science* 42, 3 (March 1996), 321–334.
- [22] Gill, J.M., and Mainous, A. The role of provider continuity in preventing hospitalizations. *Archives of Family Medicine* 7, 4 (July 1999), 352–357.
- [23] Gill, J.M., Mainous, A., and Nsereko, M. The effect of continuity of care on emergency department use. *Archives of Family Medicine* 9, 4 (April 2000), 333–338.
- [24] Graves, S.C., and Tomlin, B.T. Process flexibility in supply chains. *Management Science* 49, 7 (July 2003), 907–919.

- [25] Green, L. V., Savin, S., and Murray, M. Providing timely access to care: What is the right patient panel size? *Joint Commission Journal on Quality and Patient Safety* 33, 4 (April 2007), 211–218.
- [26] Green, L.V., and Savin, S. Reducing delays for medical appointments: A queueing approach. *Operations Research* 56, 6 (December 2008), 1526–1538.
- [27] Gupta, D., Potthoff, S., Blowers, D., and Corlett, J. Performance metrics for advanced access. *Journal of Healthcare Management* 51, 4 (July-August 2006), 246–259.
- [28] Gupta, D., and Wang, L. Capacity management for contract manufacturing. *Operations Research* 55, 2 (March-April 2007), 367–377.
- [29] Gupta, D., and Wang, L. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research* 56, 3 (May-June 2008), 576–592.
- [30] Gurumurthi, S., and Benjaafar, S. Modeling and analysis of flexible queueing systems. *Naval Research Logistics* 51, 5 (June 2004), 755–782.
- [31] Hippchen, J. Flexibility in primary care. *Masters Thesis, Accessible at: <http://people.umass.edu/hbalasub/FlexibilityThesis.pdf>* (2009).
- [32] Hopp, W., Tekin, E., and Van Oytten, M.P. Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science* 50, 1 (January 2004), 83–98.
- [33] Huh, WT., Liu, N., and Truong, VA. Multi-resource allocation scheduling in dynamic environments. *Manufacturing and Service Operations Management* 15, 2 (Spring 2013), 280–291.
- [34] Jordan, W.C., and Graves, S.C. Principles and benefits of manufacturing process flexibility. *Management Science* 41, 4 (April 1995), 577–594.
- [35] Kopach, R., DeLaurentis, P., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Wan, H., Intrevado, P., Qu, X., and Willis, D. Effects of clinical characteristics on successful open access scheduling. *Health Care Manage Science* 10, 2 (March 2007), 111–124.
- [36] LaGanga, L., and Lawrence, S. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences* 38, 2 (May 2007), 251–276.
- [37] Liu, N. Optimal choice for appointment scheduling window under patient no-show behavior. *Under revision. Notes: Third place in the 2013 INFORMS Junior Faculty Interest Group (JFIG) Paper Competition.*
- [38] Liu, N., and Ziya, S. Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management Article first published online* (March 2014).

- [39] Liu, N., Ziya, S., and Kulkarni, V. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing and Services Operations Management* 12, 2 (Spring 2010), 347–365.
- [40] Modarres, M., and Sharifyazdi, M. Revenue management approach to stochastic capacity allocation problem. *European Journal of Operational Research* 192, 2 (January 2009), 442–459.
- [41] Muriel, A., Somasundaram, A., and Zhang, Y. Impact of partial manufacturing flexibility on production variability. *Manufacturing and Service Operations Management* 8, 2 (Spring 2006), 192–205.
- [42] Murray, M., and Berwick, D. M. Advanced access: Reducing waiting and delays in primary care. *Journal of the American Medical Association* 289, 8 (February 2003), 1035–1040.
- [43] Murray, M., Bodenheimer, T., Rittenhouse, D., and Grumbach, K. Improving timely access to primary care: Case studies of the advanced access model. *Journal of the American Medical Association* 289, 3 (February 2003), 1042–1046.
- [44] Muthuraman, K., and Lawley, M. Stochastic overbooking model for outpatient clinical scheduling with no shows. *IIE Transactions* 40, 9 (July 2008), 820–837.
- [45] Nelson, J., Banning, T., Kroll, C., and Bailey, C.J. Fractured: the state of health care in texas. *Austin (TX): Primary Care Coalition* (2006).
- [46] Oh, H-J., Muriel, A., Balasubramaniana, H., Atkinson, K., and Ptaszkiewicz, T. 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 (December 2013), 263–279.
- [47] Ozen, A., and Balasubramanian, H. The impact of case mix on timely access to appointments in a primary care group practice. *Health care management science* 16, 2 (June 2013), 101–118.
- [48] Qu, X., Rardin, R., Williams, J.A.S., and Willis, D. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research* 183, 2 (December 2007), 812–826.
- [49] Robinson, L., and Chen, R. A comparison of traditional and open access policies for appointment scheduling. *Manufacturing and Services Operations Management* 12, 2 (Spring 2010), 330–347.
- [50] Sethi, A.K., and Sethi, S.P. Flexibility in manufacturing : A survey. *The International Journal of Flexible Manufacturing Systems* 2, 4 (July 1990), 289–328.
- [51] Sheikhzadeh, M., Benjaafar, S., and Gupta, D. Machine sharing in manufacturing systems: Total flexibility versus chaining. *International Journal of Flexible Manufacturing Systems* 10, 4 (October 1998), 351–378.

- [52] Simchi-levi, D., and Wei, Y. Understanding the performance of the long chain and sparse designs in process flexibility. *Operations Research* 60, 5 (September-October 2012), 1125–1141.
- [53] Wang, W., and Gupta, D. Adaptive appointment systems with patient preferences. *Manufacturing and Service Operations Management* 13, 3 (Summer 2011), 373–389.