University of Massachusetts Amherst

The Impact of Flexibility And Capacity Allocation On The Performance of Primary Care Practices

Item Type	thesis
Authors	Wang, Liang
DOI	10.7275/1483325
Download date	2025-04-20 02:21:20
Link to Item	https://hdl.handle.net/20.500.14394/47374

THE IMPACT OF FLEXIBILITY AND CAPACITY ALLOCATION ON THE PERFORMANCE OF PRIMARY CARE PRACTICES

A Master's Thesis Presented

by

LIANG WANG

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

MASTER OF SCIENCE IN INDUSTRIAL ENGINEERING AND OPERATION RESEARCH

September 2010

Department of Mechanical and Industrial Engineering

THE IMPACT OF FLEXIBILITY AND CAPACITY ALLOCATION ON THE PERFORMANCE OF PRIMARY CARE PRACTICES

A Master's Thesis Presented

by

LIANG WANG

Approved as to style and content by:

_____

Hari Balasubramanian, Co-chair

_____

Ana Muriel, Co-chair

_____

Senay Solak, Member

_____

Donald L. Fisher, Department Head
Department of Mechanical and Industrial Engineering

ACKNOWLEDGEMENTS

I sincerely thank my advisors, Dr. Hari J. Balasubramanian and Dr. Ana Muriel, for their countless support throughout my study and numerously insightful discussions during my work on this thesis. I also thank Dr. Senay Solak for the comments and time he devoted to improve the quality of my work.

I want to thank my parents, without their continuous support and unlimited love, this work would not have been possible.

I also want to give my thanks to my girlfriend, Linli Zhang, who is graduated from University of Massachusetts Amherst and now is attending the MBA program in University of Texas Austin. For many years, she devotes her love to encourage me to overcome the challenges impeding in the way and strive for the excellence, I am deeply grateful for all she did for me.

ABSTRACT

THE IMPACT OF FLEXIBILITY AND CAPACITY ALLOCATION ON THE PERFORMANCE OF PRIMARY CARE PRACTICES

SEPTEMBER 2010

LIANG WANG

B.S., HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Dr. Hari Balasubramanian and Professor Dr. Ana Muriel

The two important metrics for any primary care practice are: (1) *Timely Access* and (2) *Patient-physician Continuity*. Timely access focuses on the ability of a patient to get access to a physician as soon as possible. Patient-physician continuity refers to building a strong or permanent relationship between a patient and a specific physician by maximizing patient visits to that physician. In the past decade, a new paradigm called advanced access or open access has been adopted by practices nationwide to encourage physician to "do today's work today." However, most clinics still reserve pre-scheduled appointments for long lead-time appointments due to patient preference and clinical necessities. Therefore, an important problem for clinics is how to optimally manage and allocate limited physician capacities as much as possible to meet the two types of demand – pre-scheduled (non-urgent) and open access (urgent) – while simultaneously maximizing timely access and patient-physician continuity. In this study we use a quantitative approach to apply the ideas of manufacturing process flexibility to capacity management in a primary care practice. We develop a closed form expression for

capacity allocation for an individual physician and a two physician practice. In the case of multiple physicians, we use a two-stage stochastic integer programming approach to investigate the value of flexibility under different levels of flexibility and provide the optimal capacity allocation solution for each physician. We find that flexibility has the greatest benefit when system utilization is balanced and when the individual physicians have unequal utilizations. The benefits of flexibility also increase as the practice gets larger.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iii

ABSTRACT iv

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER

1 INTRODUCTION 1

1.1 Background on primary care 1

1.2 Current primary care practices 3

1.3 Team care and physician flexibility 5

1.4 Capacity allocation between pre-scheduling and open access 6

2 LITERATURE REVIEW 8

2.1 Quantitative models for primary care practice 8

2.2 Research related to flexibility 9

3 MODELING APPROACH 12

3.1 Assumptions 12

3.2 Model formulation 13

3.2.1 Formulation for dedicated flexibility 14

3.2.2 Formulation for two physicians with full flexibility 15

3.2.3 Formulation for general configuration 18

4 VALUE OF FLEXIBILITY 21

4.1 Practice without any flexibility 21

4.2 Two physicians with open access flexibility 24

4.3 Value of flexibility in a practice with three physicians 25

4.3.1 Results for three physicians with symmetric demand distributions 27

4.3.2 N_i^{p*} of three physicians with symmetric demand distributions 40

4.3.3 Results for three physicians with asymmetric demand distributions 47

4.3.4 N_i^{p*} of three physicians with asymmetric demand distributions 53

4.3.5 Trends in the total N_i^{p*} values for all three physicians 56

4.4 Value of flexibility in a practice with six physicians 58

4.4.1 Results for six physicians with symmetric demand distributions 58

4.4.2 Results for six physicians with asymmetric demand distributions 61

4.5 Conclusion 63

5 IMPLICATIONS FOR PRACTICE 65

6 CONCLUSIONS 68

7 FUTURE WORK 70

APPENDICES

A. THEOREMS PROOF 72

B. PROGRAMS FOR THE STUDY OF FLEXIBILITY 77

BIBLIOGRAPHY 88

LIST OF TABLES

Table **Page**

4.1 Assumptions for 3 physicians with symmetric demand distributions in Symmetric Case 1 (10/14). 28

4.2 Statistics of objective value for different flexibilities with 100% utilization in Symmetric Case 1. 29

4.3 Statistics of total demands for 3 physicians with 100% utilization in Symmetric Case 1. 30

4.4 Measurement for different flexibilities in term of system revenue in Symmetric Case 1 (10/14). 31

4.5 Measurement for different flexibilities in term of timely access rate in Symmetric Case 1 (10/14). 31

4.6 Measurement for different flexibilities in term of continuity rate in Symmetric Case 1 (10/14). 31

4.7 Measurement for different flexibilities in term of system revenue in Symmetric Case 2 (14/10). 36

4.8 Measurement for different flexibilities in term of timely access rate in Symmetric Case 2 (14/10). 36

4.9 Measurement for different flexibilities in term of continuity rate in Symmetric Case 2 (14/10). 36

4.10 Measurement for different flexibilities in term of system revenue in Symmetric Case 3 (6/18). 37

4.11 Measurement for different flexibilities in term of timely access rate in Symmetric Case 3 (6/18). 37

4.12 Measurement for different flexibilities in term of continuity rate in Symmetric Case 3 (6/18). 37

4.13 Measurement for different flexibilities in term of system revenue in Symmetric Case 4 (18/6). 38

4.14 Measurement for different flexibilities in term of timely access rate in Symmetric Case 4 (18/6). 38

4.15 Measurement for different flexibilities in term of continuity rate in Symmetric Case 4 (18/6). 38

4.16 Assumptions for 3 physicians with asymmetric demand distributions in Asymmetric Case 1. 47

4.17 Measurements of system revenue with asymmetric demands in Asymmetric Case 1. 48

4.18 Measurements of timely access rate with asymmetric demands in Asymmetric Case 1. 48

4.19 Measurements of continuity rate with asymmetric demands in Asymmetric Case 1. 48

4.20 Assumptions for 3 physicians with asymmetric demand distributions in Asymmetric Case 2. 49

4.21 Measurements of system revenue with asymmetric demands in Asymmetric Case 2. 50

4.22 Measurements of timely access rate with asymmetric demands in Asymmetric Case 2. 50

4.23 Measurements of continuity rate with asymmetric demands in Asymmetric Case 2. 50

4.24 Assumptions for 6 physicians with symmetric demand distributions. 58

4.25 Measurement of system revenue for 6 physicians (symmetric). 59

4.26 Measurement of timely access rate for 6 physicians (symmetric). 59

4.27 Measurement of continuity rate for 6 physicians (symmetric). 59

4.28 Assumptions for 6 physicians with asymmetric demand distributions. 61

4.29 Measurement of system revenue for 6 physicians (asymmetric). 62

4.30 Measurement of timely access rate for 6 physicians (asymmetric). 62

4.31 Measurement of continuity rate for 6 physicians (asymmetric). 62

LIST OF FIGURES

Figure **Page**

3.1 System configuration for dedicated flexibility. 14

3.2 System configuration for two physicians sharing open access demands..... 16

3.3 System configuration for partial and full flexibility. 18

4.1 Dedicated case with demand rates 10 and 14 for pre-scheduling and open access respectively. And a closer view of the value near the optimal point. 22

4.2 Dedicated case with demand rates 16 and 8 for pre-scheduling and open access respectively. And a closer view of the value near the optimal point. 23

4.3. Two physicians have flexibility in open access practice where N1p = 19 and N2p = 14..... 24

4.4. Two physicians have flexibility in open access practice where N1p = 16, N2p = 16. 25

4.5 Box-Whisker Plot comparison of objective values for different flexibilities with 100% utilization in Symmetric Case 1. 29

4.6 Comparisons of different flexibilities in term of system revenue in Symmetric Case 1 (10/14). 32

4.7 Comparisons of different flexibilities in term of timely access rate in Symmetric Case 1 (10/14). 32

4.8 Comparisons of different flexibilities in term of continuity rate in Symmetric Case 1 (10/14). 33

4.9 An example of diversion process in 2-chain and full flexibility..... 35

4.10 2-chain flexibility improvement under different demand ratios for all symmetric cases..... 39

4.11 Full flexibility improvement under different demand ratios for all symmetric cases..... 39

4.12 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 40% utilized in Symmetric Case 3 (6/18)..... 41

4.13 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 80% utilized in Symmetric Case 3 (6/18)..... 42

4.14 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 100% utilized in Symmetric Case 3 (6/18)..... 42

4.15 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 120% utilized in Symmetric Case 3 (6/18)..... 43

4.16 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 160% utilized and in Symmetric Case 3 (6/18).. 44

4.17 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 40% utilized in Symmetric Case 2 (14/10)..... 44

4.18 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 80% utilized in Symmetric Case 2 (14/10)..... 45

4.19 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 100% utilized in Symmetric Case 2 (14/10)..... 45

4.20 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 120% utilized in Symmetric Case 2 (14/10)..... 46

4.21 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 160% utilized in Symmetric Case 2 (14/10)..... 46

4.22 System revenue comparison between Asymmetric Case 1 and 2
for flexible configurations..... 51

4.23 Timely access comparison between Asymmetric Case 1 and 2
for flexible configurations..... 52

4.24 Continuity comparison between Asymmetric Case 1 and 2
for flexible configurations..... 52

4.25 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 40% utilized in Asymmetric Case 2. 53

4.26 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 80% utilized in Asymmetric Case 2. 54

4.27 Distributions of the differences of Nps between flexible configurations
and dedicated case when the system is 100% utilized in Asymmetric Case 2. 54

4.28 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 120% utilized in Asymmetric Case 2. 55

4.29 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 160% utilized in Asymmetric Case 2. 55

4.30 Average Nps values for three physicians with asymmetric demand distributions.... 57

4.31 Comparison of system revenue improvement between 3 and 6 physicians..... 60

4.32 Comparison of timely access improvement between 3 and 6 physicians. 60

4.33 Comparison of continuity improvement between 3 and 6 physicians. 61

CHAPTER 1

INTRODUCTION

The US healthcare system, by all accounts, is in a state of crisis and cannot be alleviated without fundamental change and reform. With expenditures of about $2.2 trillion, or 16.2% of the GDP [1], the US healthcare system ranks the second among the members of the World Health Organization (WHO) and ranks at the top among industrialized countries [2]. This expenditure is expected to increase continuously to around 20% of the GDP in less than a decade [1, 3]. One might think that, given this immense spending, health outcomes would improve correspondingly. However, the current situation is that about 40-50 million Americans lack health insurance. Most of them believe the insurance is too expensive to afford. The WHO ranks the US as 37th in overall system performance and 72nd among the 192 member states in terms of overall level of health [4].

A solution to the current crisis in healthcare requires a multi-pronged effort involving multiple aspects of the healthcare system. Healthcare policy makers agree that one of the key areas that needs to be addressed is primary care. The World Health Report 2008 [5] is appropriately titled "Primary Health Care Now More Than Ever."

1.1 Background on primary care

Primary care providers (PCP) form the backbone of most modern health care systems and are typically the first point of contact between patients and systems. They manage a patient's general health issues and provide preventive medicine, patient education and routine physical exams, In addition, they review a patient's medical history and take care

of referrals to medical specialists for secondary and tertiary care. 94% of patients value their PCP as a “source of first contact care” and approximately 90% are satisfied with their coordinated referrals [6]. The important benefits of an effective primary care system are well documented in the clinical literature. For instance, Starfield, Shi and Macinko (2005), among others [7, 8, 9], show that improving primary care generates several promising results:

- Improves access to health services for relatively deprived population groups.
- Assist in the prevention and early management of health problems due to education and early detection.
- Builds stronger relationships between patients and their PCP and reduces the amount of wasteful expenditures by minimizing inappropriate referrals to secondary and tertiary care providers.

Despite its pivotal role in the overall system, primary care is “at grave risk due to a dysfunctional financing and delivery system” [7]. A study by the American College of Physicians (2006) points out the current dilemma faced by the primary care: the demand for healthcare grows steadily and dramatically with an estimated growth rate of 38% from 2000 to 2020, yet the number of students specializing in primary care keeps declining due to lower salaries combined with higher workloads [7, 10]. This imbalanced situation involving increasing demand and shortage of supply leads to worse quality of care, longer waiting times, and increased dissatisfactions, all of which aggravate the crisis in the healthcare system.

To improve primary care practices and overcome the problems that are impeding the healthcare system from performing optimally, two important metrics are introduced: (1)

Timely Access and (2) *Patient-physician Continuity*. These are two of the six recommended aims by the Institute of Medicine (2001) [11].

Timely Access focuses on the ability of a patient to get access to care as soon as possible. Not getting timely appointments lowers patient satisfaction and increases the likelihood of sending the patients to the Emergency Room (ER) more frequently [12, 13]. The inability to get a timely appointment especially hinders the appropriate management of chronic diseases that could have been effectively treated in a primary care practice.

Patient-physician continuity refers to building a strong or permanent relationship between a patient and a specific physician so that the patient can see his/her own PCP as much as possible. Continuity is considered as one of the hallmarks of primary care. Gill and Mainous (1998) point to several studies which show that patients who regularly see their own PCP are (1) more satisfied with their care; (2) more likely to take medications correctly; (3) more likely to have problems correctly detected; and (4) less likely to be hospitalized [14]. Continuity is more important for patients with a complex medical history and chronic problems since they can be treated more appropriately by their own physicians who are familiar with their conditions. From the physician's perspective, continuity is also beneficial since workloads are more focused.

1.2 Current primary care practices

Various types of primary care practices currently exist in the U.S., for example, those consisting of family physicians, general internists and pediatricians. Though many of them are conducted by one single physician, more than 65% of primary care practices are group practices consisting of more than one physician [15]. To establish the connection

between patient and physician, each physician has a *panel*, which is the set of patients he/she is responsible for. The physician takes appointments from his/her respective panel and only treat patients from other panels in exceptional cases. Physician appointments are usually scheduled into 15- or 20-minute slots. Reimbursement to physicians in primary care is based largely on 20-minute visits, and a full-time physician typically has 24 appointments in a working day based on eight hours.

Broadly speaking, appointments for primary care can be classified into two types: (1) Non-urgent or pre-scheduled appointments and (2) Urgent or acute appointments. Non-urgent appointments come from patients with chronic conditions who need regular treatments, and patients requiring annual exams or a first time assessment. Urgent appointments are demands that come in on a daily basis from patients requiring immediate attention from their PCPs. If their own physicians are unavailable at the walk-in time, patients have to get their care at an emergency room.

In traditional practices of appointment scheduling, urgent appointments received higher priority and were scheduled as soon as possible, while non-urgent requests were usually postponed up to several weeks or even months. To address the issue of long backlogs and intolerant waiting times, a new paradigm called *advanced access* or *open access* has been adopted by practices nationwide [16]. Under open access, all patients, regardless of urgent or non-urgent status, are given same-day appointments with their own physician who are encouraged to “do today’s work today.” The key of a successful implementation of open access is to balance the demand and supply appropriately, which means panels should be sized properly and physicians might work overtime occasionally

[17]. In common practice, open access schemes are implemented simultaneously with traditional pre-scheduling methods in most clinics.

1.3 Team care and physician flexibility

Another approach to overcome deficiencies in primary care practices is to allow the concept of *team care* to play a central role to improve quality of care, something which is recommended by the Institute of Medicine (2001) in its report *Crossing the Quality Chasm: New Health System for the Twenty First Century* [18]. Team care brings with it the idea of *physician flexibility*, which implies that patients will not only be seen by their dedicated physician, but also by support staff or other physicians in the team. This actually happens routinely in practice without any "installation" or "special configuration". While, the flexibility of allowing a physician to see patients from any of other physician panels might improve timely access, physician flexibility can be detrimental to continuity and increase the chances of misdiagnosis. One question that arises naturally is: what is the maximum level of flexibility that will still provide an acceptable level of continuity given two different demand streams? The levels of flexibility that will be compared and investigated in this thesis are shown in Figure 1.1.

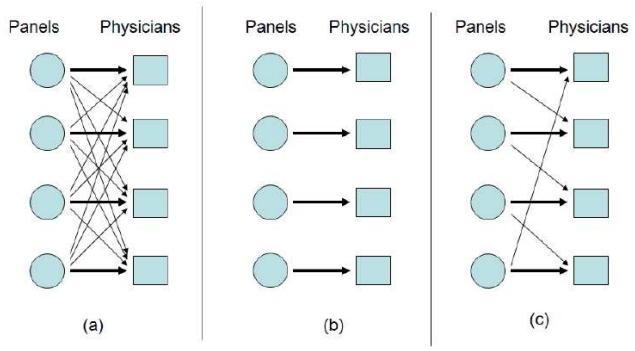


Figure 1.1 Different flexibility configurations that tradeoff continuity with timeliness.

In (a), patients may see any other physician (full flexibility). This configuration leads to the highest level of timely access as resources are pooled, but may not ensure continuity. In (b), patients can only see their own dedicated physician (no flexibility), which leads to the highest level of continuity, although timely access might not be guaranteed. Combing these two levels leads to configuration (c) partial flexibility, where patients and physicians are *chained* such that each patient in addition to having his/her own physician, also has one *auxiliary physician* (AP).

Having laid out the main issues, below we examine in more detail how the inherent flexibility of primary care physicians can be best managed, at different levels of the planning hierarchy, to improve timely access and continuity.

1.4 Capacity allocation between pre-scheduling and open access

Though open access has been successfully implemented and adopted in primary care practices, most clinics still reserve pre-scheduled appointments for long lead-time appointments due to patient preference and clinical necessities. Therefore, the most urgent problem becomes finding how to optimally manage and allocate limited physician capacities as much as possible to meet the two types of demand—pre-scheduling and open access. Qu and Shi (2009) proposed a two-level physician capacities management scheme which combines the high level total capacity of the clinic and low level capacity of individual physician care to find the optimal capacity allocation method for current open access clinics with one physician, or a physician team that has capacities pooled [19]. We will use an alternative approach to find the best allocation scheme for multiple

physicians and investigate the optimal allocation method for primary care practices with different levels of flexibility.

CHAPTER 2

LITERATURE REVIEW

2.1 Quantitative models for primary care practice

The application of optimization approaches to primary care is limited, yet growing. With the advent of *advanced access* proposed by Murray and Tantau [20, 21], research focusing on capacity planning and allocating in primary care is booming. For instance, Green et al. (2007) [17] develop a simple probability model to investigate the number of overtime appointments that a physician could be expected to engage as a function of his/her panel size. To offset the effect of variability, they conclude that physician capacity should be sufficiently higher than patient demand. Using a queuing model, Green and Savin (2008) determine the effect of no-shows on a physician's panel size. This queuing model demonstrates an ability to estimate the relationship between a physician's backlog and his/her panel size, as well as patient no-show rates.

Qu et al (2007) [22] develop an expression for the optimal number of slots that should be reserved for pre-scheduled appointments in a day for a single physician practice. They find the optimal solution depends on the no-show rates of pre-scheduled demand and open access demand, as well as the distribution of open access demand. In chapter 3, we provide a simpler approach for the same quantity, which in turn leads to more complex and yet unexplored two physician practices. Kopach et al (2007) [23] use discrete event simulation in an open access scheduling environment to analyze the effects of clinical characteristics on continuity of care and clinic performance. One primary conclusion relevant to this research is that continuity of care is adversely affected as the fraction of

patients on open access increases. They also propose that physician team practice would be the solution to the problem.

Gupta and Wang (2008) [24] develop a model to establish appointment booking policies that can maximize a clinic's revenue. They use a Markov decision process (MDP) that explicitly accounts for patient preferences with respect to specific appointment times and multiple physicians, and also for different types of demand: pre-scheduling and open access. The main differences between their research and ours are: 1) In their approach, the booking of pre-scheduled appointments is driven by patient preference; by contrast, we try to balance pre-scheduled demand and same-day demand. 2) The same-day demand in their model arrives before the beginning of the day and can be treated as deterministic information, while we focus on more dynamic behavior and provide optimal bound for the patient flow management.

2.2 Research related to flexibility

Lots of research investigating the benefits of flexibility has focused mainly on the manufacturing, but more recently has extended to include the service system and worker training and allocation. Jordan and Graves (1995) [25] have studied the improvements arising from using a flexibility configuration in sales and capacity utilization in multi-product and plant networks. They were the first to compare the benefits of partial flexibility to full flexibility in the field of assembly lines, and they concluded that partial flexibility (chaining), delivers almost the same benefits of a fully flexible system, yet needs only a small fraction of links and costs. Graves and Tomlin (2003) [26] extend this research to multi-stage supply chains and to a make-to-order environment where

flexibility is also used to hedge against variability (Muriel et al. (2006) [27]). Brusco and Johns (1998) [28] find that the benefits of partial flexibility decrease with additional cost. Similarly, Chou et al (2008) [29] distinguish between *range* (the different scenarios a system can adapt to) and *response* (the cost of using additional flexibility links) and show that improving response outperforms improving range. This conclusion suggests that in primary care practice, the benefits of limiting the number of physicians that can see a patient is likely to outweigh the higher range provided by a fully flexible practice where any physician can see the patient.

Flexible queuing systems have been studied by Sheikhzadeh et. al. (1998) [30] using a similar chaining configuration. They compare full flexibility, or "pooling", with a 2-chain configuration, i.e., one where two “neighboring” queues are linked to each server and two neighboring servers are connected to each queue. They find that the chained system works almost as well as the fully flexible system if the assumption of homogenous demand and service rate holds. The analysis is generalized in Gurumurthi and Benjaafar (2004) [31] to flexible queuing systems with general customer and service flexibility under Poisson-distributed demand and service rates. They show that the optimal allocation depends on the characteristics of the demand and particular policy implemented.

As in the case of cross-training in serial production lines (Hopp et al. (2004) [32]), flexibility has been found to be beneficial when implementing (1) capacity balancing, or balancing the expected workload among physicians. In this case flexibility will allow the load to be shared among physicians, which improves overall timely access and physician utilization; and (2) variability buffering, which refers to a flexibility configuration that

accommodates to variability in patient demand. They used a MDP to compare different strategies of cross-training and found that configurations parallel to chaining “have the potential to be robust and efficient methods” [32].

Though extensive studies have been conducted in manufacturing flexibility and its more recent application to other areas, there are, however, key operational differences that make the application of flexibility to primary care more complex and worthy of further analysis, as we explore in the next chapter.

CHAPTER 3

MODELING APPROACH

3.1 Assumptions

As we model a practice that implements a pre-scheduled appointment paradigm and an open access scheme at the same time, we assume that the daily capacity for each physician is the same and known in advance. In primary care practice, each appointment usually takes 20 minutes and practitioners are paid by the number of 20 minute appointments. Since each physician normally works eight hours a day, this leads to a capacity of 24 slots per physician per day.

We assume that the demands of pre-scheduled and open access appointments in practice are independent of each other, and for each physician, demands for pre-scheduled appointments and open access appointments are also independent. Further, we assume that demand distributions of pre-scheduled appointments and open access appointments are known (can be estimated by historical records) and belong to the Poisson distribution.

The open access paradigm increases the timely access effect that leads to much lower patient no show rates [33]. To include the no-show effects in our model, we treat the actual show-up rate as a *revenue* associated with each accessing paradigm. Thus we consider the revenue associated with meeting one open access demand to be higher than that of satisfying one pre-scheduled appointment.

To investigate the value of flexibility in primary care practice, we configure a system with three different flexibilities: full flexibility, partial flexibility (2-chain) and no flexibility (dedicated). To encourage continuity, we assume that seeing a patient from

another physician's panel will generate a slightly less revenue for a physician compared to satisfying a demand from his/her own panel.

3.2 Model formulation

We model the problem as a stochastic integer programming problem with stationary probability distribution and contribution (i.e. revenue). Below we show the notation for the dedicated cases (i.e. no flexibility) and for a scenario of 2 physicians with full flexibility. The notation is as follows:

N : Capacity of each physician.

M : Number of physicians and therefore panel. We index physicians with $i \in [1..M]$.

C_p : Cost of missing one pre-scheduled demand.

C_o : Cost of missing one open access demand.

N_i^p : Number of slots allocated for pre-scheduled demand of physician *i*.

d_i^p : Demand for pre-scheduled appointments of physician *i*.

d_i^o : Demand for open access appointments of physician *i*.

$p_i(\cdot)$: Probability mass function of pre-scheduled demand for physician *i*.

$q_i(\cdot)$: Probability mass function of open access demand for physician *i*.

$F_i(\cdot)$: Cumulative distribution function of pre-scheduled demand for physician *i*.

$\Phi_i(\cdot)$: Cumulative distribution function of open access demand for physician *i*.

$EC_i^p(\cdot)$: Expected cost of missing pre-scheduled demand for physician *i*.

$EC_i^o(\cdot)$: Expected cost of missing open access demand for physician *i*.

$EC_i(\cdot)$: Total expected cost of missing demands for physician *i*.

The notation for a general formulation (i.e. more than 2 physicians with any configuration of flexibility) will be demonstrated in the subsequent sections.

3.2.1 Formulation for dedicated flexibility

An individual physician without any flexibility is defined as one who can only serve the patients from his/her own panel. The system configuration is shown below in Figure 3.1:

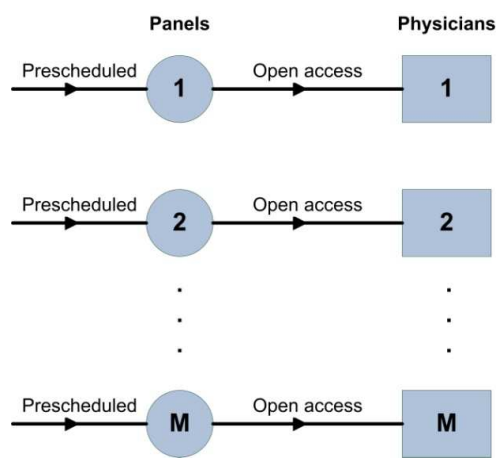


Figure 3.1 System configuration for dedicated flexibility.

For each number of slots that are allocated for a given pre-scheduled demand $N_i^p \in \{0,1,2,...,N\}$, the expected cost of missing the pre-scheduled demand for each physician is:

$$EC_i^p = \sum_{d_i^p = N_i^p + 1}^{\infty} C_p \cdot (d_i^p - N_i^p) \cdot p_i(d_i^p) \tag{3.2.1}$$

and the expected cost of missing a given open access demand for each physician is:

$$EC_i^o = \sum_{d_i^p=0}^{N_i^p} p_i\left(d_i^p\right)\cdot \sum_{d_i^o=N-d_i^p+1}^{\infty} C_o\cdot\left[d_i^p-\left(N-d_i^p\right)\right]\cdot q_i\left(d_i^o\right)+$$
$$\left[1-F_i\left(N_i^p\right)\right]\cdot \sum_{d_i^o=N-N_i^p+1}^{\infty} C_o\cdot\left[d_i^o-\left(N-N_i^p\right)\right]\cdot q_i\left(d_i^o\right) \tag{3.2.2}$$

The total expected cost of missing demands for the panel of physician *i* is equal to the sum of equation (3.2.1) and (3.2.2). Our objective is to find the optimal number of slots reserved for pre-scheduled appointments N_i^{p*} that minimizes the total expected cost of missing demands for physician *i*. For the dedicated flexibility configuration, we can use theorem 1 to find N_i^{p*}:

Theorem 1. *For the dedicated case, the optimal number of slots allocated for pre-scheduled appointments of each individual physician does not depend on the distribution of the pre-scheduled demand but relies on the total capacity N, the costs scale* C_p / C_o *and the inverse cumulative distribution function of his/her own open access demand, specifically:*

$$N_i^{p*} = N - \Phi_i^{-1}(1-\frac{C_p}{C_o}) \tag{3.2.3}$$

The proof is shown in the appendix.

3.2.2 Formulation for two physicians with full flexibility

In a fully flexible practice, patients can be seen by any available physician. We divide the case of two physicians with full flexibility into two scenarios: (1) the physicians also have full flexibility in pre-scheduled appointments; (2) the physicians only have full flexibility in open access appointments.

For physicians that have both full flexibility in pre-scheduled and open access practices, the optimal value of N_i^{p*} can be determined by theorem 2:

Theorem 2. *For a system that has both full flexibility in pre-scheduled and open access practices, the optimal value of each* N_i^{p*} *should satisfy:* $\sum_{i=1}^{M} N_i^{p*} = M \cdot N - \Phi^{-1}\left(1 - C_p / C_o\right)$, *where* $\Phi^{-1}(\cdot)$ *is the inverse cumulative distribution function where the mean rate equals to the sum of each individual open access demand mean rate.*

With full flexibility in the pre-scheduled and open access practice, both the demand and capacity of *M* physicians can be aggregated proportionally. This means that we can use a single system, with aggregated capacity and demand, to substitute for the case of multiple physicians, and the optimal value of N_i^{p*} can be obtained from equation (3.2.3). Further, considering each physician individually, the number of N_i^{p*} can be any value that is no larger than *N*, but the sum of these N_i^{p*} should be always equal to the value indicated in theorem 2.

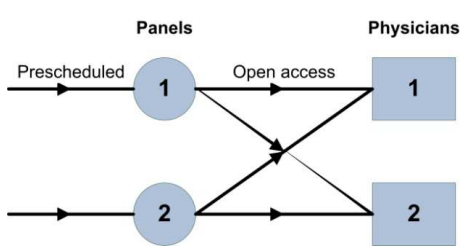


Figure 3.2 System configuration for two physicians sharing open access demands.

For the scenario where pre-scheduled patients see their own physician, but the time-sensitive open access patients can be seen by more than one physician (the system

configuration is shown in figure 3.2), we use the following theorem to determine the optimal values of N_1^{p*} and N_2^{p*} :

Theorem 3. *The optimal number of appointment slots for each physician i to make available to pre-scheduled patients in a two-physician partially flexible practice, where the two physicians share open access demands, is the smallest integers* N_1^p *and* N_2^p *that satisfy:*

$$\frac{C_p}{C_o} \le [1-F_2(N_2^p)]\cdot[1-\Phi(2N-N_1^p-N_2^p-1)]+ \sum_{d_2^p=0}^{N_2^p} p_2(d_2^p)\cdot[1-\Phi(2N-N_1^p-d_2^p-1)] \tag{3.2.4}$$

and

$$\frac{C_p}{C_o} \le [1-F_1(N_1^p)]\cdot[1-\Phi(2N-N_1^p-N_2^p-1)]+ \sum_{d_1^p=0}^{N_1^p} p_1(d_1^p)\cdot[1-\Phi(2N-N_2^p-d_1^p-1)] \tag{3.2.5}$$

where $\Phi(\cdot)$ *is the cumulative distribution function where the mean rate equals to the sum of each individual open access demand mean rate. If both physicians have the same distribution of pre-scheduled demand (symmetric), then the optimal numbers of* N_1^{p*} *and* N_2^{p*} *are the same and equal to the smallest integer* N^p *such that:*

$$\frac{C_p}{C_o} \le [1-F_i(N^p)]\cdot[1-\Phi(2N-2N^p-1)]+ \sum_{d_i^p=0}^{N^p} p_i(d_i^p)\cdot[1-\Phi(2N-N^p-d_i^p-1)] \tag{3.2.6}$$

where *i can be any one of the two physicians.*

The proof can be found in the appendix. Observe that N_i^{p*} does not depend on the distribution of pre-scheduled demand for physician *i*.

3.2.3 Formulation for general configuration

We investigate a primary care practice involving more than two physicians with full flexibility, partial (2-chain), and no flexibility using a stochastic integer programming approach. The system configuration is demonstrated in figure 3.3.

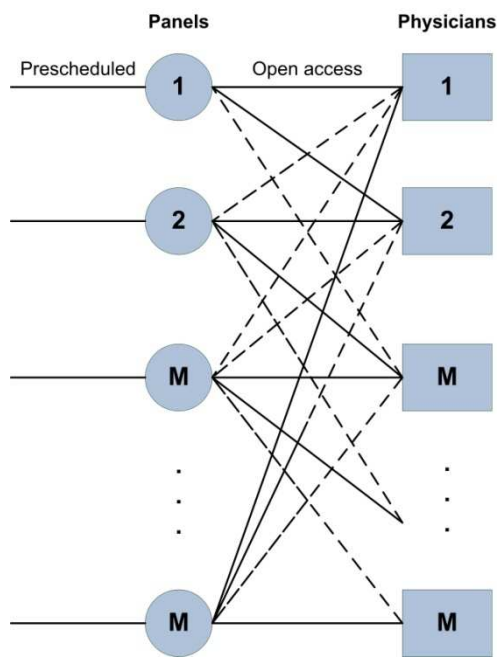


Figure 3.3 System configuration for partial and full flexibility.

Let *A* be the set of all possible links (*i*, *j*) such that patients in panel *i* can be served by physician *j*, R_i^p is the revenue associated with physician *i* seeing one of his pre-scheduled patients, and R_{ij}^o is the revenue associated with physician *j* seeing an open-access patient of panel *i*. Let *U* be the upper bound of the realization of pre-scheduled demand d_{is}^p and

open access demand d_{is}^{o} for scenario *s*, for instance, $U = 50$, which means $d_{is}^{p} \in \{0,1,2,.,50\}$ and $d_{is}^{o} \in \{0,1,2,.,50\}$ .We introduce the following variables:

$\phi_{iu_{is}} = 1$ if $d_{is}^{p} < N_i^{p}$, otherwise $\phi_{iu_{is}} = 0$. where $u_{is} = d_{is}^{p}$ and $u_{is} \in \{0,1,2,.,U\}$. $\phi_{iu_{is}}$ is introduced for pushing unused slots from pre-scheduled appointments to open access demands. The total number of binary variables $\phi_{iu_{is}}$ equals the number of physicians times the value of the upper bound of the demand realization. But these binary variables don't depend on the number of scenarios, since they only depend on the realization of pre-scheduled demand and have no relationship with the open access demand.

x_{is}^{p} : Number of patients pre-scheduled with physician *i* under demand scenario *s*.

x_{ijs}^{o} : Number of open access patients of panel *i* assigned to physician *j* under demand scenario *s*. For all $i = 1, 2, \ldots, M$ and $(i, j) \in A$.

We will consider demand scenarios *s* associated with a particular realization $(d_{1s}^{p}, d_{1s}^{o}, ..., d_{Ms}^{p}, d_{Ms}^{o})$ of demand and with a probability q_s. Our goal is to maximize the revenue of satisfying appointments, and following the notation previously introduced, we can formulate the problem as follows:

$$\textit{Objective:} \qquad Max \sum_{s=1}^{S} \sum_{i=1}^{M} q_s \left[R_i^p x_{is}^p + \sum_{(i,j) \in A} R_{ij}^o x_{ijs}^o \right] \tag{3.2.7}$$

$$\textit{Subject to:} \qquad N_i^p \le N \qquad \forall i = 1, 2, \ldots, M \tag{3.2.8}$$

$$N_i^p \le d_{is}^p + N\phi_{iu_{is}} \qquad \forall i = 1, 2, \ldots, M, \quad s = 1, 2, \ldots, S \tag{3.2.9}$$

$$N_i^p \ge d_{is}^p \phi_{iu_{is}} \qquad \forall i = 1, 2, \ldots, M, \quad s = 1, 2, \ldots, S \tag{3.2.10}$$

$$x_{is}^p \le N_i^p \qquad \forall i = 1, 2, \ldots, M, \quad s = 1, 2, \ldots, S \tag{3.2.11}$$

$$x_{is}^{p} \le d_{is}^{p} \qquad \forall i = 1,2,\ldots,M, \quad s = 1,2,\ldots,S \tag{3.2.12}$$

$$\sum_{i:(i,j)\in A} x_{ijs}^{o} \le N - d_{js}^{p}\phi_{ju_{is}} \qquad \forall j = 1,2,\ldots,M, \quad s = 1,2,\ldots,S \tag{3.2.13}$$

$$\sum_{i:(i,j)\in A} x_{ijs}^{o} \le N - N_{j}^{p} + \phi_{ju_{i}} N \qquad \forall j = 1,2,\ldots,M, \quad s = 1,2,\ldots,S \tag{3.2.14}$$

$$\sum_{j:(i,j)\in A} x_{ijs}^{o} \le d_{is}^{o} \qquad \forall i = 1,2,\ldots,M, \quad s = 1,2,\ldots,S \tag{3.2.15}$$

$$\phi_{iu_{is}} \in \{0,1\} \qquad \forall i = 1,2,\ldots,M, \quad u_{is} = 1,2,\ldots,U \tag{3.2.17}$$

$$N_{i}^{p}, x_{is}^{p}, x_{ijs}^{o} \ge 0 \qquad \forall i, j = 1,2,\ldots,M, \quad (i,j) \in A, \quad s = 1,2,\ldots,S \tag{3.2.18}$$

Equation (3.2.9) ensures that $\phi_{iu_{is}} = 1$ if $d_{is}^{p} < N_{i}^{p}$. Equation (3.2.10) ensures that $\phi_{iu_{is}} = 0$ if $d_{is}^{p} > N_{i}^{p}$. Equation (3.2.11) limits the number of pre-scheduled appointments to the desired capacity. Equations (3.2.13) and (3.2.14) ensure that the total open access appointments for physician *i* do not exceed remaining capacity when $\phi_{iu_{is}} = 1$ and $\phi_{iu_{is}} = 0$ respectively. Equation (3.2.17) is the binary constraint.

CHAPTER 4

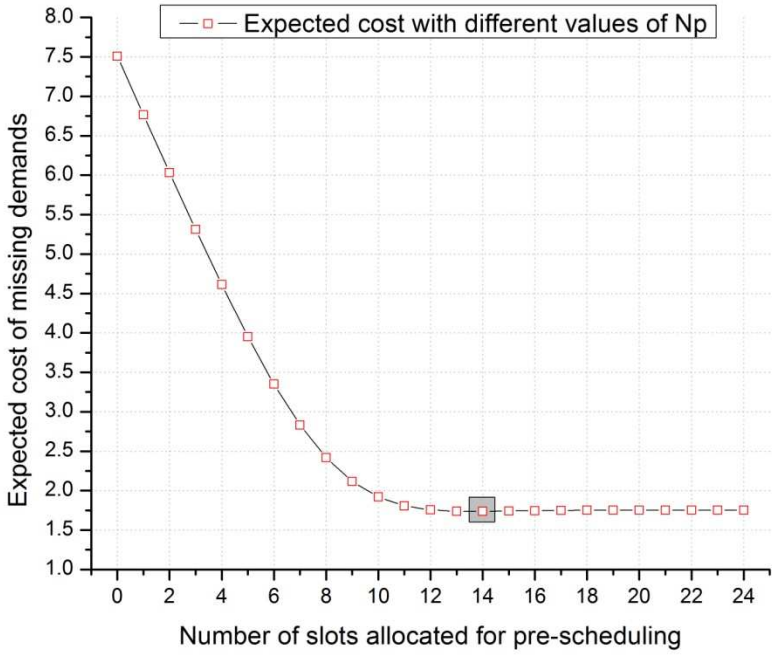
VALUE OF FLEXIBILITY

4.1 Practice without any flexibility

We refer to the primary care practice without any flexibility as the *dedicated* case. Each physician can only see the patients come from his/her own panel. If the capacity, i.e. the capacity for pre-scheduled demand or the capacity for open access demand, is used up, the remaining demand will have to be turned away and a cost will incurred. We can use equation (3.2.3) to directly decide the optimal number of slots that should be allocated for pre-scheduled appointments of each physician in the dedicated case. Notice that equation (3.2.3) has a newsvendor type solution which does not depend on the distribution of pre-scheduled demand.

$$N_i^{p*} = N - \Phi_i^{-1}(1 - \frac{C_p}{C_o}) \tag{3.2.3}$$

Figure 4.1 and 4.2 show the total expected costs of missing demands in two instances for the dedicated case: the capacity of each physician is 24 slots per day, and the cost of missing one pre-scheduled appointment is set to 0.75 and the cost of missing an open access demand is 0.9; these costs are equal to the typical show rates of each type of demand as indicated by Bennett and Baxley (2009) [33]. All demands belong to Poisson distribution. In Figure 4.1, the demand rates for pre-scheduled and open access appointments are 10 and 14 respectively. In Figure 4.2, we change them to 16 and 8. We can see that since the cost of missing one open access demand is higher than missing a pre-scheduled demand, the marginal gain of increasing the value of N_p is significant at the beginning but trends to be flat when it approaches the optimal point.



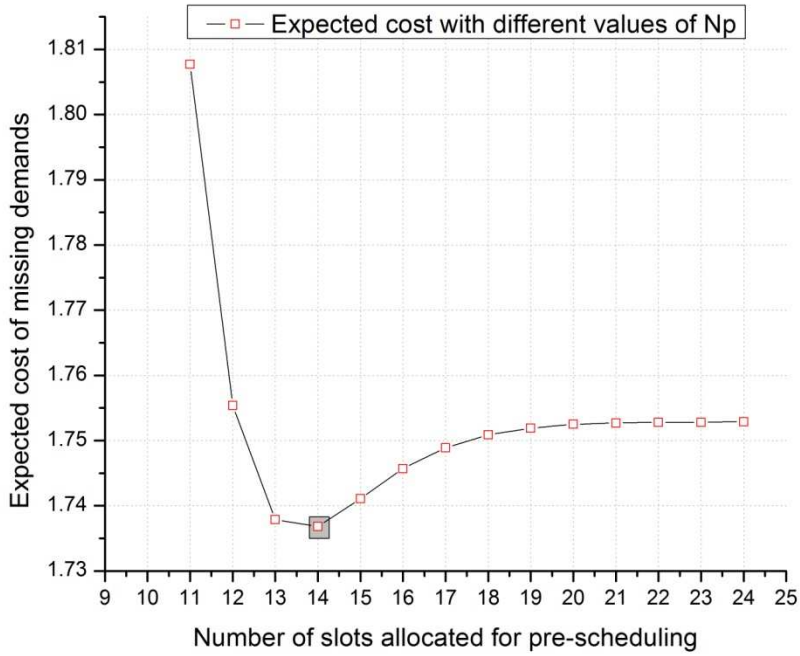
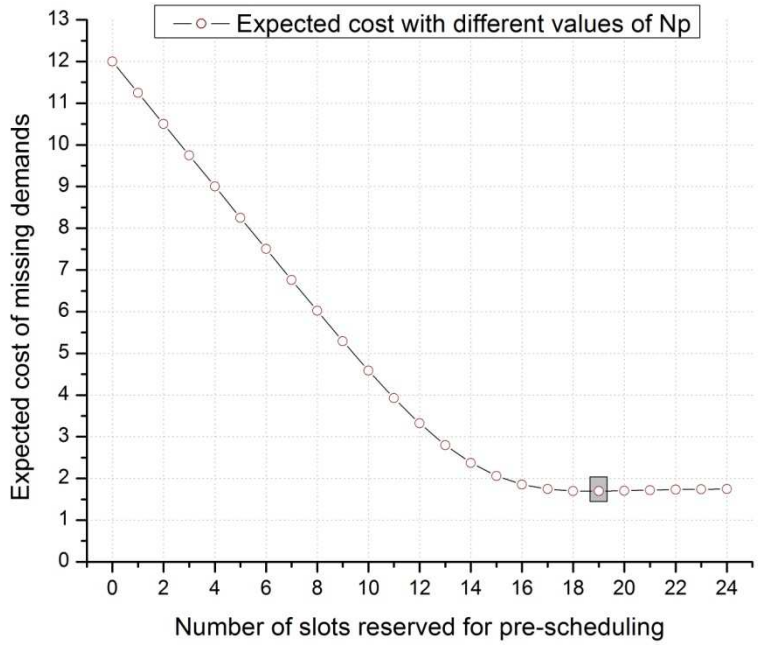


Figure 4.1 Dedicated case with demand rates 10 and 14 for pre-scheduling and open access respectively. And a closer view of the value near the optimal point.



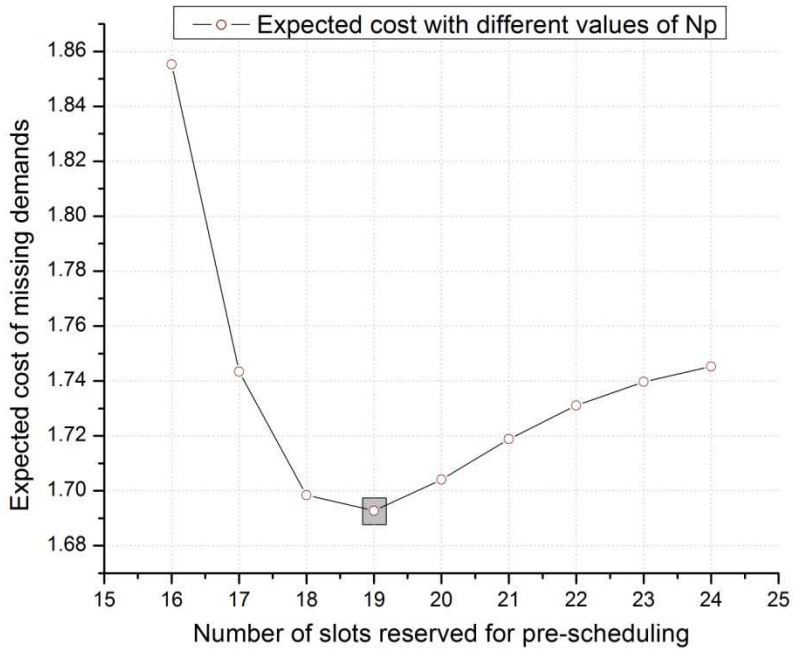


Figure 4.2 Dedicated case with demand rates 16 and 8 for pre-scheduling and open access respectively. And a closer view of the value near the optimal point.

4.2 Two physicians with open access flexibility

When physicians have full flexibility to share both pre-scheduled and open access patients, the practice can be treated as a dedicated system with pooled demands and capacities. For the case that physicians only have flexibility in open access practice shown in Figure 3.2, we can use conditions (3.2.4) and (3.2.5) to search the optimal value of N_1^{p*} and N_2^{p*} directly. The running complexity is $O(N^2)$, where N is the capacity of each physician. Particularly, if two physicians have the same demand rate of pre-scheduled appointments, we can use the condition (3.2.6) to search the optimal value of N_1^{p*} (N_2^{p*}) in $O(N)$ time. Figures 4.3 and 4.4 illustrates two examples:

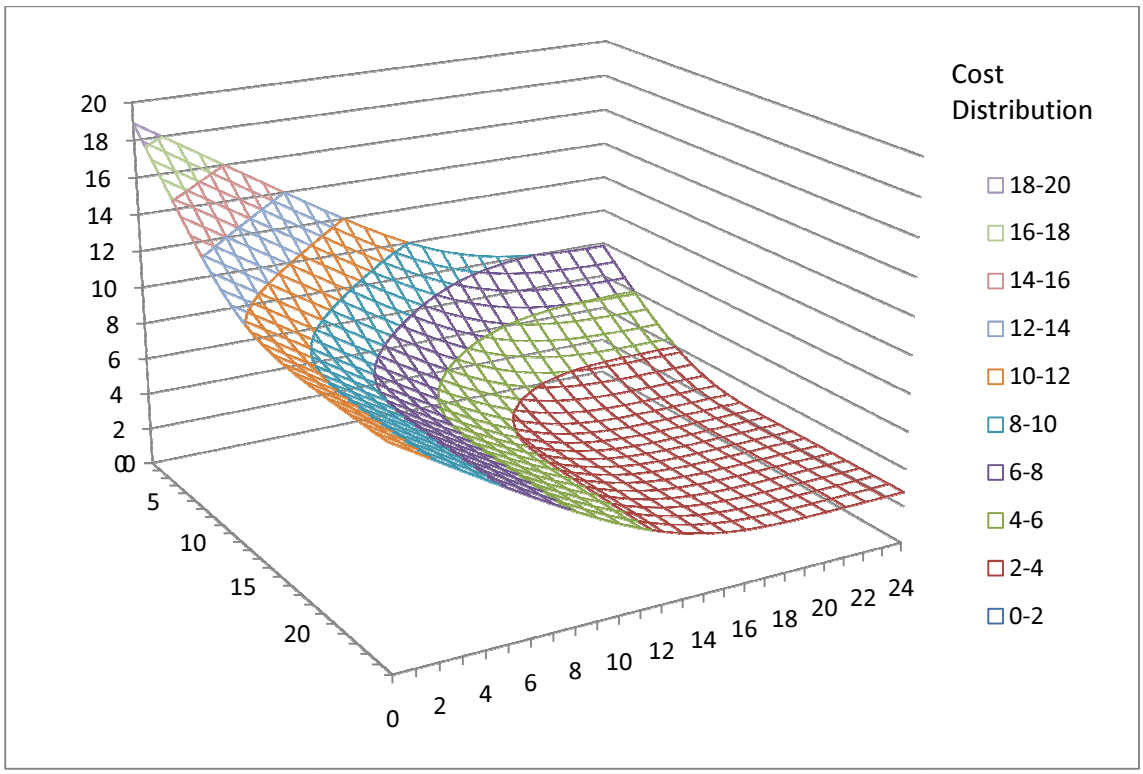


Figure 4.3. Two physicians have flexibility in open access practice.
where N1p = 19 and N2p = 14.

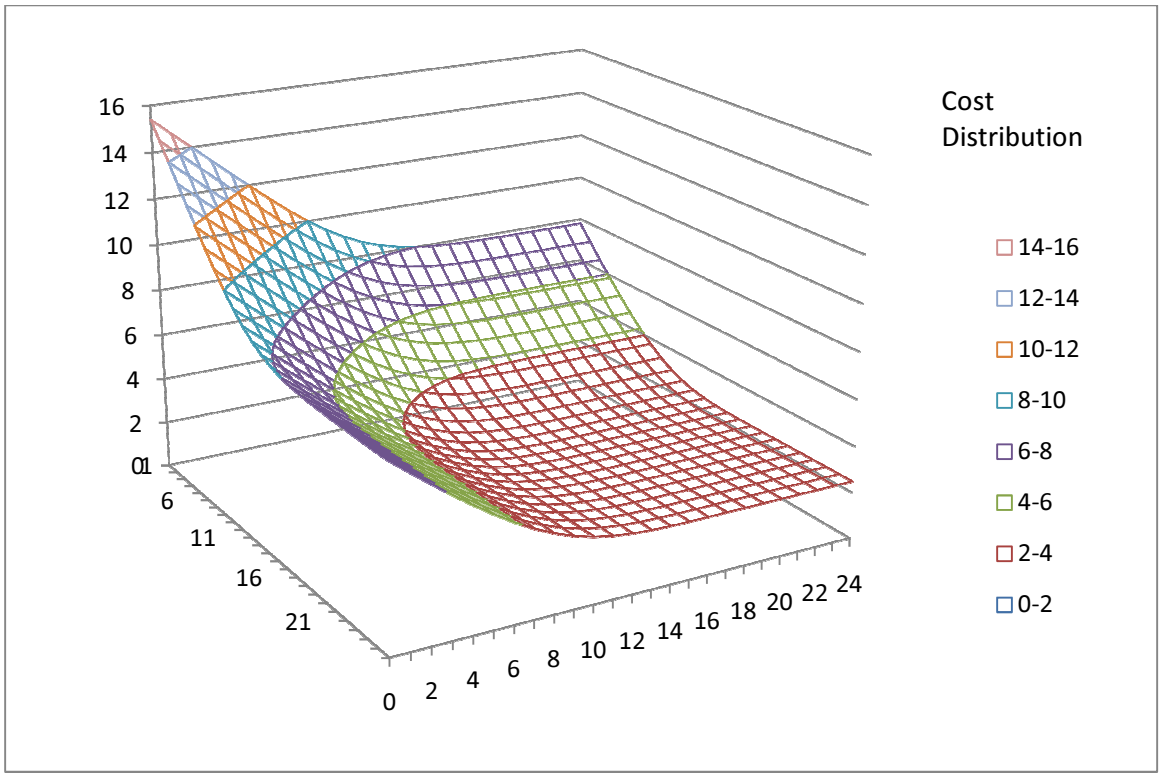


Figure 4.4. Two physicians have flexibility in open access practice.
where N1p = 16, N2p = 16.

4.3 Value of flexibility in a practice with three physicians

For a primary care practice with three physicians or more, it is too complicated to get any closed form or condition for the optimal value of N_i^{p*}. To investigate the value of flexibility in this circumstance, we will use the stochastic integer programming model introduced in section 3.2.3. Three different levels of flexibility will be evaluated—full flexibility, partial flexibility (2-chain) and no flexibility (dedicated)—for a variety of settings with symmetric and asymmetric demand distributions and different levels of system utilization (from 40% up to 160%). The system utilization refers to the scale between expected demands and available capacities. We will focus on three measures: system revenue, timely access rate and continuity rate. The system revenue stands for the

total revenue of meeting patient demands; timely access rate is the percentage of patients can who get access to care; and continuity rate presents the percentage of patients who see their own physician. Our model provides the optimal value of $N_1^{p^*}, N_2^{p^*}, ..., N_M^{p^*}$, and the optimal allocation of patients to physicians (i.e., for each panel that how many patients should see their own physician, and how many of them should be diverted to a different physician).

The computational complexity of our model heavily depends on the number of scenarios, which is the most influential factor, and the number of physicians. We tested the model of the general formulation using IBM ILOG OPL 6.3 on a PC with Intel 2 Cores Dual 2×3G CPU and 8GB memory. For three physicians with 100,000 scenarios, it takes 50 hours to get the results when the relative MIP (Mixed Integer Programming) gap tolerance is set to 1%. Although our stochastic integer programming model can theoretically investigate the value of flexibility for any flexibility configuration with any number of physicians, the time-consuming nature of the optimization and evaluation makes it impractical. Fortunately, a computationally effective sample average approximation method was proposed by S. Solak [34] to provide an efficient solution approach for two-stage stochastic integer programming problems. The basic idea of the sample average approximation method used in our research is to create a manageable number of samples/scenarios to produce an estimation of the optimal objective value and corresponding first stage solutions. We then further run a large number of scenarios to have a precise estimation of the objective value based on the fixed first stage solution. This process is repeated over a number of replications to provide confidence intervals and

statistical guarantees on the quality of the estimation. To allow for a fair comparison, the 2-chain, full flexibility and dedicated case use the same set of scenarios.

To investigate the value of flexibility for three physicians under different levels of system utilization, we first focus on the symmetric demand distributions (i.e., all panels generate identically distributed demands) to gain insights on its effectiveness to hedge against demand uncertainty. We then analyze the impact of asymmetric demand distributions, where flexibility additionally helps to balance the average supply and demand across providers. We also use several cases in which the demand ratio between prescheduled and open access demand changes significantly.

4.3.1 Results for three physicians with symmetric demand distributions

Following the findings of Bennett and Baxley (2009) [33], we assume a typical no show rate for pre-scheduled demand of 25%, and a 10% no show rate for open access demand. Thus, we assign the revenue of scheduling one pre-scheduled demand as 0.75, and 0.9 for seeing one open access patient. These values stand for the actual show rates. To encourage continuity in the system, we assume that there is a 0.05 cost of seeing patients from another physician's panel. System utilization in our model is defined as the ratio of the expected total demand for the clinics and total available capacity. For instance, in a practice with three physicians, suppose each physician has a demand rate of 10 for prescheduled appointment and 14 for open access demand. The total expected demand is $10\times3+14\times3=72$, and the total capacity is $24\times3 = 72$, therefore, the system utilization is 100%. To make the system under-/over-utilized, a factor varying from 0.4 to 1.6 will be multiplied to the mean demands rate to generate different levels of utilization. We use

four cases with demand ratios of 10/14, 14/10, 6/18 and 18/6 to investigate the value of flexibility for a practice with three physicians having symmetric demand distributions.

Symmetric Case 1 (10/14). Table 4.1 summarizes the assumptions for the first case where the demand ratio between prescheduled and open access demands is 10/14.

Physician capacity	24
Number of physicians in practice	3
Scenarios for each replication	1000
Number of replications	50
Revenue of seeing one pre-scheduled demand	0.75
Revenue of seeing one owned open access demand	0.90
Revenue of seeing one diverted open access demand	0.85
Mean demand rate for pre-scheduled appointments	[10, 10, 10]
Mean demand rate for open access appointments	[14, 14, 14]
Relative MIP tolerance gap	0.01%

Table 4.1 Assumptions for 3 physicians with symmetric demand distributions in Symmetric Case 1 (10/14).

In our experiments, one interesting and promising phenomena is that the 95% confidence interval of the objective values (system revenue) resulting from 50 replications lies in a very narrow range, the variance over the mean is less than 1%. Therefore, we can use the mean objective value of 50 replications to achieve an accurate estimation of the real objective value over the whole population of scenarios. Computational effort for the second step of stochastic integer program can be saved due to this. Table 4.2 shows an instance of the objective value statistics for different

flexibilities when the system is balanced. Figure 4.5 presents the corresponding Box-Whisker plot.

	2-chain	Full Flex	Dedicated
Conf. Intervals (One-Sample)	**100% utilization Obj**	**100% utilization Obj**	**100% utilization Obj**
Sample Size	50	50	50
Sample Mean	57.115	57.1535	55.0977
Sample Std Dev	0.1399	0.1402	0.1367
Confidence Level (Mean)	95.0%	95.0%	95.0%
Degrees of Freedom	49	49	49
Lower Limit	57.0753	57.1137	55.0588
Upper Limit	57.1548	57.1934	55.1365
Confidence Level (Std Dev)	95.0%	95.0%	95.0%
Degrees of Freedom	49	49	49
Lower Limit	0.1168	0.1172	0.1142
Upper Limit	0.1743	0.1748	0.1703

Table 4.2 Statistics of objective value for different flexibilities with 100% utilization in Symmetric Case 1.

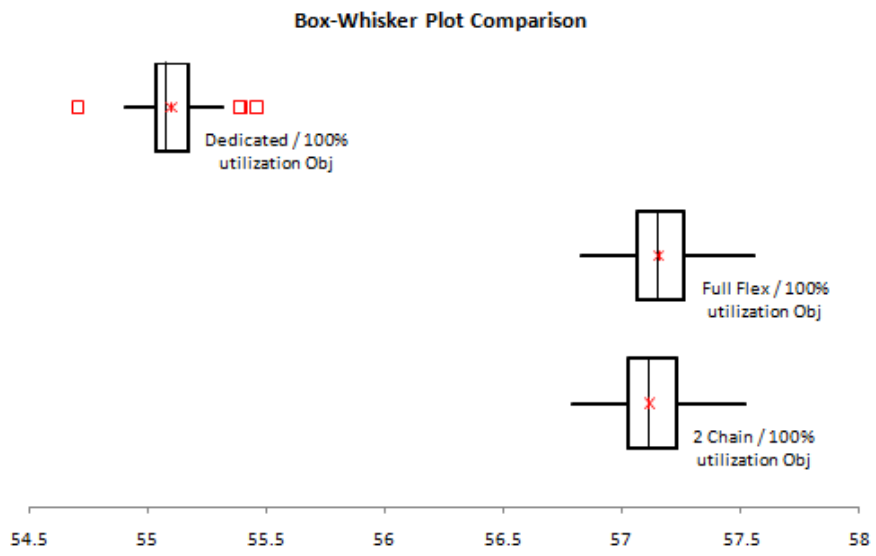


Figure 4.5 Box-Whisker Plot comparison of objective values for different flexibilities with 100% utilization in Symmetric Case 1.

A possible explanation for this concentrated distribution of objective values might be the low variation of the aggregate system demand distribution. Table 4.3 demonstrates the distribution of total arrival demand of 50 replications when the system is balanced (i.e., 100% utilization).

	Total demand
Conf. Intervals (One-Sample)	**100% utilization Demand**
Sample Size	50
Sample Mean	71.9517
Sample Std Dev	0.2838
Confidence Level (Mean)	95.0%
Degrees of Freedom	49
Lower Limit	71.8711
Upper Limit	72.0323
Confidence Level (Std Dev)	95.0%
Degrees of Freedom	49
Lower Limit	0.2370
Upper Limit	0.3536

Table 4.3 Statistics of total demands for 3 physicians with 100% utilization in Symmetric Case 1.

We can see that the value of total demand varies very little among the replications. Though the demands are sampled from Poisson distribution and the realization varies dramatically in each scenario, for a sum of 1000 scenarios, the averaged total demand will closely approximate the sum of mean demand rates. Since the objective value is equal to the revenue of demands which the system could satisfy, a "flat" total demand distribution among the replications will produce a "concentrated" objective value estimation. As mentioned earlier, we will use the mean objective value estimated from 50 replications to approximate the actual value over the whole scenario space.

Tables 4.4, 4.5, and 4.6 give the measurement and comparison of 2-chain flexibility, full flexibility and dedicated case under different levels of system utilization in the three dimensions of interest: system revenue, timely access rate and continuity rate.

System Revenue Utilization	40%	80%	100%	120%	160%
2-chain	25.2142	47.574	57.115	59.89385	62.00081
Full Flex	25.2142	47.5819	57.1535	59.91734	62.02412
Dedicated	25.2141	46.8694	55.0977	58.63243	60.85155
2-chain vs Dedicated	0.00%	1.50%	3.66%	2.15%	1.89%
Full vs Dedicated	0.00%	1.52%	3.73%	2.19%	1.93%

Table 4.4 Measurement for different flexibilities in term of system revenue in Symmetric Case 1 (10/14).

Timely Access Rate Utilization	40%	80%	100%	120%	160%
2-chain	100%	99.88%	95.29%	82.01%	62.66%
Full Flex	100%	99.88%	95.29%	81.99%	62.65%
Dedicated	100%	98.40%	91.78%	80.72%	62.24%
2-chain vs Dedicated	0.00%	1.50%	3.82%	1.59%	0.69%
Full vs Dedicated	0.00%	1.50%	3.82%	1.58%	0.66%

Table 4.5 Measurement for different flexibilities in term of timely access rate in Symmetric Case 1 (10/14).

Continuity Rate Utilization	40%	80%	100%	120%	160%
2-chain	100%	98.24%	95.29%	97.03%	96.97%
Full Flex	100%	98.52%	96.41%	97.68%	97.59%
Dedicated	100%	100.00%	100.00%	100.00%	100.00%
2-chain vs Dedicated	0.00%	-1.76%	-4.71%	-2.97%	-3.03%
Full vs Dedicated	0.00%	-1.48%	-3.59%	-2.32%	-2.42%

Table 4.6 Measurement for different flexibilities in term of continuity rate in Symmetric Case 1 (10/14).

And Figures 4.6, 4.7 and 4.8 are the comparisons illustrated in plot form respectively.

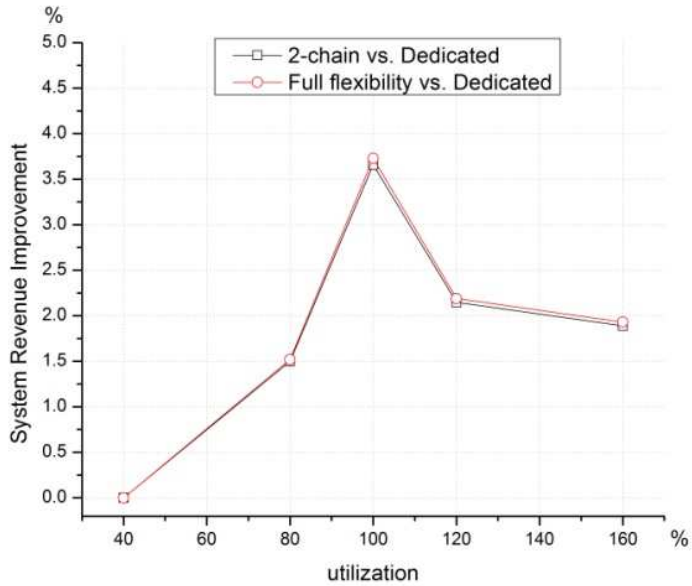


Figure 4.6 Comparisons of different flexibilities in term of system revenue in Symmetric Case 1 (10/14).

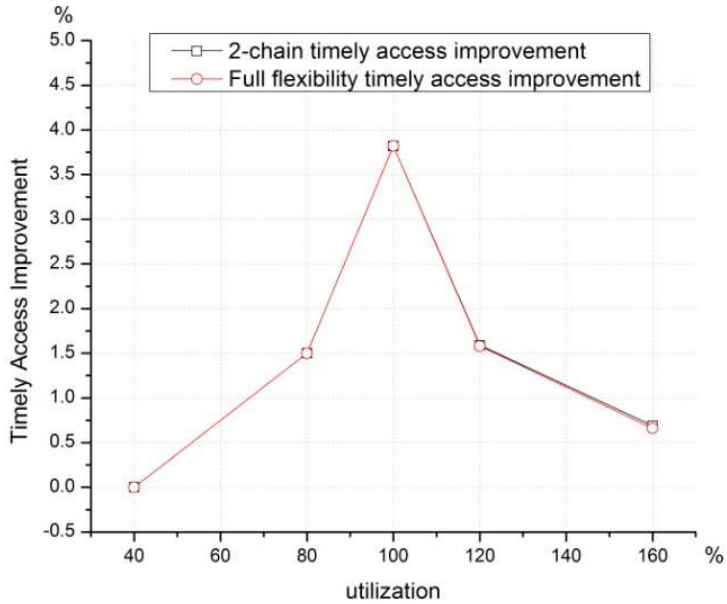


Figure 4.7 Comparisons of different flexibilities in term of timely access rate in Symmetric Case 1 (10/14).

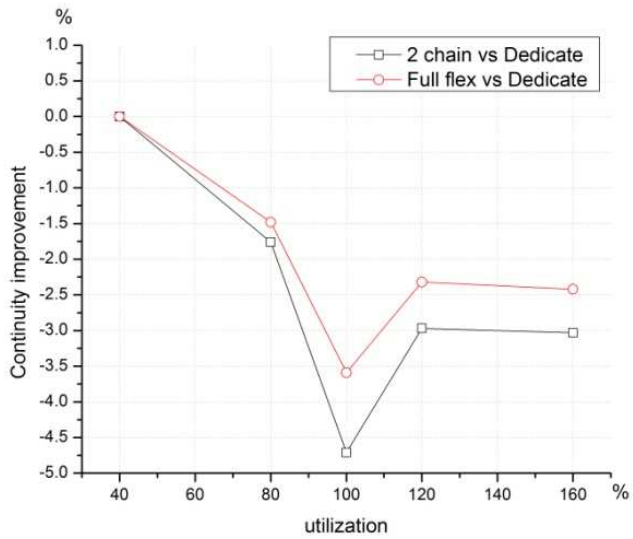


Figure 4.8 Comparisons of different flexibilities in term of continuity rate in Symmetric Case 1 (10/14).

We can see that the highest benefit of both system revenue and timely access rate is achieved in the case where the system is balanced, i.e. when the expected demand equals the available capacity. When the system is under-utilized, most of the demands can be met and therefore result in lower benefits of flexibility. By contrast, when the system is over-utilized and more likely to miss the demand, flexibility still has the ability to shift demand to a less utilized physician. Therefore, the graph of system performance improvement is not symmetric.

The benefits of 2-chain flexibility are almost as high as those of full flexibility, with only a 0.07% detriment in terms of system revenue. One interesting result is that the timely access rates of 2-chain flexibility and full flexibility are nearly the same no matter what the level of utilization of the system is. This is consistent with the results reported in the literature on flexibility in manufacturing settings. The difference in revenue is even

lower in our healthcare setting, since the prescheduled demand cannot be shared between physicians; flexibility can only be used on the open access demand.

Intuition tells us that since full flexibility has more "outbound" links than 2-chain flexibility, it should have a better ability to absorb incoming demands and yield a higher timely access rate than 2-chain flexibility. This is indeed true for the dynamic setting of patient scheduling where allocation decisions are made as requests arrive, with limited knowledge of the overall demand that will need to be serviced (Hippchen (2009) [35]). By contrast, in the aggregate demand setting captured by our two-stage stochastic integer programming approach, the patient allocation is only performed after the full system demand is known. Although, 2-chain flexibility achieves almost the same benefits as full flexibility, in our aggregate setting, there are instances where full flexibility will clearly dominate. For instance, consider a practice with four physicians, where each has 10 slots left for open access, and the demands for open access are 20, 20, 0 and 0 respectively. In this extreme case, the 2-chain flexibility can only meet 30 open access demands the full flexibility can satisfy all of them. Since this type of instance would occur with a low probability, from a statistical point of view, the 2-chain flexibility has almost the same effectiveness to absorb the demand as full flexibility.

Another phenomena that deserves our attention is that the diversion rate, which equals one minus the continuity rate, of 2-chain flexibility is higher than that of full flexibility. Our initial intuition tells us that since full flexibility has more "outbound" links than 2-chain flexibility, it should have a higher probability that the demand will be diverted to other physicians. In reality, however, a single patient redirection to an available physician, which can be made directly under full flexibility, may require redirecting several patients

along the 2-chain if the initial patient's panel and available physician involved are not connected. For example, Figure 4.9 shows a case of three physicians where each physician has 10 slots left for open access, and the demands are 16, 10 and 4 respectively. We can see that the total number of diversions under 2-chain flexibility is 12, but only 6 under the full flexibility. Since 2-chain flexibility requires more "jumps" to shift the demands, the diversion rate of 2-chain is higher than that of full flexibility in our model.

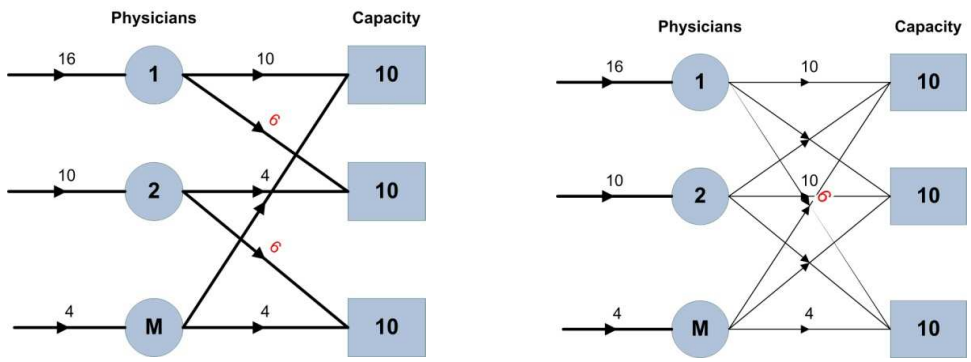


Figure 4.9 An example of diversion process in 2-chain and full flexibility.

While the number of redirections is greater in the 2-chain system, it is important to note that each patient will always see either one of two physicians. We believe this results in stronger continuity and efficiency from the perspective of both the patient (who could quickly get to be familiar and comfortable with both physicians) and the physician (who would be able to follow the other's panel relatively well and share cases with only one other physician).

Symmetric Case 2 (14/10). To further study the impact of the demand ratio on system performance, we reverse the ratio from 10/14 used in case 1 to 14/10. Tables 4.7, 4.8, and 4.9 give the measurement and comparison of 2-chain flexibility, full flexibility and

dedicated case under different levels of system utilization. We can see that the system performs nearly the same as in case 1 where the demand ratio is 10/14.

System Revenue Utilization	40%	80%	100%	120%	160%
2-Chain	24.25121	46.24628	55.28167	57.84773	59.5821
Full Flex	24.25121	46.25338	55.32369	57.86957	59.60268
Dedicated	24.25103	45.53759	53.34754	56.66859	58.62003
2-Chain vs Dedicated	0.00%	1.56%	3.63%	2.08%	1.64%
Full vs Dedicated	0.00%	1.57%	3.70%	2.12%	1.68%

Table 4.7 Measurement for different flexibilities in term of system revenue in Symmetric Case 2 (14/10).

Timely Access Rate Utilization	40%	80%	100%	120%	160%
2-Chain	100.00%	99.87%	95.30%	82.02%	62.72%
Full Flex	100.00%	99.86%	95.32%	82.01%	62.68%
Dedicated	100.00%	98.36%	91.80%	80.70%	62.68%
2-Chain vs Dedicated	0.00%	1.53%	3.82%	1.64%	0.06%
Full vs Dedicated	0.00%	1.53%	3.84%	1.63%	0.00%

Table 4.8 Measurement for different flexibilities in term of timely access rate in Symmetric Case 2 (14/10).

Continuity Rate Utilization	40%	80%	100%	120%	160%
2-Chain	100.00%	98.23%	95.37%	97.28%	97.41%
Full Flex	100.00%	98.51%	96.44%	97.87%	97.92%
Dedicated	100.00%	100.00%	100.00%	100.00%	100.00%
2-Chain vs Dedicated	0.00%	-1.77%	-4.63%	-2.72%	-2.59%
Full vs Dedicated	0.00%	-1.49%	-3.56%	-2.13%	-2.08%

Table 4.9 Measurement for different flexibilities in term of continuity rate in Symmetric Case 2 (14/10).

Symmetric Case 3 (6/18). Further, we change the demand ratio to 6/18, a "polarized" case that the system is fulfilled with more open access demands. This represents an urgent care center, where walk-ins are more prominent than scheduled visits. Tables 4.10, 4.11, and 4.12 give the measurements of system performance under different levels of system utilization.

System Revenue Utilization	40%	80%	100%	120%	160%
2-Chain	23.40038	48.88095	58.87549	61.98912	64.56461
Full Flex	23.40038	48.88901	58.91315	62.01434	64.57808
Dedicated	23.40031	48.17918	56.83515	60.6714	63.53728
2-Chain vs Dedicated	0.00%	1.46%	3.59%	2.17%	1.62%
Full vs Dedicated	0.00%	1.47%	3.66%	2.21%	1.64%

Table 4.10 Measurement for different flexibilities in term of system revenue in Symmetric Case 3 (6/18).

Timely Access Rate Utilization	40%	80%	100%	120%	160%
2-Chain	100.00%	99.86%	95.25%	81.96%	61.41%
Full Flex	100.00%	99.86%	95.25%	81.96%	61.41%
Dedicated	100.00%	98.39%	91.81%	80.75%	60.45%
2-Chain vs Dedicated	0.00%	1.49%	3.74%	1.50%	1.58%
Full vs Dedicated	0.00%	1.49%	3.74%	1.50%	1.58%

Table 4.11 Measurement for different flexibilities in term of timely access rate in Symmetric Case 3 (6/18).

Continuity Rate Utilization	40%	80%	100%	120%	160%
2-Chain	100.00%	98.27%	95.33%	96.75%	97.91%
Full Flex	100.00%	98.53%	96.42%	97.44%	98.29%
Dedicated	100.00%	100.00%	100.00%	100.00%	100.00%
2-Chain vs Dedicated	0.00%	-1.73%	-4.67%	-3.25%	-2.09%
Full vs Dedicated	0.00%	-1.47%	-3.58%	-2.56%	-1.71%

Table 4.12 Measurement for different flexibilities in term of continuity rate in Symmetric Case 3 (6/18).

Symmetric Case 4 (18/6). Again, we reverse the demand ratio from 6/18 to 18/6 where the system has more prescheduled demands coming in. This demand profile represents a family medicine clinic. Tables 4.13, 4.14, and 4.15 show the system performance under different levels of system utilization.

System Revenue Utilization	40%	80%	100%	120%	160%
2-Chain	21.13375	44.83722	53.43987	55.83865	57.23444
Full Flex	21.13375	44.85259	53.54503	55.86207	57.25016
Dedicated	21.13375	44.16648	51.69283	54.82082	56.48597
2-Chain vs Dedicated	0.00%	1.52%	3.38%	1.86%	1.33%
Full vs Dedicated	0.00%	1.55%	3.58%	1.90%	1.35%

Table 4.13 Measurement for different flexibilities in term of system revenue in Symmetric Case 4 (18/6).

Timely Access Rate Utilization	40%	80%	100%	120%	160%
2-Chain	100.00%	99.80%	95.00%	81.97%	61.19%
Full Flex	100.00%	99.82%	95.16%	81.98%	61.18%
Dedicated	100.00%	98.36%	91.69%	80.78%	60.91%
2-Chain vs Dedicated	0.00%	1.47%	3.62%	1.47%	0.46%
Full vs Dedicated	0.00%	1.49%	3.79%	1.49%	0.45%

Table 4.14 Measurement for different flexibilities in term of timely access rate in Symmetric Case 4 (18/6).

Continuity Rate Utilization	40%	80%	100%	120%	160%
2-Chain	100.00%	98.33%	95.75%	97.71%	97.86%
Full Flex	100.00%	98.55%	96.53%	98.22%	98.28%
Dedicated	100.00%	100.00%	100.00%	100.00%	100.00%
2-Chain vs Dedicated	0.00%	-1.67%	-4.25%	-2.29%	-2.14%
Full vs Dedicated	0.00%	-1.45%	-3.47%	-1.78%	-1.72%

Table 4.15 Measurement for different flexibilities in term of continuity rate in Symmetric Case 4 (18/6).

Comparing the respective measurements of system improvement in all four symmetric cases, we can observe that the system performs similarly under different demand ratios of prescheduled and open access appointments. Figures 4.10 and 4.11 give comparisons of the system revenue improvement under different demand ratios.

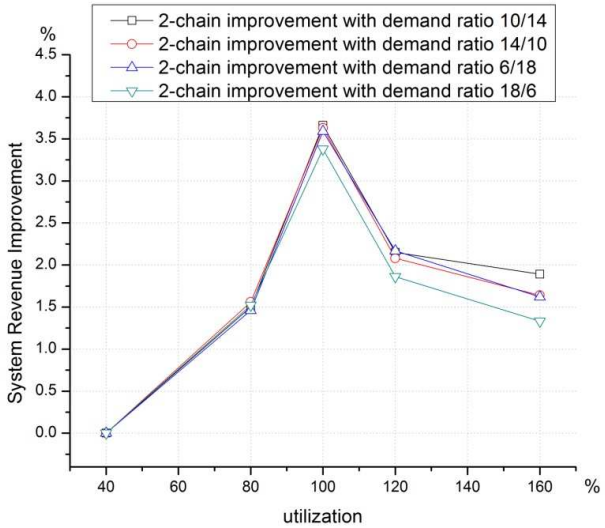


Figure 4.10 2-chain flexibility improvement under different demand ratios for all symmetric cases.

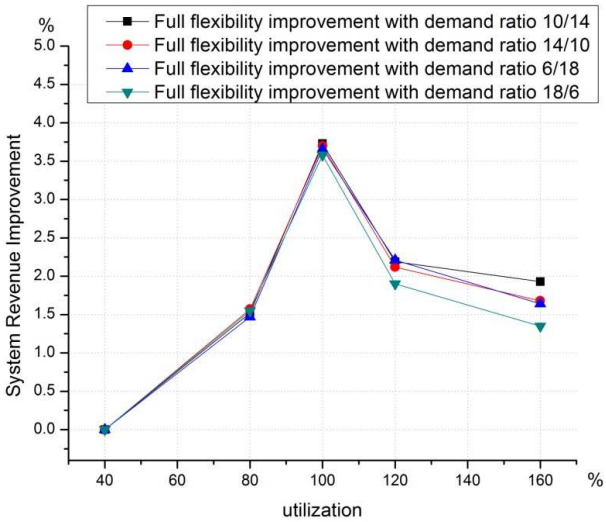


Figure 4.11 Full flexibility improvement under different demand ratios for all symmetric cases.

The system performance slightly downgrades when the demand ratio is 18/6, where the proportion of open access demand is reduced. Since flexibility is only implemented in the open access phase, the benefit of using flexibility to balance the demands among physicians has been reduced slightly due to lower in-bound open access demand.

Other system measures show the same properties. Although the absolute values of these metrics vary among different demand ratios due to the inequality of the revenues of the two types of demand, the improvements of flexible configurations are not very sensitive to the change of the demand ratio between prescheduled and open access appointments. The system uses the N_i^{p*} as a tool to accommodate as many demands as possible. In symmetric cases, the system performance mainly depends on the total demand, but doesn't rely on the demand ratio when the N_i^{p*} can be adjusted effectively.

4.3.2 N_i^{p*} of three physicians with symmetric demand distributions

For the primary care practice with dedicated flexibility, we can use equation (3.2.3) to find the optimal capacity allocation decision for each physician in a closed form expression. When the system involves three physicians or more, the stochastic integer programming model demonstrated in section 3.2.3 can be used to find the optimal capacity allocation between pre-scheduled and open access demands for the physicians in a practice. However, as we demonstrated, the computational effort required makes it impractical for practices with a large number of physicians. To reduce the computational burden and improve the search efficiency, we would like to identify underlying properties of the values of N_i^{p*} under flexible system configurations, and use the results of the dedicated case as initial references to guide the search, if possible.

Interestingly, from the data, we find that the values of N_i^{p*} for 2-chain and full flexibility are almost equal to each other in all levels of system utilization. Comparing the N_i^{p*} under flexible system configurations to the ones of the dedicated case we find the following:

- When the system is under-utilized, such as 40% utilization, the N_i^{p*} under flexible system configurations are approximately the same as the values of dedicated case.
- As demand grows toward a balanced system, the N_i^{p*} under the flexible configurations, in most cases, are greater than the ones in dedicated case.
- As the system becomes over-utilized, the N_i^{p*} under the flexible configurations, in most cases, are smaller than those in the dedicated case.

Figures 4.12, 4.13 and 4.14 show the distributions of the differences between N_i^{p*} under flexible configurations and the ones in dedicated case in Symmetric Case 3 when the system is 40%, 80%, and 100% utilized respectively.

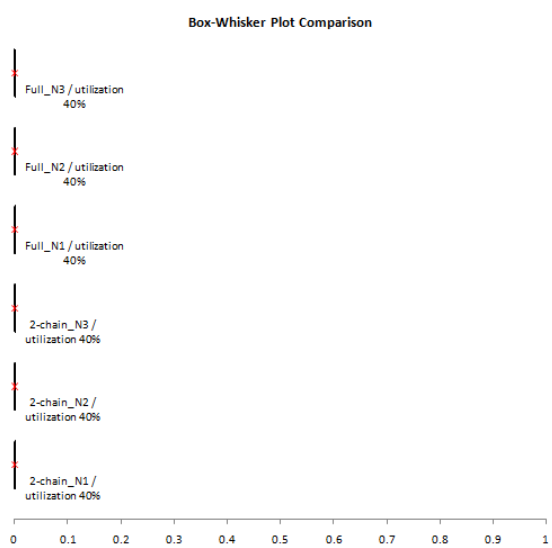


Figure 4.12 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 40% utilized in Symmetric Case 3 (6/18).

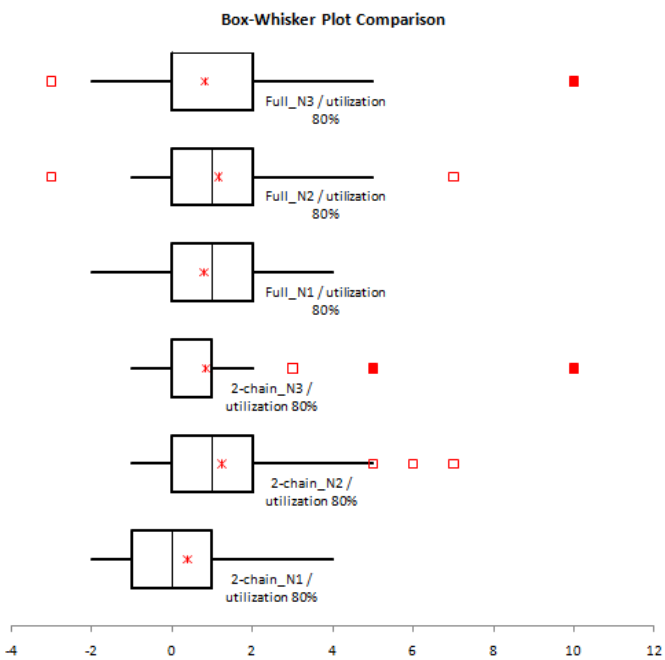


Figure 4.13 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 80% utilized in Symmetric Case 3 (6/18).

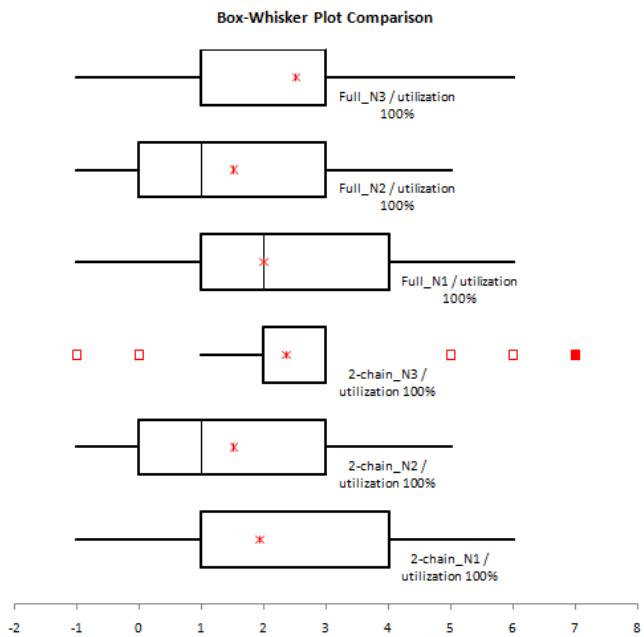


Figure 4.14 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 100% utilized in Symmetric Case 3 (6/18).

In Figure 4.12, when the system is quite under-utilized (40%), the N_i^{p*} of flexibility cases have the same values as the dedicated case. In Figures 4.13 to 4.14, as the demand and supply in the system become better balanced, we can see that the N_i^{p*} under the flexible configurations are greater than the ones of dedicated case from a statistical view, however, there are some "outliers" that behave conversely. We find that the values of N_i^{p*} for the 2-chain and full flexibility are close to each other in all levels of utilization. When the system is quite under-utilized, the values of N_i^{p*} calculated by the stochastic integer model are noticeably smaller than the theoretical values. This is due to the optimal gap set in cplex and "flat tail" effect shown in Figure 4.1 and 4.2. The model terminates the search of N_i^{p*} when it reaches the optimal gap. And when the system is fulfilled with more demands, the N_i^{p*} values become the same as the theoretical results.

Figure 4.15 and 4.16 show the distributions of the differences between N_i^{p*} under flexible configurations and the ones in dedicated case in Symmetric Case 3 when the system is 120% and 160% utilized respectively.

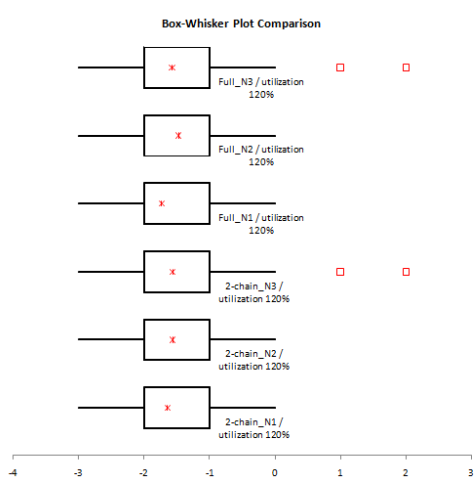


Figure 4.15 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 120% utilized in Symmetric Case 3 (6/18).

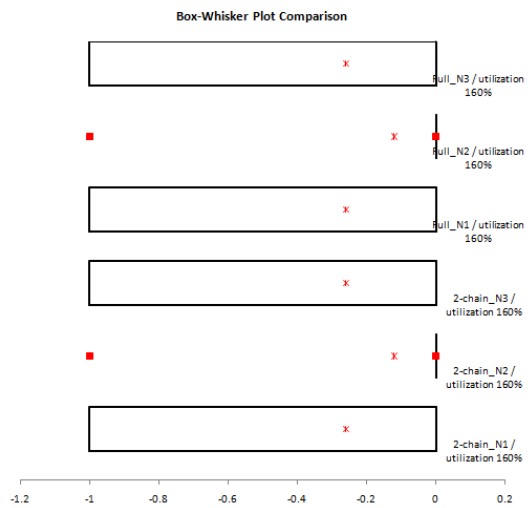


Figure 4.16 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 160% utilized and in Symmetric Case 3 (6/18).

In Figures 4.15 and 4.16, we can observe that when the system goes from balanced stage to over-utilized, the N_i^{p*} of flexibility cases are statistically smaller than the ones of dedicated case, and the "outliers" are negligible.

Figures 4.17, 4.18, 4.19, 4.20 and 4.21 give another instance of the directional structure of N_i^{p*} under flexible configurations in Symmetric Case 2 (14/10).

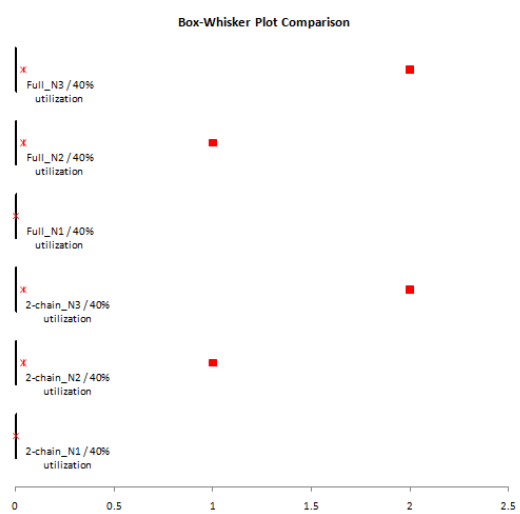


Figure 4.17 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 40% utilized in Symmetric Case 2 (14/10).

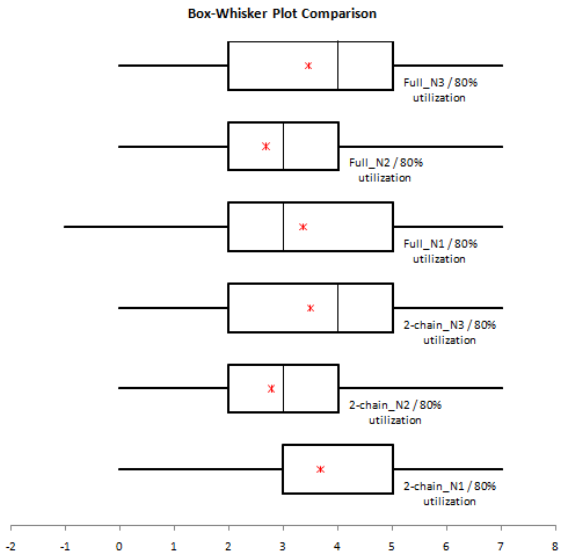


Figure 4.18 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 80% utilized in Symmetric Case 2 (14/10).

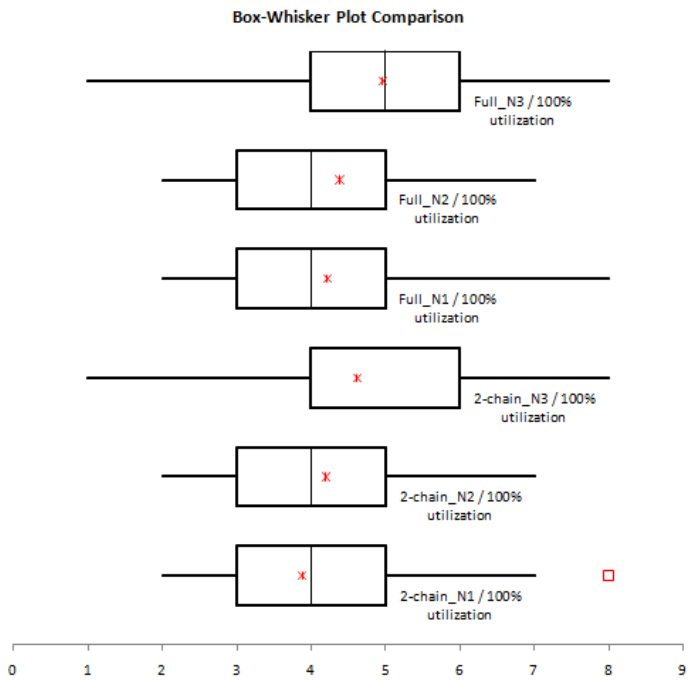


Figure 4.19 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 100% utilized in Symmetric Case 2 (14/10).

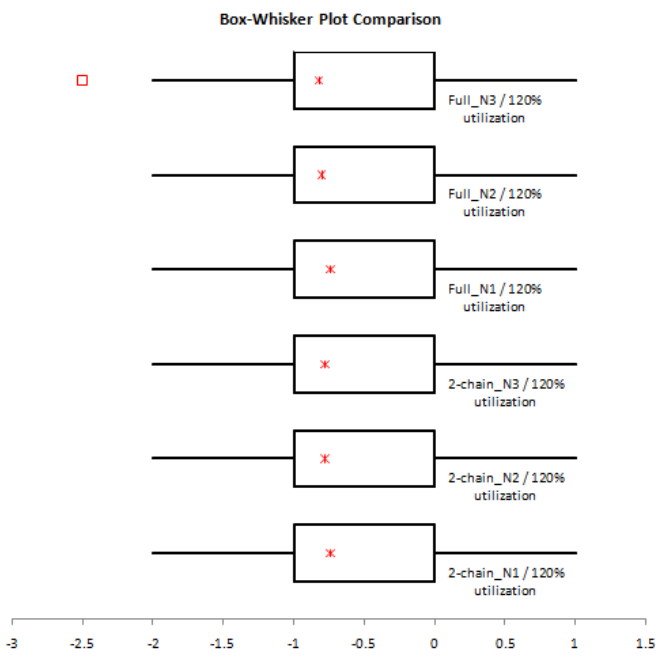


Figure 4.20 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 120% utilized in Symmetric Case 2 (14/10).

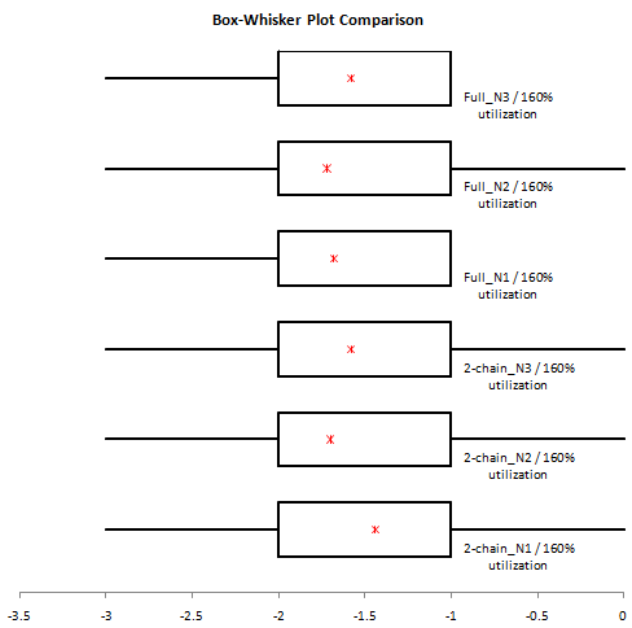


Figure 4.21 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 160% utilized in Symmetric Case 2 (14/10).

In summary, the directional structure of N_i^{p*} holds when the system is very under-/over-utilized, but is not strongly conclusive enough when the system approaches the balanced situation from both directions. It is possible that this loosely directional structure of the optimal solution could save the computational efforts for capacity allocation problem in our stochastic integer programming approach. It can be used as a heuristic, but not a firm property.

4.3.3 Results for three physicians with asymmetric demand distributions

Asymmetric Case 1. Table 4.16 summarizes the assumptions used in the Asymmetric Case 1 for three physicians with asymmetric demand distributions. Although each physician has different demand rates, the expected demand and available capacity for each physician are balanced, which means, each physician is equally utilized.

Physician capacity	24
Number of physicians in practice	3
Scenarios for each replication	1000
Number of replications	50
Revenue of seeing one pre-scheduled demand	0.75
Revenue of seeing one owned open access demand	0.90
Revenue of seeing one diverted open access demand	0.85
Mean demand rate for pre-scheduled appointments	[6, 10, 14]
Mean demand rate for open access appointments	[18, 14, 10]
Relative MIP tolerance gap	0.01%

Table 4.16 Assumptions for 3 physicians with asymmetric demand distributions in Asymmetric Case 1.

Tables 4.17, 4.18, and 4.19 demonstrate the measurements for 2-chain flexibility, full flexibility and dedicated in terms of system revenue, timely access rate and continuity rate in Asymmetric Case 1.

System Revenue Utilization	40%	80%	100%	120%	160%
2-chain	24.305	47.5985	57.10803	59.93379	62.12353
Full Flex	24.305	47.6065	57.14862	59.95717	62.14829
Dedicated	24.3048	46.8888	55.1161	58.6715	60.99828
2-chain vs Dedicated	0.00%	1.51%	3.61%	2.15%	1.84%
Full vs Dedicated	0.00%	1.53%	3.69%	2.19%	1.89%

Table 4.17 Measurements of system revenue with asymmetric demands in Asymmetric Case 1.

Timely Access Rate Utilization	40%	80%	100%	120%	160%
2-chain	100%	99.87%	95.29%	81.96%	62.09%
Full Flex	100%	99.87%	95.30%	81.94%	62.07%
Dedicated	100%	98.38%	91.81%	80.66%	61.66%
2-chain vs Dedicated	0.00%	1.52%	3.79%	1.61%	0.70%
Full vs Dedicated	0.00%	1.52%	3.79%	1.60%	0.66%

Table 4.18 Measurements of timely access rate with asymmetric demands in Asymmetric Case 1.

Continuity Rate Utilization	40%	80%	100%	120%	160%
2-chain	100%	98.23%	95.33%	97.02%	96.73%
Full Flex	100%	98.51%	96.43%	97.66%	97.35%
Dedicated	100%	100.00%	100.00%	100.00%	100.00%
2-chain vs Dedicated	0.00%	-1.77%	-4.67%	-2.98%	-3.27%
Full vs Dedicated	0.00%	-1.49%	-3.57%	-2.34%	-2.65%

Table 4.19 Measurements of continuity rate with asymmetric demands in Asymmetric Case 1.

If we make a comparison of the results between asymmetric case 1 and symmetric cases (notice that in all symmetric cases, each physician is equally utilized), we will find that the corresponding measurements are approximately the same, which means, the system is insensitive to the demand distributions among physicians when each physician has balanced/enough capacity to meet expected demands.

Asymmetric Case 2. To study how the system performs when each physician is unequally utilized, we test another case that one physician is under-utilized, the other one is balanced and the third physician is over utilized. Table 4.20 summarizes the assumptions used in the Asymmetric Case 2 for three physicians with asymmetric demand distributions.

Physician capacity	24
Number of physicians in practice	3
Scenarios for each replication	1000
Number of replications	50
Revenue of seeing one pre-scheduled demand	0.75
Revenue of seeing one owned open access demand	0.90
Revenue of seeing one diverted open access demand	0.85
Mean demand rate for pre-scheduled appointments	[6, 8, 10]
Mean demand rate for open access appointments	[12, 16, 20]
Relative MIP tolerance gap	0.1%

Table 4.20 Assumptions for 3 physicians with asymmetric demand distributions in Asymmetric Case 2.

In this case, the first physician is 75% utilized, the second physician is 100% utilized, and the third one is 125% over-utilized. Tables 4.21, 4.22, and 4.23 demonstrate the

measurements for 2-chain flexibility, full flexibility and dedicated in terms of system revenue, timely access rate and continuity rate in Asymmetric Case 2.

System Revenue Utilization	40%	80%	100%	120%	160%
2-Chain	23.83159	49.12353	57.86721	60.59101	63.42315
Full Flex	23.83159	49.13562	57.92525	60.63111	63.46042
Dedicated	23.82978	47.31722	53.80867	57.57554	62.0599
2-Chain vs Dedicated	0.01%	3.82%	7.54%	5.24%	2.20%
Full vs Dedicated	0.01%	3.84%	7.65%	5.31%	2.26%

Table 4.21 Measurements of system revenue with asymmetric demands in Asymmetric Case 2.

Timely Access Rate Utilization	40%	80%	100%	120%	160%
2-Chain	100.00%	99.80%	95.25%	82.73%	61.60%
Full Flex	100.00%	99.79%	95.26%	82.72%	61.59%
Dedicated	99.99%	96.06%	87.97%	78.11%	61.11%
2-Chain vs Dedicated	0.01%	3.89%	8.28%	5.90%	0.82%
Full vs Dedicated	0.01%	3.88%	8.29%	5.89%	0.80%

Table 4.22 Measurements of timely access rate with asymmetric demands in Asymmetric Case 2.

Continuity Rate Utilization	40%	80%	100%	120%	160%
2-Chain	99.99%	95.74%	90.95%	92.90%	95.08%
Full Flex	99.99%	96.28%	92.62%	93.97%	96.06%
Dedicated	100.00%	100.00%	100.00%	100.00%	100.00%
2-Chain vs Dedicated	-0.01%	-4.26%	-9.05%	-7.10%	-4.92%
Full vs Dedicated	-0.01%	-3.72%	-7.38%	-6.03%	-3.94%

Table 4.23 Measurements of continuity rate with asymmetric demands in Asymmetric Case 2.

Compare to the results in Asymmetric Case 1, we can see that the flexible configurations gain more improvement when each physician is differently utilized, which means, the flexibility system is more effective in a practice when the utilizations among physicians are unequal or unbalanced, especially some physicians are over-utilized. Figures 4.22 and 4.23 show the comparison between Asymmetric Case 1 and 2 in terms of system revenue and timely access improvement of flexible configurations. Figure 4.24 compares the continuity detriment between Asymmetric Case 1 and 2, we can see that a better system performance comes with a higher patient diversion rate.

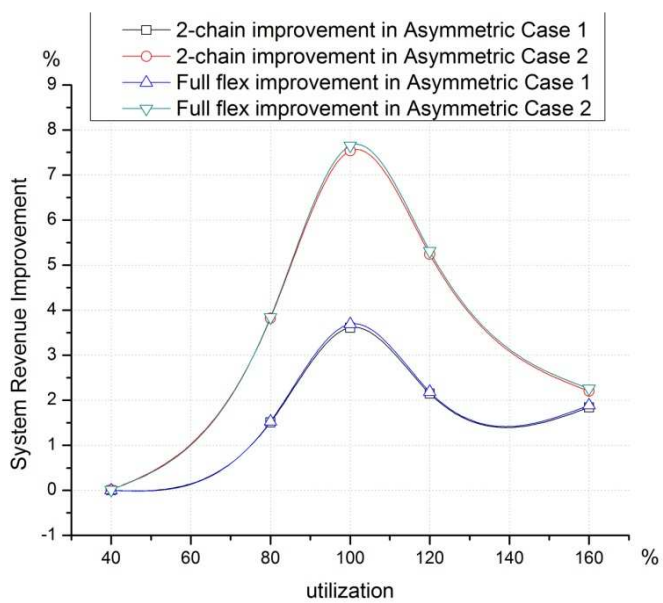


Figure 4.22 System revenue comparison between Asymmetric Case 1 and 2 for flexible configurations.

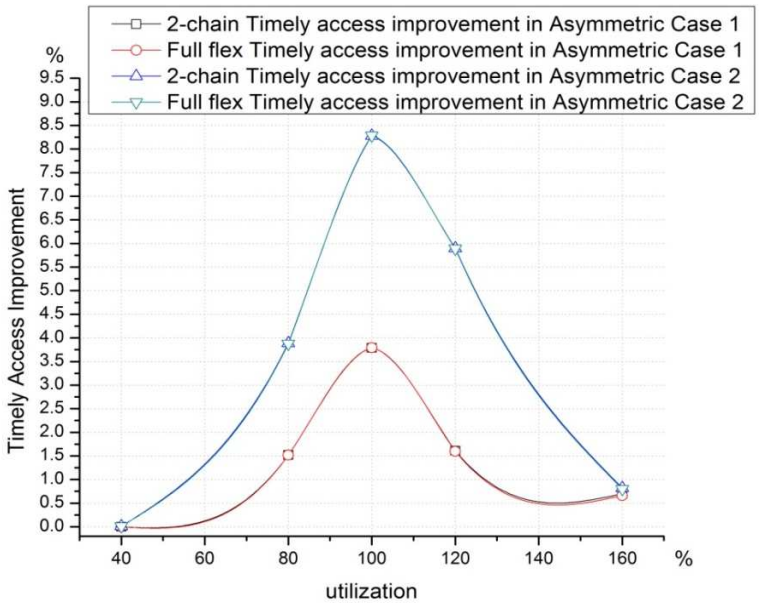


Figure 4.23 Timely access comparison between Asymmetric Case 1 and 2 for flexible configurations.

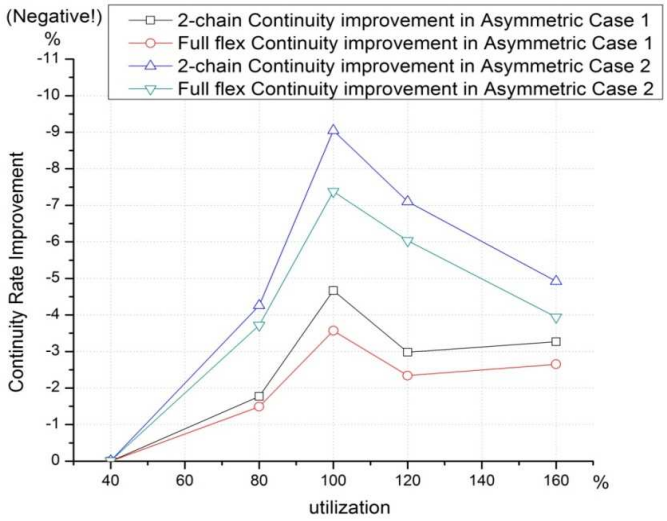


Figure 4.24 Continuity comparison between Asymmetric Case 1 and 2 for flexible configurations.

4.3.4 N_i^{p*} of three physicians with asymmetric demand distributions

When the demands are asymmetrically distributed and each physician has different utilization, for instance, in Asymmetric Case 2, from Figures 4.25 to 4.29, we can see that the structure of optimal solution we discussed in section 4.3.2 becomes worse for the asymmetric demand distributions. In under-utilized circumstances, the N_i^{p*} of flexibility cases are statistically equal or greater than the ones of dedicated case, but come with more counter examples; when the system goes to over-utilized, the N_i^{p*} of flexibility cases become smaller than the values of dedicated case, but don't hold for all cases. For instance, in 120% utilization, the N_3^{p*} is greater than the value of dedicate case. This is due to fact that the third physician is always over-utilized (125% utilized), and in a over-utilized configuration (120% utilization), the open access demand is so overwhelmed that the third physician in the dedicate case has to assign all the capacity for the open access demand and the N_3^{p*} becomes zero. However, with flexible configuration, the system has "extra" ability to accommodate the open access demands without the need to allocate all capacity to open access appointments.

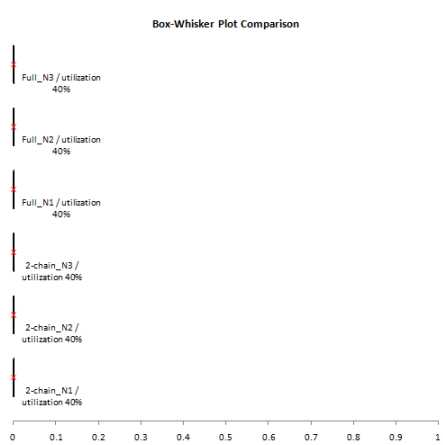


Figure 4.25 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 40% utilized in Asymmetric Case 2.

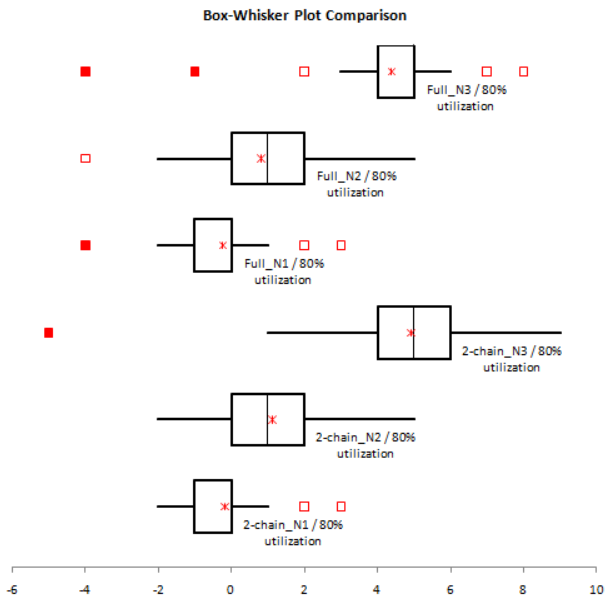


Figure 4.26 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 80% utilized in Asymmetric Case 2.

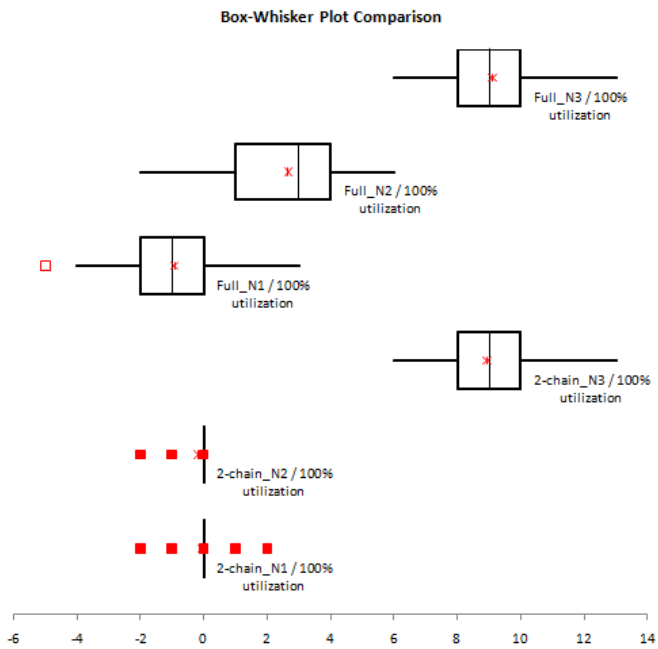


Figure 4.27 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 100% utilized in Asymmetric Case 2.

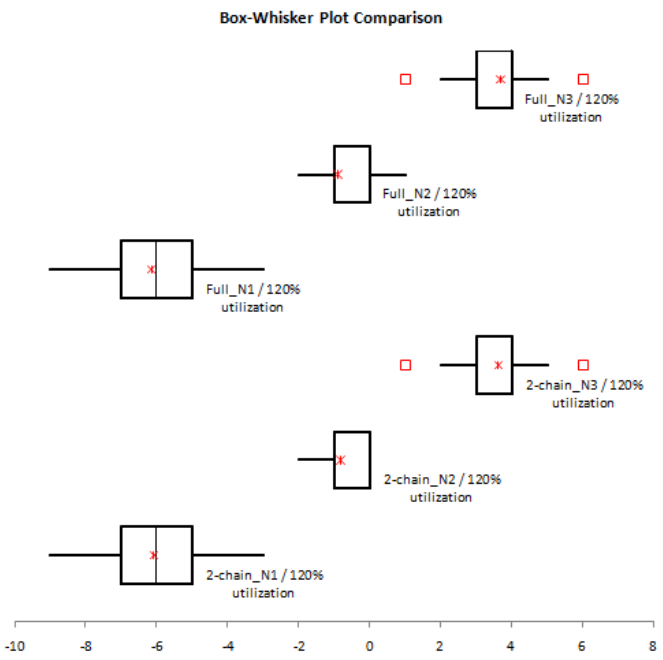


Figure 4.28 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 120% utilized in Asymmetric Case 2.

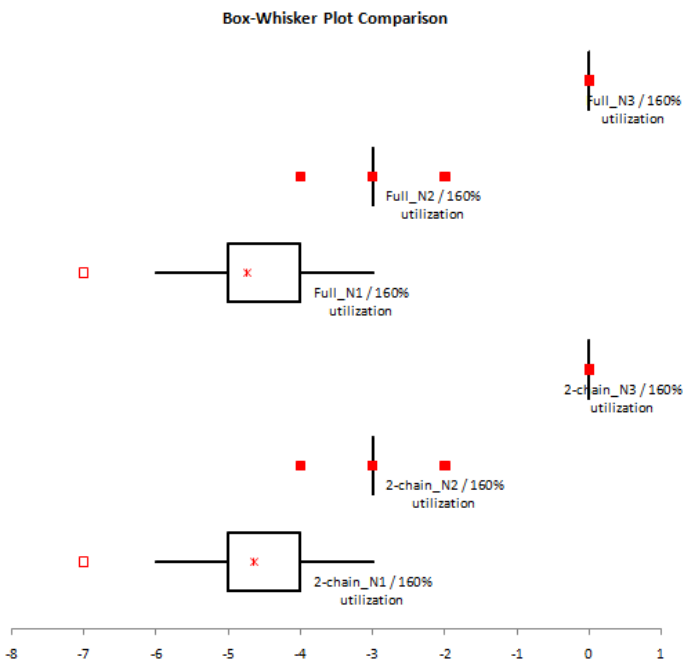


Figure 4.29 Distributions of the differences of Nps between flexible configurations and dedicated case when the system is 160% utilized in Asymmetric Case 2.

An explanation for this structure of optimal solution is that since the revenue of satisfying an open access demand is higher than meeting one pre-scheduled appointment, compared with the dedicated case, the system with flexibility will be more confident and capable of absorbing open access demands. In the balanced or under-utilized situations, the possibility that the open access demands will overflow the available capacities is not very high, therefore, the system will feel more "safe" to reserve more slots for pre-scheduled demands compared with the dedicated case who lacks the flexibility to deal with the occasional overflow of open access demand. By contrast, when over-utilized, the system with flexibility will struggle to meet all the open access demands. Since satisfying a open access demand will generates a higher revenue, the system will be more "greedily" to capture the open access demands, which means, the number of slots reserved for pre-scheduled demands will be reduced, compared with the dedicated case.

Again, this "directional" structure is currently not a very robust guideline for conducting a quick search of N_i^{p*} by using the values of the dedicated case as references. A further study is needed to validate the structure on a more comprehensive basis.

4.3.5 Trends in the total N_i^{p*} values for all three physicians

Figure 4.30 shows the average N_i^{p*} values for the entire clinic (that is for all the physicians) under different utilizations and for the three flexibility configurations. The trends observed by looking at the individual physicians' N_i^{p*} values are summarized concisely here. In general, for the highly underutilized case, the total N_i^{p*} values for the dedicated and flexibility configurations, not surprisingly, are identical. Since the demands

are so low, the N_i^{p*} values are likely to be fairly robust at this level. As the utilization increases to 80% and 100%, the clinic as a whole reserves more prescheduled appointments in the flexibility cases than the dedicated case. This is a direct consequence of flexibility: open access appointments can be absorbed effectively by pooling the (lower) capacity of all physicians together. In the high utilization cases (120% and 160%), there is enough demand for the high revenue open access appointments for the total N_i^{p*} of the clinic to be lower. The flexibility cases have a lower total N_i^{p*} value than the dedicated case, reserving more capacity for open access, since there is a higher probability of using the additional capacity when physicians are able to see each others' open access appointments.

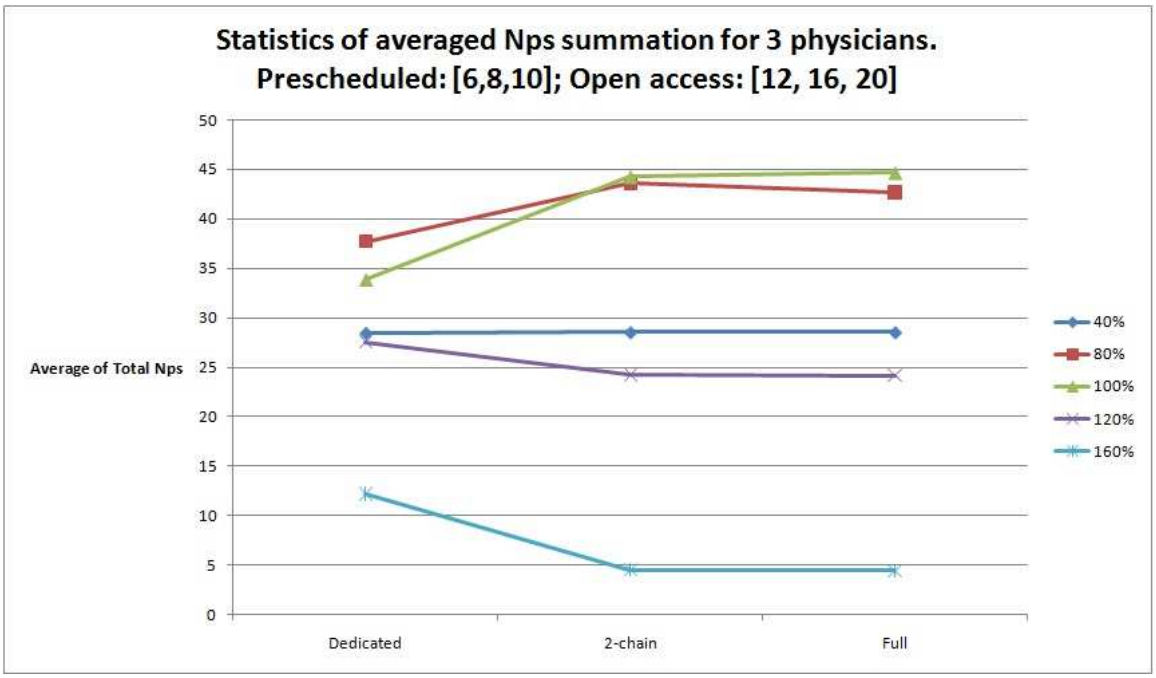


Figure 4.30 Average Nps values for three physicians with asymmetric demand distributions.

4.4 Value of flexibility in a practice with six physicians

In larger practices (academic practices for instance), there are typically more than ten physicians working at a clinic. But they often subdivide their practices into smaller groups or teams. The number of such physicians in a group may be up to five or six. We will emphasize on studying the value of flexibility for six physicians to gain insights about the system performance in the practice.

4.4.1 Results for six physicians with symmetric demand distributions

Table 4.24 summarizes the assumptions used in the study of six physicians with symmetric demand distributions.

Physician capacity	24
Number of physicians in practice	6
Scenarios for each replication	1000
Number of replications	50
Revenue of seeing one pre-scheduled demand	0.75
Revenue of seeing one owned open access demand	0.90
Revenue of seeing one diverted open access demand	0.85
Mean demand rate for pre-scheduled appointments	[10,10,10,10,10,10]
Mean demand rate for open access appointments	[14,14,14,14,14,14]
Relative MIP tolerance gap	0.5%

Table 4.24 Assumptions for 6 physicians with symmetric demand distributions.

Tables 4.25, 4.26, and 4.27 illustrate the measurements for 2-chain flexibility, full flexibility and dedicated case for a practice with six physicians.

System Revenue Utilization	60%	80%	100%	120%	140%
2-chain	70.16151	95.23072	115.5911	120.5238	123.1126
Full Flex	70.16153	95.26475	115.9736	120.747	123.3331
Dedicated	70.11055	93.71649	110.1977	117.2515	120.1987
2-chain vs Dedicated	0.07%	1.62%	4.89%	2.79%	2.42%
Full vs Dedicated	0.07%	1.65%	5.24%	2.98%	2.61%

Table 4.25 Measurement of system revenue for 6 physicians (symmetric).

Timely Access Rate Utilization	60%	80%	100%	120%	140%
2-chain	100.00%	99.99%	96.65%	82.37%	70.23%
Full Flex	100.00%	99.99%	96.68%	82.29%	70.18%
Dedicated	99.93%	98.39%	91.79%	80.72%	69.49%
2-chain vs Dedicated	0.07%	1.63%	5.29%	2.05%	1.06%
Full vs Dedicated	0.07%	1.63%	5.32%	1.95%	0.99%

Table 4.26 Measurement of timely access rate for 6 physicians (symmetric).

Continuity Rate Utilization	60%	80%	100%	120%	140%
2-chain	99.93%	97.83%	90.13%	93.73%	93.57%
Full Flex	99.93%	98.40%	95.05%	96.52%	96.47%
Dedicated	100.00%	100.00%	100.00%	100.00%	100.00%
2-chain vs Dedicated	-0.07%	-2.17%	-9.87%	-6.27%	-6.43%
Full vs Dedicated	-0.07%	-1.60%	-4.95%	-3.48%	-3.53%

Table 4.27 Measurement of continuity rate for 6 physicians (symmetric).

If we compare these measures to the associated values of three physicians (Symmetric Case 1), we can see that the improvement of flexibility configuration is higher in a practice with a larger number of physicians. Figures 4.31 and 4.32 give the comparisons of system performance between three physicians and six physicians.

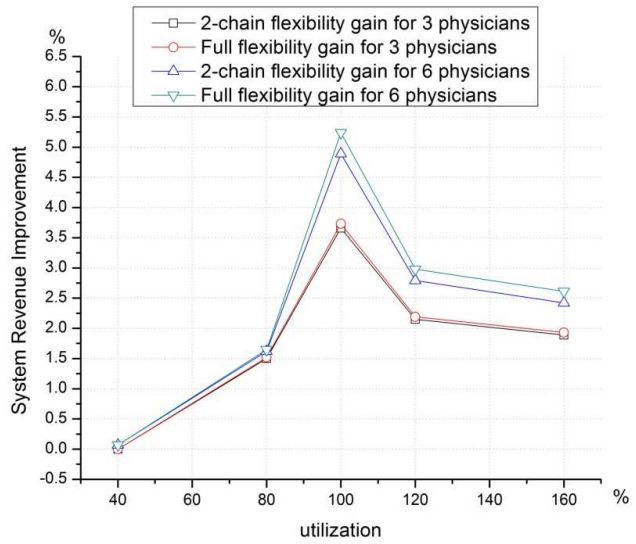


Figure 4.31 Comparison of system revenue improvement between 3 and 6 physicians.

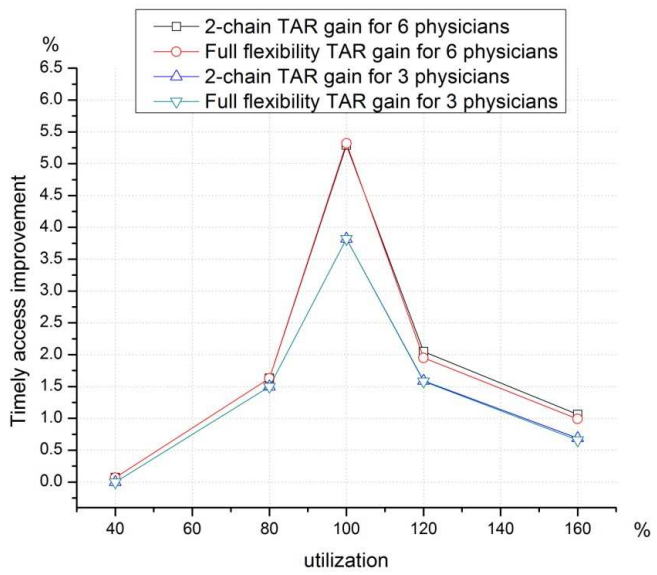


Figure 4.32 Comparison of timely access improvement between 3 and 6 physicians.

One thing deserves an attention is that the better improvements come with a higher diversion rate for six physicians, as shown in Figure 4.33.

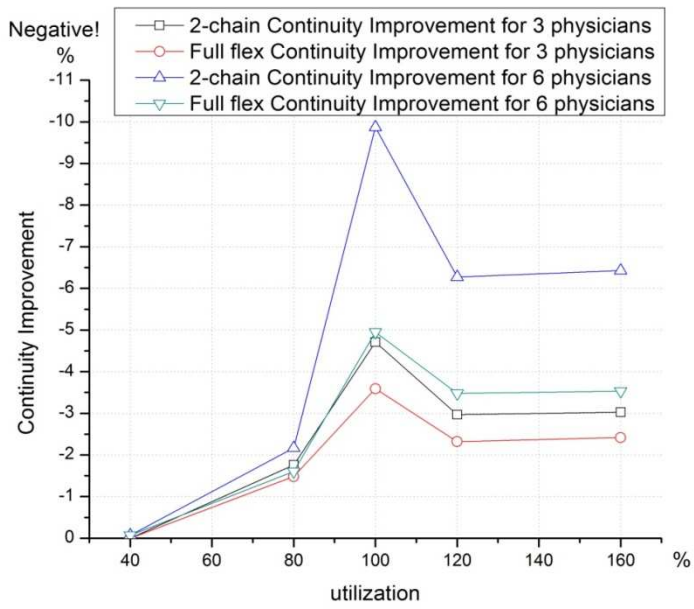


Figure 4.33 Comparison of continuity improvement between 3 and 6 physicians.

4.4.2 Results for six physicians with asymmetric demand distributions

Table 4.28 summarizes the assumptions for six physicians with asymmetric demand distributions.

Physician capacity	24
Number of physicians in practice	6
Scenarios for each replication	1000
Number of replications	50
Revenue of seeing one pre-scheduled demand	0.75
Revenue of seeing one owned open access demand	0.90
Revenue of seeing one diverted open access demand	0.85
Mean demand rate for pre-scheduled appointments	[6,10,14,6,10,14]
Mean demand rate for open access appointments	[18,14,10,18,14,10]
Relative MIP tolerance gap	0.5%

Table 4.28 Assumptions for 6 physicians with asymmetric demand distributions.

Tables 4.29, 4.30, and 4.31 give the measurements for 2-chain flexibility, full flexibility and dedicated case in terms of system revenue, timely access rate and continuity rate for six physicians with asymmetric demands.

System Revenue Utilization	60%	80%	100%	120%	140%
2-chain	72.06577	95.21054	115.5284	120.5605	122.8373
Full Flex	72.06373	95.24583	115.984	120.7883	123.0649
Dedicate	71.99092	93.70374	110.2392	117.3445	119.9602
2-chain vs Dedicated	0.10%	1.61%	4.80%	2.74%	2.40%
Full vs Dedicated	0.10%	1.65%	5.21%	2.93%	2.59%

Table 4.29 Measurement of system revenue for 6 physicians (asymmetric).

Timely Access Rate Utilization	60%	80%	100%	120%	140%
2-chain	99.99%	99.96%	96.57%	82.27%	70.92%
Full Flex	99.98%	99.96%	96.69%	82.21%	70.84%
Dedicated	99.89%	98.36%	91.81%	80.67%	70.10%
2-chain vs Dedicated	0.10%	1.63%	5.19%	1.98%	1.17%
Full vs Dedicated	0.10%	1.63%	5.31%	1.91%	1.06%

Table 4.30 Measurement of timely access rate for 6 physicians (asymmetric).

Continuity Rate Utilization	60%	80%	100%	120%	140%
2-chain	99.89%	97.85%	90.40%	93.76%	93.66%
Full Flex	99.89%	98.40%	95.05%	96.52%	96.45%
Dedicated	100.00%	100.00%	100.00%	100.00%	100.00%
2-chain vs Dedicated	-0.11%	-2.15%	-9.60%	-6.24%	-6.34%
Full vs Dedicated	-0.11%	-1.60%	-4.95%	-3.48%	-3.55%

Table 4.31 Measurement of continuity rate for 6 physicians (asymmetric).

A further look at the results in Table 4.29, 4.30 and 4.31, pluses a comparison between the corresponding values in Table 4.25, 4.26 and 4.27, deliver the same message: the system yields almost the same performance with symmetric and asymmetric demands when each physician is equally utilized and there is no physician who is obviously over-utilized. The improvement of flexibility is higher in a practice with a larger number of physicians. The loss of continuity in 2-chain flexibility is due to, in reality, a single patient redirection to an available physician, which can be made directly under full flexibility, may require redirecting several patients along the 2-chain if the initial patient's panel and available physician involved are not connected.

4.5 Conclusion

In this chapter, we use quantitative methods to demonstrate the value of flexibility for single physician, two physicians, three physicians and six physicians with symmetric and asymmetric demand distributions. Introducing flexibility is obviously always improving the performance of our tested system, even with a 5% cost for using flexibility links (i.e., the revenue of seeing a patient from owned open access panel is 0.9, but meeting a patient from another physician's open access panel is 0.85), the system revenue can be increased by up to 7.5%. With more physicians, flexibility becomes more beneficial, this can be found by comparing the corresponding results of three and six physicians. Our two-stage stochastic integer programming model can be used for the analysis of a systems with a larger amount of physicians.

Not surprisingly, the system achieves the maximum gain when the demand and supply are balanced (100% utilization). For under-/over-utilized systems, while still yielding

improvements, flexibility is less beneficial. In all cases, the 2-chain flexibility has a similar performance compared with full flexibility in aspects of system revenue, timely access rate, and interestingly, it has a higher diversion rate than full flexibility. As explained in section 4.3, in the aggregate demand setting captured by our two-stage stochastic integer programming approach, the patient allocation is only performed after the full system demand is known.

An important observation is that, by using the loosely directional structure of the optimal solution of flexibility, the computational efforts of searching optimal capacity allocation decision might be reduced significantly by using the values of the dedicated case as references.

CHAPTER 5

IMPLICATIONS FOR PRACTICE

We study primary care practices with three physicians or more by using the two-stage stochastic integer programming model developed in section 3.2.3. The performance of the flexibility configurations studied and the structure of the optimal solution depend on several parameters: the revenues associated with satisfying each type of demand, open access or prescheduled; the cost of a patient diversion; and the demand distributions. Our goal in this thesis was to explore the general value of flexibility and the factors that may affect it. For that purpose, we took some representative parameter values, which are justified below.

- **Revenues associated with satisfying demands.** In our numerical tests, we consider the revenues of scheduling patients to see a physician as the typical show rates for prescheduled and open access demand. Therefore, the system revenue actually stands for the expected total number of patients that the system will satisfy, given that some scheduled patients will not show up. To effectively capture the revenue improvement gained by introducing flexibility into a clinic practice, a monetary value of seeing prescheduled and same-day appointments could be used in our model. The patient no-show rate is typically a key factor and can be estimated from the historical data of the clinic. The overall revenue associated with each patient type, however, needs to be investigated and better understood.
- **The cost of a patient diversion.** We add a 5% cost to a patient diversion to encourage patient-physician continuity in the system. However, in a real clinic

practice, the diversion cost is very hard to estimate and quantify. Physicians tend to spend more time on examining the history of a diverted patient they are not familiar with. The system revenue will be reduced accordingly; not to mention the increased chance of misdiagnosis and patient's dissatisfaction. To evaluate the influence of patients diversion on the system performance, a clinic practice needs to capture the diversion cost quantitatively. A possible way is to estimate the average time that a physician spends on a patient from his/her own panel, and compare it with the average time that the physician takes on a patient from other panels. The difference of the time is the reflection of the increased operation cost. This will make the diversion cost easier to understand and more convincing for the clinic management team. It is important to note that the diversion cost may depend on how we manage the flexibility in the system. In a two-chain, each patient can only see two physicians and each physician only receives patients from two panels. The loss in familiarity is going to be minimal, as compared to a large practice with full flexibility where patients may see any of the doctors.

- **The demand distributions** vary from clinic to clinic, therefore, the best way to implement the flexibility modeling approach on a practice is to use the real data estimated from historical records as the inputs. Since each clinic focuses on different types of patients in different regional areas with different physical capacity, the exact benefit of flexibility will accordingly vary. Out study however provides insight on the general value of flexibility for primary care practices and how it varies with some characteristics of the demand distributions.

Physicians are inherently flexible to see each other's patients. In contrast with manufacturing, there is no cost associated with "installing" flexibility; but flexibility needs to be implemented and managed. In most clinics, a certain level of the flexibility, especially the full flexibility, has already been implemented in practice. The patient usually asks to see his/her own physician; if the physician is not available, the patient will be advised to see any other physician in the practice. In our study, we find that the 2-chain flexibility yields nearly as much benefit as full flexibility, but with reduced complexity. A natural question arises: how to implement the 2-chain or other flexibilities in the practice? That is, how do we decide which two physicians should be connected? The answer to this question depends on lots of factors, but an easy and effective approach is to connect physicians with different utilizations, such as over-utilized to under-utilized physicians, to make the system more balanced. It is important to note, however, that the connection configuration heavily relies on the clinic's working structure and policy, as well as its daily operational process. It might be possible that a clinic cannot be configured as a particular flexible system we discussed.

In summary, our models, which are developed for the primary care practices, focus mainly on the theoretical aspects of allowing flexibility in appointment scheduling. To more accurately evaluate the performance of flexible configurations, we need to test them in a real clinic practice, gather feedback from physicians, and more importantly, work with them to address the issues that may impede their widespread implementation.

CHAPTER 6

CONCLUSIONS

To find the optimal capacity allocation decision between pre-scheduled demands and open access demands for physicians in the primary care practice, we develop formulations and find closed form solutions for individual, dedicated physicians and for two physicians with flexibility links; for multiple physicians with different levels of flexibility, we use a stochastic integer programming approach to provide the optimal capacity allocation decision for any number of physicians in a practice and with any flexibility configuration.

The results of our study confirm that introducing flexibility yields benefits even if there is a cost for using flexibility links. Similarly, we find that the benefits are the highest when the system is balanced, and decreasing for higher or lower levels of system utilization. The 2-chain flexibility yields almost all the benefits of full flexibility in terms of system revenue and timely access rate, but comes with a higher rate of patient diversion; due to the limited outbound links in the 2-chain system, more "jumps" may be required to shift and absorb the demands.

By using the stochastic integer programming model, we investigate the three- and six-physician cases. As we expected, flexibility is more beneficial with increased number of physicians. Our model is not sensitive to the change of demand ratio between prescheduled and open access demands when physicians are equally utilized. The flexible configurations become more beneficial when physicians are unequally utilized.

Our computational experiments show that the optimal capacity allocation decision for flexibility configuration yields a *directional* structure in some cases: The optimal

capacity to reserve for prescheduled appointments under flexible configurations tends to be higher when the system is under-utilized and lower when it is over-utilized, as compared to the values gained from the dedicated case. This interesting characteristic, which also needs further investigation, might reduce the computational efforts and make the search be conducted in a small fraction of the feasible space.

CHAPTER 7

FUTURE WORK

While we developed the closed form formulation and stochastic integer programming model to investigate the basic properties of physician flexibility and performed analysis of the structure of optimal capacity allocation decision, there are still open questions that deserve attention in future research.

- We assigned a 5% cost for using flexibility links in our analysis. A more comprehensive study with different levels of cost, such as 0%, 10% and 15%, needs to be investigated in future.
- The revenues of meeting one pre-scheduled demand and seeing an open access patient are based on the typical show rates for each access scheme. We wonder how the change of these revenues will influence the allocation decision and the solution structure.
- The demand rates need to be estimated from historical data. A case study based on a real clinic practice will be more convincing to demonstrate the benefits of introducing flexibility.
- Though deduced from a reasonable explanation and confirmed with experimental results, the directional or monotonic structure of the optimal allocation solution of flexibility needs to be validated on a more comprehensive basis. And a new algorithm that uses the values gained from dedicated case as a starting point and searches the solution only in one direction needs to answer the following question: how many steps we have to go further to achieve an acceptably near optimal

solution while not increase the complexity noticeably. In other words, what is the best point that to stop the search.

APPENDIX A

THEOREMS PROOF

A.1 Proof of Theorem 1

For any individual physician *i*, the expected cost of missing pre-scheduled demand $EC_i^p(N_i^p)$ is non-increasing with N_i^p, which means $EC_i^p(N_i^p+1) \le EC_i^p(N_i^p)$ for any $N_i^p \in \{0,1,2,.,N-1\}$, and the expected cost of missing open access demand $EC_i^o(N_i^p)$ is non-decreasing with N_i^p, that is, $EC_i^o(N_i^p+1) \ge EC_i^o(N_i^p)$ for any $N_i^p \in \{0,1,2,.,N-1\}$.

For a given N_i^p, if N_i^p increases by 1, the reduced expected cost of missing pre-scheduled demand is equal to $EC_i^p(N_i^p) - EC_i^p(N_i^p+1)$, which is:

$$\begin{aligned}
EC_i^p(N_i^p) - EC_i^p(N_i^p+1) = & \sum_{d_i^p=N_i^p+1}^{\infty} C_p(d_i^p - N_i^p)\, p_i(d_i^p) - \\
& \sum_{d_i^p=N_i^p+2}^{\infty} C_p[d_i^p - (N_i^p+1)]\, p_i(d_i^p) \\
& = C_p[\sum_{d_i^p=N_i^p+1}^{\infty} d_i^p\, p_i(d_i^p) - \sum_{d_i^p=N_i^p+1}^{\infty} N_i^p\, p_i(d_i^p)] - \\
& C_p[\sum_{d_i^p=N_i^p+2}^{\infty} d_i^p\, p_i(d_i^p) - \sum_{d_i^p=N_i^p+2}^{\infty} (N_i^p+1)\, p_i(d_i^p)] \\
& = C_p[(N_i^p+1) \sum_{d_i^p=N_i^p+1}^{\infty} p_i(d_i^p) - N_i^p \sum_{d_i^p=N_i^p+1}^{\infty} p_i(d_i^p)] \\
& = C_p \sum_{d_i^p=N_i^p+1}^{\infty} p_i(d_i^p) \\
& = C_p[1 - F(N_i^p)]
\end{aligned} \tag{A.1}$$

And the increased expected cost of missing open access demand $EC_i^o(N_i^p+1) - EC_i^o(N_i^p)$ equals to:

$$
\begin{aligned}
EC_i^o\left(N_i^p+1\right)-EC_i^o\left(N_i^p\right)=&\sum_{d_i^p=0}^{N_i^p+1}p_i(d_i^p)\sum_{d_i^o=N-d_i^p+1}^{\infty}C_o\left[d_i^o-\left(N-d_i^p\right)\right]q_i\left(d_i^o\right)+\\
&[1-F_i\left(N_i^p+1\right)]\sum_{d_i^o=N-N_i^p}^{\infty}C_o\left[d_i^o-\left(N-d_i^p\right)\right]q_i\left(d_i^o\right)-\\
&\sum_{d_i^p=0}^{N_i^p}p_i(d_i^p)\sum_{d_i^o=N-d_i^p+1}^{\infty}C_o\left[d_i^o-\left(N-d_i^p\right)\right]q_i\left(d_i^o\right)-\\
&[1-F_i\left(N_i^p\right)]\sum_{d_i^o=N-N_i^p+1}^{\infty}C_o[d_i^o-\left(N-N_i^p\right)]q_i(d_i^o)\\
=&C_o p_i(N_i^p+1)\sum_{d_i^o=N-N_i^p}^{\infty}\left(d_i^o-N+N_i^p+1\right)q_i\left(d_i^o\right)-\\
&C_o p_i(N_i^p+1)\sum_{d_i^o=N-N_p^i}^{\infty}\left(d_i^o-N-N_i^p+1\right)q_i\left(d_i^o\right)+\\
&C_o\left[1-F\left(N_i^p\right)\right][1-\left(N-N_i^p\right)]\\
=&C_o\left[1-F\left(N_i^p\right)\right][1-\left(N-N_i^p\right)]
\end{aligned}
\tag{A.2}
$$

The optimal N_i^{p*} comes out when N_i^{p*} increases by 1, the marginal reduced cost of missing pre-scheduled demand should be less or equal to the marginal increased cost of missing open access demand, which means, N_i^{p*} should satisfy:

$$
EC_i^p\left(N_i^{p*}\right)-EC_i^p\left(N_i^{p*}+1\right)\le EC_i^o\left(N_i^{p*}+1\right)-EC_i^o(N_i^{p*})
\tag{A.3}
$$

Using the above derivations, we have:

$$
C_p\left[1-F\left(N_i^{p*}\right)\right]\le C_o\left[1-F\left(N_i^{p*}\right)\right]\left[1-\left(N-N_i^{p*}\right)\right]
\tag{A.4}
$$

That is:

$$
N_i^{p*}\ge N-\Phi_i^{-1}(1-\frac{C_p}{C_o})
\tag{A.5}
$$

And if N_i^{p*} decreases by 1, the marginal increased cost of missing pre-scheduled demand should be larger than the marginal decreased cost of missing open access demand, similarly, we get:

$$N_i^{p*} < N + 1 - \Phi_i^{-1}(1 - \frac{C_p}{C_o}) \tag{A.6}$$

Therefore:

$$N_i^{p*} = N - \Phi_i^{-1}(1 - \frac{C_p}{C_o}) \tag{A.7}$$

Proof done.

A.2 Proof of Theorem 3

For two physicians with partial flexibility, the total expected cost of missing pre-scheduled demands is equals to:

$$\begin{aligned} EC^p\left(N_1^p, N_2^p\right) &= EC^p\left(N_1^p\right) + EC_2^p\left(N_2^p\right) \\ &= \sum_{i=1}^{2} \sum_{d_i^p = N_i^p + 1}^{\infty} C_p(d_i^p - N_i^p) p_i(d_i^p) \end{aligned} \tag{A.8}$$

And the total expected cost of missing open access demands equals to:

$$\begin{aligned} & EC^o\left(N_1^p, N_2^p\right) = \\ & C_o\left[1 - F_1\left(N_1^p\right)\right]\left[1 - F_2\left(N_2^p\right)\right] \sum_{d^o = 2N - \left(N_1^p + N_2^p\right) + 1}^{\infty} \left[d^o - \left(2N - N_1^p - N_2^p\right)\right] q\left(d^o\right) + \\ & C_o[1 - F_2\left(N_2^p\right)] \sum_{d_1^p = 0}^{N_1^p} p_1(d_1^p) \sum_{d^o = 2N - \left(d_1^p + N_2^p\right) + 1}^{\infty} \left[d^o - \left(2N - d_1^p - N_2^p\right)\right] q\left(d^o\right) + \\ & C_o[1 - F_1\left(N_{12}^p\right)] \sum_{d_2^p = 0}^{N_2^p} p_2(d_2^p) \sum_{d^o = 2N - \left(d_2^p + N_1^p\right) + 1}^{\infty} \left[d^o - \left(2N - d_2^p - N_1^p\right)\right] q\left(d^o\right) + \end{aligned} \tag{A.9}$$

$$C_o \sum_{d_1^p=0}^{N_1^p} p_1(d_1^p) \sum_{d_2^p=0}^{N_2^p} p_2(d_2^p) \sum_{d^o=2N-\left(d_1^p+d_2^p\right)+1}^{\infty} [d^o - \left(2N - d_1^p - d_2^p\right)] q(d^o)$$

Where d^0 is the aggregated open-access demand.

Similar to the proof of Theorem 1, for a given N_1^p and N_2^p, if N_1^p increases by 1, the reduced total expected cost of missing pre-scheduled demand is equal to $EC^p(N_1^p, N_2^p) - EC^p(N_1^p + 1, N_2^p)$, which is:

$$EC^p\left(N_1^p, N_2^p\right) - EC^p\left(N_1^p + 1, N_2^p\right) = C_p[1 - F_1\left(N_1^p\right)] \tag{A.10}$$

And the increased total expected cost of missing open access demand $EC^o(N_1^p + 1, N_2^p) - EC^o(N_1^p, N_2^p)$ equals to:

$$\begin{aligned} &EC^o\left(N_1^p + 1, N_2^p\right) - EC^o\left(N_1^p, N_2^p\right) = \\ &C_o\left[1 - F_2\left(N_2^p\right)\right]\left[1 - F_1\left(N_1^p\right)\right]\left[1 - \Phi\left(2N - N_1^p - N_2^p - 1\right)\right] + \\ &[1 - F_1\left(N_1^p\right)] \sum_{d_2^p=0}^{N_2^p} p_2\left(d_2^p\right)[1 - \Phi\left(2N - N_1^p - d_2^p - 1\right)] \end{aligned} \tag{A.11}$$

If N_2^p increases by 1, the reduced total expected cost of missing pre-scheduled demand is equal to $EC^p(N_1^p, N_2^p) - EC^p(N_1^p, N_2^p + 1)$, which is:

$$EC^p\left(N_1^p, N_2^p\right) - EC^p\left(N_1^p, N_2^p + 1\right) = C_p[1 - F_2\left(N_2^p\right)] \tag{A.12}$$

The increased total expected cost of missing open access demand $EC^o(N_1^p, N_2^p + 1) - EC^o(N_1^p, N_2^p)$ equals to:

$$\begin{aligned} &EC^o\left(N_1^p, N_2^p + 1\right) - EC^o\left(N_1^p, N_2^p\right) = \\ &C_o\left[1 - F_1\left(N_1^p\right)\right]\left[1 - F_2\left(N_2^p\right)\right]\left[1 - \Phi\left(2N - N_1^p - N_2^p - 1\right)\right] + \\ &\left[1 - F_2\left(N_2^p\right)\right] \sum_{d_1^p=0}^{N_1^p} p_1\left(d_1^p\right)[1 - \Phi(2N - d_1^p - N_2^p - 1)] \end{aligned} \tag{A.13}$$

The optimal N_1^{p*} and N_2^{p*} come out when either N_1^{p*} or N_2^{p*} increases by 1, the marginal reduced total cost of missing pre-scheduled demand should be less or equal to the marginal increased total cost of missing open access demand, which means, N_1^{p*} should satisfy:

$$EC^p\left(N_1^{p*}, N_2^{p*}\right) - EC^p\left(N_1^{p*}+1, N_2^{p*}\right) \le EC^o\left(N_1^{p*}+1, N_2^{p*}\right) - EC^o\left(N_1^{p*}, N_2^{p*}\right) \tag{A.14}$$

which is:

$$\begin{aligned}\frac{C_p}{C_o} \le & [1-F_2(N_2^p)]\cdot[1-\Phi(2N-N_1^p-N_2^p-1)]+ \\ & \sum_{d_2^p=0}^{N_2^p} p_2(d_2^p)\cdot[1-\Phi(2N-N_1^p-d_2^p-1)]\end{aligned} \tag{A.15}$$

and similarly, N_2^{p*} should satisfy:

$$\begin{aligned}\frac{C_p}{C_o} \le & [1-F_1(N_1^p)]\cdot[1-\Phi(2N-N_1^p-N_2^p-1)]+ \\ & \sum_{d_1^p=0}^{N_1^p} p_1(d_1^p)\cdot[1-\Phi(2N-N_2^p-d_1^p-1)]\end{aligned} \tag{A.16}$$

The optimal combination of N_1^{p*} and N_2^{p*} are the smallest integers of N_1^p and N_2^p that satisfy the above conditions simultaneously.

Proof done.

APPENDIX B

PROGRAMS FOR THE STUDY OF FLEXIBILITY

```
% This program is used to generate the data for Flex_Model project sloved
% in OPL. All parameters used in Flex_Model can be changed and generated here.

% First, changes the desired parameters and run this program, it will update the corresponding data file
used in the Flex_Model
% Second, run Flex_Model to solve the LP problem with updated data.

clear;
clc;

% Number of replications for frist stage evaluation
DataNum     =   50;

for replication = 1: DataNum,

% Change the data file path and name if you have changed the Flex_Model position
file_name       =       sprintf   ('C:\\Users\\Liang\\Desktop\\Flex_Model_NewSample\\Flex_Model_%d.dat',
replication );
fid =   fopen( file_name, 'w' );


% ------------------------ Setting the parameters -------------------------

N             =   24;                     %   Capacity of each physician;
                                          %   Change the scale of revenue accordingly with number of physicians,
otherwise, all solutions will be zeros
RevPresche      =   0.75;                 %   Revenue of meeting one pre-scheduled demand
RevOpenOwn      =   0.9;                  %   Revenue of meeting one owned open-access demand
RevOpenOther    =   0.85;                 %   Revenue of meeting one open-access demand from other's panel
% recommended. 3:e7, 4: e10, 5: e14, 6: e17, 7: e18, 8: e20, 9: e23, 10:e25

M             =   6;                      %   Number of physicians modeled
Scenario        =   500;                  %   Number of scenarios calculated
Utilization     =   1.4;                  %   Utilization of demand 0.2-1.6; default: 1.0
DemandUpper     =   80;                   %   The maximum realization of a demand
Scale           =   0;                    %   1/(sum of probabilities)

PreDemandRate   =   [ 6, 10, 14, 6, 10, 14 ];
OpenDemandRate  =   [ 18, 14, 10, 18, 14, 10 ];

% ---------------------------------------------------------------------------
                                          %   Set different level of utilization
PreDemandRate   =   round( Utilization .* PreDemandRate );
OpenDemandRate  =   round( Utilization .* OpenDemandRate );

% Average demand rate for pre-scheduling and open access appointment.
% ***** The dimension must be equal to M, the number of physicians  *****
% ***** Change the number and size manually ******************************
% ---------------------------------------------------------------------------
```

```
% Realization of pre-scheduled and open access demand for each physician in scenarios
PreDemand          =   zeros( Scenario, M );
OpenDemand         =   zeros( Scenario, M );

% Corresponding probability of each realization;
PreProb            =   zeros( Scenario, M );
OpenProb           =   zeros( Scenario, M );
Probability        =   ones( 1, Scenario );          %Total Probability of each scenario
Temp               =   zeros( 1, M );

% Generate scenarios and corresponding probabilities
for i = 1:Scenario,
   for j = 1:M,

      PreDemand   ( i, j )    =   poissrnd ( PreDemandRate (j) );
      OpenDemand  ( i, j )    =   poissrnd ( OpenDemandRate (j) );

      PreProb     ( i, j )    =   poisspdf ( PreDemand(i,j), PreDemandRate(j) );
      OpenProb    ( i, j )    =   poisspdf ( OpenDemand(i,j), OpenDemandRate(j) );
   end
end

% Calculate the total probability of each scenario
for i = 1: Scenario,
   for j = 1:M,
      Probability (i) = Probability(i) * PreProb(i, j) * OpenProb(i,j);
   end
   Scale   =   Scale + Probability(i);
end



% ----------------- Writing variables to the data file -------------------
fprintf( fid, '//The data is generated by the program
C:\\MATLAB7\\work\\Flex_data_generator_Multiple.m\n ');
fprintf( fid, '\nN\t=\t%d;\n', N );
fprintf( fid, 'M\t=\t%d;\n', M );
fprintf( fid, 'Scenario\t=\t%ld;\n', Scenario );
fprintf( fid, 'Utilization\t=\t%.2f;\n', Utilization  );
fprintf( fid, 'DemandUpper\t=\t%d;\n', DemandUpper );
%fprintf( fid, 'Scale\t=\t%.4f;\n', 1/Scale );
fprintf( fid, '\n' );

fprintf( fid, 'RevPresche\t\t=\t%f;\n', RevPresche );
fprintf( fid, 'RevOpenOwn\t\t=\t%f;\n', RevOpenOwn );
fprintf( fid, 'RevOpenOther\t=\t%f;\n', RevOpenOther );
fprintf( fid, '\n' );

fprintf( fid, 'OutputFile\t=\t"Output_%d.txt";\n\n', replication );

% ---------------------- write the array structure -----------------------
% write the data array of PreDemand
fprintf( fid, 'PreDemand\t=\t[\n' );
for i = 1:Scenario,
   fprintf( fid, '\t\t\t\t[');
```

```
    for j = 1:M,
        if ( j < M )
            fprintf( fid, '%d, ', PreDemand(i,j) );
        else
            fprintf( fid, '%d ', PreDemand(i,j) );
        end
    end

    if ( i < Scenario )
        fprintf( fid, '],\n');
    else
        fprintf( fid, ']\n' );
    end
end
fprintf( fid, '\t\t\t\t];\n\n');

% write the data array of OpenDemand
fprintf( fid, 'OpenDemand\t=\t[\n' );
for i = 1:Scenario,
    fprintf( fid, '\t\t\t\t[');

    for j = 1:M,
        if ( j < M )
            fprintf( fid, '%d, ', OpenDemand(i,j) );
        else
            fprintf( fid, '%d ', OpenDemand(i,j) );
        end
    end

    if ( i < Scenario )
        fprintf( fid, '],\n');
    else
        fprintf( fid, ']\n' );
    end
end
fprintf( fid, '\t\t\t\t];\n\n');

% write the data array of PreProb
% fprintf( fid, 'PreProb\t\t=\t[\n' );
% for i = 1:Scenario,
%     fprintf( fid, '\t\t\t\t[');
%
%     for j = 1:M,
%         if ( j < M )
%             fprintf( fid, '%f, ', PreProb(i,j) );
%         else
%             fprintf( fid, '%f ', PreProb(i,j) );
%         end
%     end
%
%     if ( i < Scenario )
%         fprintf( fid, '],\n');
%     else
%         fprintf( fid, ']\n' );
%     end
% end
```

```
% fprintf( fid, '\t\t\t\t];\n\n'); %}


% write the data array of OpenProb
% fprintf( fid, 'OpenProb\t=\t[\n' );
% for i = 1:Scenario,
%     fprintf( fid, '\t\t\t\t[');
%
%     for j = 1:M,
%         if ( j < M )
%             fprintf( fid, '%f, ', OpenProb(i,j) );
%         else
%             fprintf( fid, '%f ', OpenProb(i,j) );
%         end
%     end
%
%     if ( i < Scenario )
%         fprintf( fid, '],\n');
%     else
%         fprintf( fid, ']\n' );
%     end
% end
% fprintf( fid, '\t\t\t\t];\n\n');


% % write the data array of probabilites of scenarios
% fprintf( fid, 'Probability\t=\t[ \n' );
% for i = 1:Scenario,
%     if ( i < Scenario )
%         fprintf( fid, '\t\t\t\t%g,\n ', Probability(i) );
%     else
%         fprintf( fid, '\t\t\t\t%g\n ', Probability(i) );
%     end
% end
% fprintf( fid, '\t\t\t\t];\n\n' );


% Close the data file
fclose( fid );

end


/*****************************************
```

```
 * OPL 6.3 Model
 * Author: Liang
 * Creation Date: Apr 20, 2010 at 7:55:31 PM
 * This program is used to solve the LP problem for 2Chain flexibility
 *********************************************/

 int N   =   ...;                                    // Physician Capacity
 int M   =   ...;                                    // Number of physicians
 int Scenario   =   ...;                             // Number of scenarios calculated
 float   Utilization=   ...;                         // Demand utilization. 0.2-1.6, default: 1.0
 int DemandUpper     =   ...;                        // Upper bound of demand realization
 //float Scale       =   ...;                        // 1/total probability
string   OutputFile     =   ...;                     //Outputfile name

 float   RevPresche    =   ...;                      // Revenue of meeting one pre-scheduled demand
 float   RevOpenOwn=   ...;                          // Revenue of meeting one owned open access
demand
 float   RevOpenOther = ...;                         // Revenue of meeting one open access demand of
other's

 range  DocNum        =   1..M;
 range  scenario    =   1..Scenario;
 range  demandupper =   0..DemandUpper;              // the second index of Phi

 int PreDemand  [scenario][DocNum]   =   ...;        // Pre-scheduled demand for each physician in
scenarios
 int OpenDemand [scenario][DocNum] =   ...;          // Open access demand for each physician in
scenarios
 //float Probability[scenario]          =   ...;     // Total probability of each scenario


 dvar   float    Np[DocNum]                      in  0..N;       // Decision variables that how many slots
should be reserved for pre-scheduling
 dvar   float    Xp[scenario][DocNum]            in  0..N;       // Decision variables that how many pre-
scheduled appointments should be met for each scenarios
 dvar   float    Xo[scenario][DocNum][DocNum]  in  0..N;         // Decision variables that how many open
access demand should be met ( own demand and diverted)
 dvar   boolean    Phi[DocNum][demandupper];                     // Binary variables that make sure the
unused pre-scheduled capacity could be pushed to open access


    //   Objective: maximize the revenue of satisfying demands

 maximize sum ( s in scenario, i in DocNum ) (  RevPresche * Xp[s][i] ) +
             sum( s in scenario, i,j in DocNum: j==i ) ( RevOpenOwn * Xo[s][i][j] ) +
             sum( s in scenario, i,j in DocNum: j!=i ) (  RevOpenOther * Xo[s][i][j] );

 subject to{


    forall( s in scenario ){
        // Build the 2-chain flexibility configuration
        forall( i in 1..M-1, j in DocNum : j!=i && j!=(i+1) )        Xo[s][i][j] == 0;
        forall( j in DocNum : j != M && j!= 1 )                      Xo[s][M][j] == 0;

        forall( i in DocNum ){
```

```
            // constraints for decision variables Np
            Np[i] <= PreDemand[s][i] + N * Phi[i][ PreDemand[s][i] ];
            Np[i] >= PreDemand[s][i] * Phi[i][ PreDemand[s][i] ];

            // upper bound constraints for Xp
            Xp[s][i] <= Np[i];                          //Cannot larger than reserved slots
            Xp[s][i] <= PreDemand[s][i];                //Cannot larger than actual pre-scheduled demands

            // Xo cannot be larger than the actual open access demand
            sum ( j in DocNum ) Xo[s][i][j] <= OpenDemand[s][i];
        }

        forall( j in DocNum ){
            // Xo cannot be larger than the capacity left for each physisian
            sum ( i in DocNum ) Xo[s][i][j] <= N - PreDemand[s][j] * Phi[j][ PreDemand[s][j] ];
            sum ( i in DocNum ) Xo[s][i][j] <= N- Np[j] + Phi[j][ PreDemand[s][j] ] * N;
        }
    }


}   // end of constraints


execute {

    //Statistic the results      array indexed from 0
    PreDemandStat        =  new Array(M+1);          // Expected demand for pre-scheduling
    OpenDemandStat       =  new Array(M+1);          // Expected demand for open access
    PreDemandMet         =  new Array(M+1);          // Expected demand met for pre-scheduling
    OpenDemandMet        =  new Array (M+1);         // Expected demand met for open access
    OpenDemandDiverted   =  new Array (M+1);         // Expected demand diverted for open access


    for ( var i=1; i<=M+1; i++ ){
        PreDemandStat[i]        =  0;
        OpenDemandStat[i]       =  0;
        PreDemandMet[i]         =  0;
        OpenDemandMet[i]        =  0;
        OpenDemandDiverted[i]=  0;
    }

    // Begin statistic calculation
    for ( var s=1; s<=Scenario; s++ ){
        for ( i=1; i <= M; i++ ){
            PreDemandStat[i]     =  PreDemandStat[i] +  PreDemand[s][i];
            OpenDemandStat[i]  =  OpenDemandStat[i] + OpenDemand[s][i];
            PreDemandMet[i]      =  PreDemandMet[i]     + Xp[s][i];
            for ( var j=1; j <= M; j++ ){
                OpenDemandMet[i]  =   OpenDemandMet[i] + Xo[s][i][j];
                if ( j!=i )
                    OpenDemandDiverted[i] = OpenDemandDiverted[i] + Xo[s][i][j];
            }
        }

    }   // end calculation
```

```
    for( i=1; i<=M; i++){
        PreDemandStat[M+1]     =   PreDemandStat[M+1] + PreDemandStat[i];
        OpenDemandStat[M+1] =   OpenDemandStat[M+1] + OpenDemandStat[i];
        PreDemandMet[M+1]          =   PreDemandMet[M+1] + PreDemandMet[i];
        OpenDemandMet[M+1] =   OpenDemandMet[M+1] + OpenDemandMet[i];
        OpenDemandDiverted[M+1]     =   OpenDemandDiverted[M+1] + OpenDemandDiverted[i];
    }

    var ofile    =    new IloOplOutputFile ( );
    ofile.open( OutputFile );
    ofile.writeln ("2-chain\tPhysicians\t",M, "\tScenario\t",Scenario, "\tUtilization\t", Utilization,
"\tRevPre\t", RevPresche,
    "\tRevOpenOwn\t",RevOpenOwn, "\tRevOpenOther\t", RevOpenOther, "\tObjective:\t",
cplex.getObjValue()/Scenario, "\tNp:\t", Np,
    "\tTotalDemand:\t", (PreDemandStat[M+1]+ OpenDemandStat[M+1])/Scenario, "\tDemandMet:\t",
(PreDemandMet[M+1] + OpenDemandMet[M+1])/Scenario,
    "\tRefusal:\t", (PreDemandStat[M+1]+ OpenDemandStat[M+1]-PreDemandMet[M+1]-
OpenDemandMet[M+1])/Scenario,
    "\tDiverted:\t", OpenDemandDiverted[M+1]/Scenario );
    ofile.close();
 }




/*********************************************
 * OPL 6.3 Model
 * Author: Liang
 * Creation Date: Apr 21, 2010 at 9:34:22 PM
 * This program is used to solve the LP problem for full flexibility
 *********************************************/

 int N   =   ...;                                    // Physician Capacity
 int M  =   ...;                                     // Number of physicians
 int Scenario   =   ...;                             // Number of scenarios calculated
 float   Utilization=   ...;                         // Demand utilization. 0.2-1.6, default: 1.0
 int DemandUpper     =   ...;                        // Upper bound of demand realization
 //float Scale        =   ...;                       // 1/total probability
 string  OutputFile     =   ...;                     //Outputfile name

 float   RevPresche    =   ...;                      // Revenue of meeting one pre-scheduled demand
 float   RevOpenOwn=   ...;                          // Revenue of meeting one owned open access
demand
 float   RevOpenOther = ...;                         // Revenue of meeting one open access demand of
other's

 range  DocNum        =   1..M;
 range  scenario    =   1..Scenario;
 range  demandupper =   0..DemandUpper;              // the second index of Phi

 int PreDemand  [scenario][DocNum]   =   ...;        // Pre-scheduled demand for each physician in
scenarios
 int OpenDemand [scenario][DocNum] =   ...;          // Open access demand for each physician in
scenarios
// float Probability[scenario]          =   ...;      // Total probability of each scenario
```

```
dvar   float    Np[DocNum]                      in  0..N;          // Decision variables that how many slots
should be reserved for pre-scheduling
dvar   float    Xp[scenario][DocNum]            in  0..N;          // Decision variables that how many pre-
scheduled appointments should be met for each scenarios
dvar   float    Xo[scenario][DocNum][DocNum]  in  0..N;        // Decision variables that how many open
access demand should be met ( own demand and diverted)
dvar   boolean    Phi[DocNum][demandupper];                     // Binary variables that make sure the
unused pre-scheduled capacity could be pushed to open access


    //   Objective: maximize the revenue of satisfying demands

maximize sum ( s in scenario, i in DocNum ) (  RevPresche * Xp[s][i] ) +
            sum( s in scenario, i,j in DocNum: j==i ) (  RevOpenOwn * Xo[s][i][j] ) +
            sum( s in scenario, i,j in DocNum: j!=i ) (  RevOpenOther * Xo[s][i][j] );

subject to{

    forall( s in scenario ){

        forall( i in DocNum ){
            // constraints for decision variables Np
            Np[i] <= PreDemand[s][i] + N * Phi[i][ PreDemand[s][i] ];
            Np[i] >= PreDemand[s][i] * Phi[i][ PreDemand[s][i] ];

            // upper bound constraints for Xp
            Xp[s][i] <= Np[i];                          //Cannot larger than reserved slots
            Xp[s][i] <= PreDemand[s][i];                //Cannot larger than actual pre-scheduled demands

            // Xo cannot be larger than the actual open access demand
            sum ( j in DocNum ) Xo[s][i][j] <= OpenDemand[s][i];
        }

        forall( j in DocNum ){
            // Xo cannot be larger than the capacity left for each physisian
            sum ( i in DocNum ) Xo[s][i][j] <= N - PreDemand[s][j] * Phi[j][ PreDemand[s][j] ];
            sum ( i in DocNum ) Xo[s][i][j] <= N- Np[j] + Phi[j][ PreDemand[s][j] ] * N;
        }
    }

}   // end of constraints

 execute {

    //Statistic the results      array indexed from 0
    PreDemandStat        =   new Array(M+1);                // Expected demand for pre-scheduling
    OpenDemandStat       =   new Array(M+1);                // Expected demand for open access
    PreDemandMet         =   new Array(M+1);                // Expected demand met for pre-scheduling
    OpenDemandMet        =   new Array (M+1);               // Expected demand met for open access
    OpenDemandDiverted   =   new Array (M+1);               // Expected demand diverted for open access


    for ( var i=1; i<=M+1; i++ ){
        PreDemandStat[i]        =   0;
        OpenDemandStat[i]       =   0;
```

```
        PreDemandMet[i]          =   0;
        OpenDemandMet[i]         =   0;
        OpenDemandDiverted[i]=   0;
    }

    // Begin statistic calculation
    for ( var s=1; s<=Scenario; s++ ){
        for ( i=1; i <= M; i++ ){
            PreDemandStat[i]     =   PreDemandStat[i] +  PreDemand[s][i];
            OpenDemandStat[i]  =   OpenDemandStat[i] + OpenDemand[s][i];
            PreDemandMet[i]      =   PreDemandMet[i]      + Xp[s][i];
            for ( var j=1; j <= M; j++ ){
                OpenDemandMet[i]  =   OpenDemandMet[i] + Xo[s][i][j];
                if ( j!=i )
                    OpenDemandDiverted[i] = OpenDemandDiverted[i] + Xo[s][i][j];
            }
        }
    }   // end calculation

    for( i=1; i<=M; i++){
        PreDemandStat[M+1]     =   PreDemandStat[M+1] + PreDemandStat[i];
        OpenDemandStat[M+1] =   OpenDemandStat[M+1] + OpenDemandStat[i];
        PreDemandMet[M+1]          =   PreDemandMet[M+1] + PreDemandMet[i];
        OpenDemandMet[M+1] =   OpenDemandMet[M+1] + OpenDemandMet[i];
        OpenDemandDiverted[M+1]     =   OpenDemandDiverted[M+1] + OpenDemandDiverted[i];
    }


    var ofile    =    new IloOplOutputFile ( );
    ofile.open( OutputFile );
    ofile.writeln ("Full Flex\tPhysicians\t",M, "\tScenario\t",Scenario, "\tUtilization\t", Utilization,
"\tRevPre\t", RevPresche,
    "\tRevOpenOwn\t",RevOpenOwn, "\tRevOpenOther\t", RevOpenOther, "\tObjective:\t",
cplex.getObjValue()/Scenario, "\tNp:\t", Np,
    "\tTotalDemand:\t", (PreDemandStat[M+1]+ OpenDemandStat[M+1])/Scenario, "\tDemandMet:\t",
(PreDemandMet[M+1] + OpenDemandMet[M+1])/Scenario,
    "\tRefusal:\t", (PreDemandStat[M+1]+ OpenDemandStat[M+1]-PreDemandMet[M+1]-
OpenDemandMet[M+1])/Scenario,
    "\tDiverted:\t", OpenDemandDiverted[M+1]/Scenario );
    ofile.close();
     }



/*********************************************
 * OPL 6.3 Model
 * Author: Liang
 * Creation Date: Apr 22, 2010 at 2:53:19 PM
 * This program is used to solve the LP problem for no flexibility
 *********************************************/

 int N   =   ...;                                    // Physician Capacity
 int M   =   ...;                                    // Number of physicians
 int Scenario   =   ...;                             // Number of scenarios calculated
 float   Utilization=   ...;                         // Demand utilization. 0.2-1.6, default: 1.0
 int DemandUpper     =   ...;                        // Upper bound of demand realization
```

```
//float Scale        =   ...;                          // 1/total probability
string OutputFile     =   ...;                          //Outputfile name

float   RevPresche    =   ...;                          // Revenue of meeting one pre-scheduled demand
float   RevOpenOwn=   ...;                              // Revenue of meeting one owned open access
demand
float   RevOpenOther = ...;                             // Revenue of meeting one open access demand of
other's

range DocNum        =   1..M;
range scenario    =   1..Scenario;
range demandupper =   0..DemandUpper;                   // the second index of Phi

int PreDemand  [scenario][DocNum]   =   ...;            // Pre-scheduled demand for each physician in
scenarios
int OpenDemand [scenario][DocNum] =   ...;              // Open access demand for each physician in
scenarios
//float Probability[scenario]          =   ...;         // Total probability of each scenario


dvar   float    Np[DocNum]                      in  0..N;   // Decision variables that how many slots should
be reserved for pre-scheduling
dvar   float    Xp[scenario][DocNum]            in  0..N;   // Decision variables that how many pre-
scheduled appointments should be met for each scenarios
dvar   float    Xo[scenario][DocNum]            in  0..N;   // Decision variables that how many open access
demand should be met
dvar   boolean    Phi[DocNum][demandupper];             // Binary variables that make sure the unused pre-
scheduled capacity could be pushed to open access


//   Objective: maximize the revenue of satisfying demands

maximize sum ( s in scenario, i in DocNum ) (  RevPresche * Xp[s][i] ) +
            sum ( s in scenario, i in DocNum ) (  RevOpenOwn * Xo[s][i] ) ;

subject to{

    forall( s in scenario ){
        forall( i in DocNum ){
            // constraints for decision variables Np
            Np[i] <= PreDemand[s][i] + N * Phi[i][ PreDemand[s][i] ];
            Np[i] >= PreDemand[s][i] * Phi[i][ PreDemand[s][i] ];

            // upper bound constraints for Xp
            Xp[s][i] <= Np[i];                      //Cannot larger than reserved slots
            Xp[s][i] <= PreDemand[s][i];            //Cannot larger than actual pre-scheduled demands

            // Xo cannot be larger than the actual open access demand
            Xo[s][i] <= OpenDemand[s][i];

            // Xo cannot be larger than the capacity left for each physisian
            Xo[s][i] <= N - PreDemand[s][i] * Phi[i][ PreDemand[s][i] ];
            Xo[s][i] <= N- Np[i] + Phi[i][ PreDemand[s][i] ] * N;
        }
    }
```

```
}   // end of constraints


execute {

    //Statistic the results      array indexed from 0
    PreDemandStat       =   new Array(M+1);             // Expected demand for pre-scheduling
    OpenDemandStat      =   new Array(M+1);             // Expected demand for open access
    PreDemandMet        =   new Array(M+1);             // Expected demand met for pre-scheduling
    OpenDemandMet       =   new Array (M+1);            // Expected demand met for open access


    for ( var i=0; i<M+1; i++ ){
        PreDemandStat[i]        =   0;
        OpenDemandStat[i]       =   0;
        PreDemandMet[i]         =   0;
        OpenDemandMet[i]        =   0;
    }

    // Begin statistic calculation
    for ( var s=1; s<=Scenario; s++ ){
        for ( i=1; i <= M; i++ ){
            PreDemandStat[i-1]  =   PreDemandStat[i-1] +  PreDemand[s][i];
            OpenDemandStat[i-1]     =   OpenDemandStat[i-1] + OpenDemand[s][i];
            PreDemandMet[i-1]  =   PreDemandMet[i-1]  + Xp[s][i];
            OpenDemandMet[i-1]     =   OpenDemandMet[i-1] +  Xo[s][i];
        }
    }   // end calculation

    for( i=0; i<M; i++){
        PreDemandStat[M]   =   PreDemandStat[M] + PreDemandStat[i];
        OpenDemandStat[M]      =   OpenDemandStat[M] + OpenDemandStat[i];
        PreDemandMet[M]        =   PreDemandMet[M] + PreDemandMet[i];
        OpenDemandMet[M]       =   OpenDemandMet[M] + OpenDemandMet[i];
    }


    var ofile    =    new IloOplOutputFile ( );
    ofile.open( OutputFile );

    ofile.writeln ("No Flex\tPhysicians\t",M, "\tScenario\t",Scenario, "\tUtilization\t", Utilization,
"\tRevPre\t", RevPresche,
    "\tRevOpenOwn\t",RevOpenOwn, "\tRevOpenOther\t", RevOpenOther, "\tObjective:\t",
cplex.getObjValue()/Scenario, "\tNp:\t", Np,
    "\tTotalDemand:\t", (PreDemandStat[M]+ OpenDemandStat[M])/Scenario, "\tDemandMet:\t",
(PreDemandMet[M] + OpenDemandMet[M])/Scenario,
    "\tRefusal:\t", (PreDemandStat[M]+ OpenDemandStat[M]-PreDemandMet[M]-
OpenDemandMet[M])/Scenario );

    ofile.close();

}
```

BIBLIOGRAPHY

[1] California HealthCare Foundation. Health care costs 101, April 2009. Accessed May 24th, 2009 at http://www.chcf.org/documents/insurance/HealthCareCosts09.pdf.

[2] World Health Organization. World Health Statistics 2008. World Health Organization, 2008.

[3] Keehan, S., Sisko, A., Truer, C., et al. Health spending projections through 2017: the babyboom generation is coming to medicare. Health Affairs 2, 27 (2008), 145-155.

[4] World Health Organization. The World Health Report 2000 - Health Systems: Improving Performance. World Health Organization, January 2000.

[5] World Health Organization. The World Health Report 2008: Primary Health Care Now More Than Ever. World Health Organization, December 2008.

[6] Grumbach, K., Selby, J. V., Damberg, C., et al. Resolving the gatekeeper conundrum: what patients value in primary care and referrals to specialists. Journal of the American Medical Association 282, 3 (Jul 1999), 261-266.

[7] American College of Physicians. The impending collapse of primary care and its implications for the state of the nation's healthcare. Tech. rep., American College of Physicians, 2006. Accessed May 25th, 2009 at http://www.acponline.org/advocacy/events/state of healthcare/statehc06 1.pdf.

[8] Arvantes, J. Health care experts describe the benefits of primary care, November 2007. Accessed May 25th, 2009 at http://www.aafp.org/online/en/home/publications/news/news-now/professional-issues/20070611pcforum.html.

[9] Starfield, B., Shi, L., Macinko, J. Contribution of primary care to health systems and health. The Milbank quarterly 83, 3 (2005), 457-502.

[10] Bodenheimer, T. Primary Care - Will It Survive? The New England Journal of Medicine 355, 9 (2006), 861-864.

[11] Committee on Quality of Health Care in America, Institute of Medicine, Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academy Press, 2001.

[12] Strunk, B. C., Cunningham P. J. Treading water: Americans' access to needed medical care, 1997-2001. Tech. rep., Center for Studying Health System Change 2002.

[13] Rust, G., Ye, J., Baltrus, P., et al. Practical barriers to timely primary care access. Archives of Internal Medicine 268, 15 (2008), 1705-1710.

[14] Gill, J. M., Mainous A. G. The role of provider continuity in preventing hospitalizations. Archives of Family Medicine 7 (1998), 352-357.

[15] Hing, E., Burt C. W. Characteristics of office-based physicians and their medical practices: United states, 20052006. Tech. rep., U.S. Department of Health and Human Services, 208.

[16] Murray, M., Bodenheimer T. Rittenhouse D. Grumbach. K. Improving timely access to primary care: Case studies of the advanced access model. Journal of the American Medical Association 289, 3 (2003), 1042-1046.

[17] Green, L. V., Savin, S., Murray, M. Providing timely access to care: What is the right patient panel size? The Joint Commission Journal on Quality and Patient Safety 33 (2007), 211-218.

[18] Committee on Quality of Health Care in America, Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington DC: National Academy Press, 2001.

[19] Xiuli Qu and Jing Shi. Effect of two-level provider capacities on the performance of open access, Health Care Manage Science, 2009 12:99-114.

[20] Murray, M., Tantau C. Redefining open access to primary care. Managed Care Quarterly 7, 3 (1999), 45-55.

[21] Murray, M., Tantau C. Same-day appointments: Exploding the access paradigm. Family Practice Management 7, 8 (2000), 45-50.

[22] Qu X, Rardin RL, Williams JAS, Willis DR. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. Eur J Oper Res, 2007 187:812-826.

[23] Kopach, R., DeLaurentis, P., Lawley, M. et al. Effects of clinical characteristics on successful open access scheduling. Health Care Management Science 10, 2(2007), 111-124.

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

[25] Jordan, W. C., Graves S. C. Principles and benefits of manufacturing process flexibility. Management Science 41, 4 (1995), 577-594.

[26] Graves, S. C., Tomlin B. T. Process flexibility in supply chains. Management Science 49, 7 (2003), 907-919.

[27] Muriel, A., Somasundaram, A., Zhang, Y. Impact of partial manufacturing flexibility on production variability. Manufacturing & Service Operations Management 8, 2 (2006), 192-205.

[28] Brusco, Michael J., Johns Tony R. Staffing a multi-skilled workforce with varying levels of productivity: An analysis of cross-training policies. Decision Sciences 29, 2 (1998), 499-515.

[29] Chou, Mabel C., Chua, Georey A., Teo, Chung-Piaw. On range and response: Dimensions of process flexibility. Working paper, NSU 2008.

[30] Sheikhzadeh, M., Benjaafar, S., Gupta, D. Machine sharing in manufacturing systems: Total flexibility versus chaining. The International Journal of Flexible Manufacturing Systems 10, 4 (1998), 351-378.

[31] Gurumurthi, S., Benjaafar S. Modeling and analysis of flexible queuing systems. Naval Research Logistics 51, 755 - 782 (2004).

[32] Hopp, W., Tekin, E., Van Oyten, M. P. Benefits of skill chaining in serial production lines with cross-trained workers. Management Science 50, 4 (2004), 83-98.

[33] Bennett, K. J., Baxley E. G. The effect of a carve-out advanced access scheduling system on no-show rates. Practice Management 41, 1 (2009), 51-56.

[34] Solak, S., Clarke, J.B., Johnson, E.L., Barnes, E.R., Optimization of R&D Project Portfolios under Endogenous Uncertainty, European Journal of Operational Research (2010).

[35] Jan Hippchen, Physician Flexibility in Primary Care Practices, Master thesis, University of Massachusetts, Amherst, 2009.