STOCHASTIC SCHEDULING OPTIMIZATION FOR SOLVING PATIENT NO-SHOW AND APPOINTMENT CANCELLATION PROBLEMS IN OUTPATIENT CLINICS

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yidong Peng

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Industrial and Manufacturing Engineering

May 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

STOCHASTIC SCHEDULING OPTIMIZATION FOR SOLVING PATIENT NO-SHOW AND APPOINTMENT CANCELLATION PROBLEMS IN OUTPATIENT CLINICS

By

Yidong Peng

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Jing Shi
Chair

Dr. Kambiz Farahmand

Dr. Xiuli Qu

Dr. Joseph Szmerekovsky

Approved:

07/29/2015
Date

Dr. Om Prakash Yadav
Department Chair

ABSTRACT

Patient non-attendance is the major challenge that reduces practice efficiency, resource utilization, and clinic accessibility, and leads to increased cost and diminished quality of care, while the clinic scheduling system is known as a determining factor for clinic efficiency, resource utilization and the accessibility of patients to healthcare facilities. A suitable and optimized scheduling system is one of the most important components for efficient care delivery to address the major challenges in the healthcare industry. Hence, reducing the adverse effect of patient no-shows and short-notice appointment notifications through the appointment scheduling approach is a strategically important matter for any healthcare systems.

In this research, three patient scheduling models are proposed to address the patient non-attendance problem in the outpatient clinics. The first model is a two-stage mixed stochastic programming model, which can be used to optimize the overbooking decisions: (1) How many appointment slots should be overbooking; (2) Which appointment slot should be overbooking. In addition, this model also considers the cooperation between providers and patients' choice. The second model is a Markov Decision Process (MDP) model, which can be used to optimize the walk-in patient admission policy in clinics with single physician by answering the four vital questions: (1) When the walk-in patient admission decisions should be made; (2) At each decision point, how many walk-in patients should be admitted; (3) Which provider should serve the admitted walk-in patients; (4) When the admitted walk-in patient should be served. By using this MDP model, heuristic optimal walk-in patient admission rules have been found for the single physician systems. For systems with more physicians, a more advanced two-stage mixed stochastic programming model (the third model) is proposed in order to make the optimal real time walk-in patient admission decisions. At last, it worthwhile to mention that novel solution approach has also been developed in order to solve these models in the efficient and effective manner.

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisor and committee chair, Dr. Jing Shi, who has the attitude and the substance of a genius; he continually and convincingly conveyed a spirit of adventure in regard to research and scholarship, and an excitement in regard of teach. Without his guidance and persistent help this thesis would not have been possible.

I would like to thank my committee members, Dr. Kambiz Farahmand, Dr. Xiuli Qu, and Dr. Joseph Szmerekovsky, for offering time and individual expertise to support and guide me though this exciting and painstaking journey.

I would also like to thank my colleague Dr. Ergin Erdem, who is a loyal friend and provides me a lot of help and knowledge during the study.

Last and most importantly, I want to thank my family for their selfless love and consistent support.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGMENTS v

LIST OF TABLES ix

LIST OF FIGURES xi

1. INTRODUCTION 1

1.1. Background 1

1.2. Problem statement 4

2. LITERATURE REVIEW 10

2.1. Patient no-shows and short-notice appointment cancellations 10

2.2. Clinic appointment scheduling optimization 14

2.3. Open access scheduling 15

2.4. Overbooking 18

2.5. Walk-in patient admission 20

3. TWO-STAGE STOCHASTIC PATIENT OVERBOOKING OPTIMIZATION MODEL . 23

3.1. Introduction 23

3.2. Terminology and problem formulation 24

3.2.1. Terminology 24

3.2.2. Problem formulation 26

3.3. Solution approach 32

3.4. Model verification 34

3.5. Numerical analysis 40

3.5.1. Data collection and study design 40

3.5.2. Numerical analysis result 45

3.5.3. Adverse effect of patient no-show and appointment cancellation 50

3.5.4. Effect of non-homogenous rate of patient no-show rate and appointment cancellation rate 51

3.6. Model extension 53

3.6.1. Problem description 53

3.6.2. Terminology 53

3.6.3. Formulation 55

3.6.4. Numerical analysis 59

4. SINGLE-PROVIDER WALK-IN PATIENT ADMISSION OPTIMIZATION MODEL ... 64

4.1. Problem definition 64

4.2. MDP model 67

4.2.1. States and actions 67

4.2.2. State transitions 71

4.2.3. Rewards and costs 72

4.3. Properties of the MDP model 74

4.4. Optimal and heuristic rules for walk-in patient admission 76

4.5. Comparison of heuristic walk-in patient admission rules 80

4.5.1. Data collection and numerical scenarios 81

4.5.2. Overbooking policies and walk-in patient admission policies 83

4.5.3. Reward and cost coefficients 85

4.5.4. Comparison of walk-in patient admission policies 87

4.5.5. Discussion 91

5. MULTI-PROVIDER WALK-IN PATIENT ADMISSION OPTIMIZATION MODEL 93

5.1. Introduction 93

5.2. Terminology and problem formulation 95

5.2.1. Terminology 95

5.2.2. Problem formulation 97

5.3. Solution approach 103

5.4. Numerical analysis 107

5.4.1. Data collection and experiment design 107

5.4.2. Numerical analysis results 115

5.4.3. Heuristic rule 125

6. CONCLUSIONS AND FUTURE RESEARCH 127

6.1. Conclusions 127

6.2. Future research 130

REFERENCES 133

APPENDIX. PROOFS OF PROPOSITIONS 147

LIST OF TABLES

Table Page

3.1: Comparison result for scenarios 1 – 5 35

3.2: Comparison result for scenarios 6 – 10 36

3.3: Comparison result for scenarios 11 – 15 38

3.4: Comparison result for scenarios 16 – 20 39

3.5: Parameters used for the base case.. 44

3.6: Parameter adjustments for Cases 1-8 compared with Base case 45

3.7: Optimal overbooking strategy for Cases 0 – 8 45

3.8: Summary of descriptive performance statistics for Cases 0 – 8 47

3.9: Optimal double-booking strategies.. 52

3.10: Optimal overbooking strategy for Case 0 – 8.. 59

3.11: Summary of descriptive performance statistics for Cases 0 - 8................................... 63

4.1: Indices and parameters in the MDP model... 69

4.2: States, actions, random variables and rewards in the MDP model............................... 70

4.3: Optimal and heuristic admission rules for the states $s^m \in$S satisfying $w^m = 0$ 77

4.4: Optimal and heuristic admission rules for the states $s^m \in$S satisfying $w^m > 0$ 80

4.5: Summary of the optimal and heuristic rules for walk-in patient admission 81

4.6: Patient arrival pattern in the clinic motivating this study 83

4.7: Patient arrival patterns for the comparison of the heuristic rules 84

4.8: Overbooking policies and walk-in patient admission policies 85

4.9: Cost coefficients in the immediate net reward.. 87

4.10: Average number of walk-in patients seen per clinic session under admission policy A6... 91

5.1: Patient non-attendance rates for numerical analysis 113

5.2: Patient arrival patterns for numerical analysis 114

5.3: Patient service time distributions for numerical analysis 114

5.4: Cost coefficients for numerical analysis 114

5.5: Parameters used for the base case 115

5.6: Parameter adjustments for Cases 1-8 compared with base case 115

5.7: Summary of descriptive performance statistics for Cases 0- 8 (current rate) 120

5.8: Summary of descriptive performance statistics for Cases 0- 8 (future rate) 120

LIST OF FIGURES

Figure Page

1.1: Adverse effect of patient non-attendance 2

3.1: Average patient waiting time – double-booked slots vs. single booked slots 48

3.2: Average waiting time over appointment slots 49

3.3: Effect of no-show rate on the objective function value 51

3.4: Effect of no-show rate on the double-booking rate 51

4.1: Patient flow and walk-in patient admission decisions in a clinic 65

4.2: Patient arrival process and walk-in patient admission process in a clinic session 65

4.3: Performance of the admission policies A2 – A8 under overbooking policy O1 88

4.4: Performance of the admission policies A2 – A8 under overbooking policy O2 88

4.5: Performance of the admission policies A2 – A8 under overbooking policy O3 89

4.6: Performance of the admission policies A1, A2, and A6 under overbooking policies O1 – O3 90

5.1: Comparison of optimal objective values between walk-in admission strategy and double booking strategy 117

5.2: Comparison of patient waiting times between walk-in patients and patients with appointments 117

5.3: Comparison of provider idle times under different walk-in patient arrival rates 118

5.4: Comparison of provider overtimes under different walk-in patient arrival rates 118

5.5: Impact of patient attendance rate on the number of walk-in admissions 121

5.6: Impact of patient arrival lead time on the number of walk-in patient admissions 122

5.7: Impact of patient service time on the number of walk-in admissions 122

5.8: Impact of cost coefficient ratio on the number of walk-in admissions 123

5.9: Effect of no-show rate on the number of walk-in admissions 124

5.10: Effect of no-show rate on the objective function value 125

1. INTRODUCTION

1.1. Background

In the United States, the national healthcare expenditure amounted to $2.7 trillion in 2011, which accounted for 17.9% of its GDP. This figure is more than all healthcare expenditures in other countries combined (WHO, 2011). However, low clinic efficiency and poor patient accessibility to care still remain the major challenges for most outpatient clinics in the United States. It is estimated by the Institute of Medicine that the U.S. healthcare systems wasted $190 billion annually on inefficient delivery of care (Campbell, 2012). Among the inefficiencies, patient non-attendance which includes the patient no-shows and short-notice appointment cancellations (usually defined as cancellations prior to less than 48 hours of the clinic appointment date/time) is the major contributor in primary care clinics. It is also believed that patient non-attendance is the major challenge that reduces practice efficiency, resource utilization, and clinic accessibility, and leads to increased cost and diminished quality of care, as summarized in Fig.1.1. On the other hand, the clinic scheduling system is known as a determining factor for clinic efficiency, resource utilization and the accessibility of patients to healthcare facilities. A suitable and optimized scheduling system is one of the most important components for efficient care delivery to address the major challenges in the healthcare industry. Hence, reducing the adverse effect of patient no-shows and short-notice appointment notifications through the appointment scheduling approach is a strategically important matter for any healthcare systems.

Under the setting of traditional appointment scheduling, patients make appointments weeks or months earlier by calling the clinic or right after their current visits. Usually, the appointments are not available in near term, since most clinics operate at their capacity. As a result, the patients need to wait several weeks or months for their clinic visits. In case of an urgent appointment, the patients may have to use the emergency department. It leads to a disruption of care continuity because they are not able to see their own providers in time. It also dramatically increases the care cost in an unnecessary way since the cost of emergency department visits is much higher than that of primary clinic visits. Meanwhile, patient no-show rate and short-notice appointment cancellation rate are likely to increase due to the long waiting list for appointments. It is well-known that patient no-shows and short-notice appointment cancellations increase the volatility to the standard clinic process, which would eventually increase the healthcare expenditure and decrease clinic efficiency and patient accessibility.

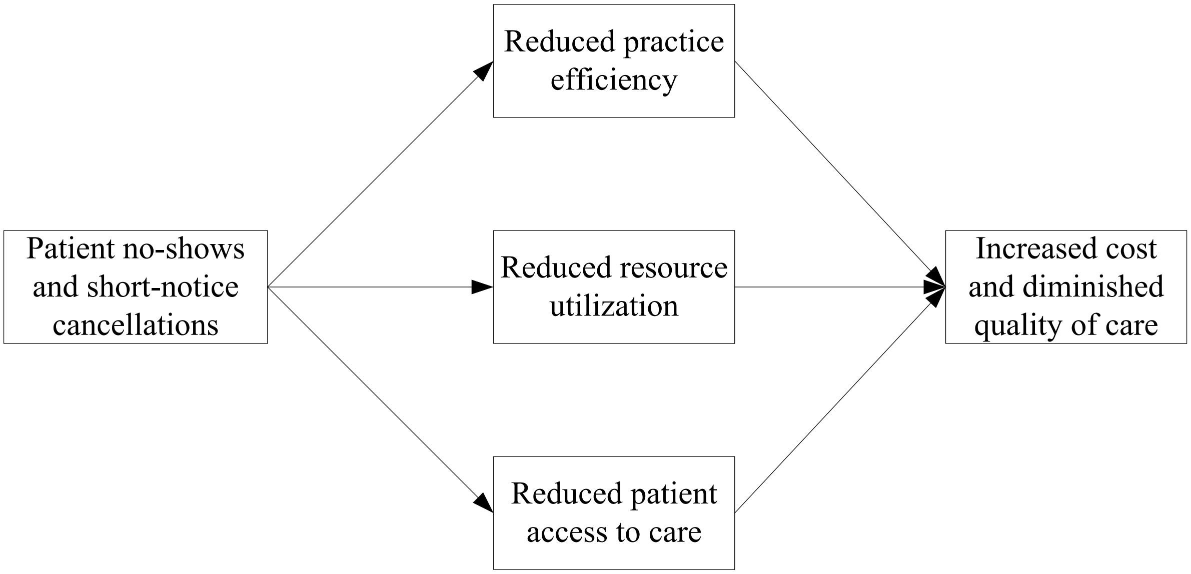
Reduced practice efficiency
Patient no-shows and short-notice cancellations
Reduced resource utilization
Increased cost and diminished quality of care
Reduced patient access to care

Fig. 1.1: Adverse effect of patient non-attendance

Due the disadvantages of the traditional appointment scheduling system, open access scheduling (or advanced access scheduling) was proposed in 1990s, which promotes timely access to care, and improve patient satisfaction. The key concept of open access scheduling is to “do today’s work today”. Under this concept, a portion of clinic slots are reserved for the patients who need same-day appointments, while the non-reserved slots are scheduled in advance for patient with non-acute illness. In order to optimize the appointment schedule template under the open access concept, clinics need to find out the optimal number and sequence of the reserved appointments according to their clinic setting and patient attendance characteristic.

Besides the open access scheduling approach, overbooking is also a well-known practice for mitigating the adverse effects of patient no-show and short-notice appointment cancellation. Generally, by using the overbooking strategy, clinics can book two or more (normally two) patients in the same appointment slots. In case of one patient being no-show or cancelling the appointment with short-notice, the provider can still see other patients in the same appointment slot. However, in case of more than one patients showing up for the same appointment, it can be expected that the patient waiting time will increase dramatically. Hence, the common problem with overbooking practice is to determine the number of overbooking slots and which appointment slots should be overbooked.

Besides the above mentioned practice, it is also believed that clinics can also reduce the adverse effects of patient no-show and short-notice appointment cancellation by admitting walk-in patients (Moore et al., 2001; Liu et al., 2010). The key concept is using walk-in patients to fill the empty appointment slots due to patient no-shows and short-notice appointment cancellations. In order to achieve the maximum profit and patient satisfaction, the clinics need to optimize their walk-in patient admission policy, since too many walk-in patient admissions will lead to reduced patient satisfaction, while too few admissions can result in the loss of profit. Hence, it is worthwhile to develop optimization models/methods that can find the optimal walk-in patient admission policy where the optimal number of walk-in patients can be admitted at the right time.

1.2. Problem statement

It can be shown that open access scheduling optimization problem and overbooking scheduling optimization problem belong to the category of static appointment scheduling problem, in which all decisions are made before the clinic session starts. The static appointment scheduling determines the scheduling template of providers, i.e., how many patients to be booked in each appointment slot. On the other hand, walk-in admission optimization problem is in the category of dynamic appointment scheduling problem, where the walk-in admission decisions are made during the clinic session. To be more specific, the walk-in admission optimization problem uses the schedule template as the input, and determines which walk-in patient should be admitted. Note that for both type of problems, we want to mitigate the adverse effect raised by patient no-show and short-notice appointment cancellation. With this understanding, the work of this dissertation can be divided into two different tasks, as shown in Fig 1.1. In the first task, we solve the static appointment scheduling problem, which finds the best overbooking strategy to deal with patient no-show and short-notice appointment cancellation. In the second task, we solve the dynamic appointment scheduling problem, which optimizes the walk-in patient admission decisions.

In the existing literature, the patient overbooking problem has drawn a lot of attention. The basic questions in the patient overbooking problems are:

1) How many appointment slots should be overbooked?

2) Which appointment slot should be overbooked?

The number of appointment slots to be overbooked is of vital importance for the patient overbooking problem, since inappropriate decisions may lead to long patient waiting time (if too many overbooking) or long provider idle time (if too few overbooking). Intuitively, the number of overbooked slots can be estimated by the total number of appointment slots times the non-attendance rate. However, many studies have shown that this kind of estimation may not provide the optimal solution. Hence, advanced techniques are needed to find the optimal solution to this problem.

Besides, the number of appointment slots to be overbooked (i.e., the second question) is also of vital importance for patient overbooking problem. Note that the optimal number of overbooked appointment slots alone does not guarantee the optimal solution to the overbooking problem. For example, assume we know that only 1 appointment slot should be overbooked, we need further decide which appointment slot should be overbooked. To answer this question, we need to consider the patient non-attendance rate, patient service time distribution, patient arrival patterns, and etc. Hence, this is a complex problem that requires advanced modeling tools and solution approach.

As for the first task of this study, we will formulate this complex patient overbooking problem, which considers patients' choice, into a two-stage stochastic programming model, where the objective is the weighted sum of patient waiting time, provider idle time and provider overtime. Note that, to some extent, the patient waiting time, provider idle time and provider overtime can indicate the operation efficiency, resource utilization and access to care, respectively. By comparing the existing overbooking studies, our study contributes to the literature in the following aspects: (1) Our study is the first one that applied the two-stage stochastic programming model to investigate the patient overbooking problem; (2) Our study does not have restriction on patient service time distribution (some studies assume constant service time) and patient arrival patterns (some studies assume punctual arrival). (3) Our study considers a clinic system with multiple providers.

As for the second task of this dissertation, it will be shown in the literature review that limited attention has been paid to the walk-in patient admission optimization problem. In such clinics, admitting some walk-in patients is usually adopted to reduce the negative impact of patient no-shows and short-notice cancellations and improve the operations efficiency, resource utilization and patients' accessibility to the clinics. To solve walk-in patient admission optimization problem, we need to give answers to the following four questions one by one:

1) When should the walk-in patient admission decisions be made?

2) At each decision point, how many walk-in patients should be admitted?

3) Which provider should serve the admitted walk-in patients?

4) When should the admitted walk-in patients be served?

An intuitive answer to the first question is making the admission decision upon the arrival of the walk-in patients. However, this will pose great difficulty in mathematically modeling the problem, since the random arrival pattern of walk-in patient will bring uncertainty to the decision time point. Hence, clinics could choose to make the decision at fixed time points (e.g., the beginning of each appointment slot), in order to simply the walk-in patient admission optimization problem. It should be noted that both scenarios are examined in this study.

The answer to the second question is of critical importance to the walk-in patient admission optimization problem, since it directly affects the objective function value. As we can imagine, excess walk-in admissions could increase the patient waiting time and provider overtime, while insufficient walk-in admissions could increase the provider idle time, given the high patient no-show rate and short-notice appointment cancellation rate. Much information has to be considered in order to make this decision, such as the number of previous admitted walk-in patients and number of providers in the clinics. In addition, the randomness of patient no-show and short-notice appointment cancellation, as well as the random service introduce more difficulties in answering this question.

The answers to the third and fourth question are also important to the walk-in patient admission optimization problem, since the arrangement of the admitted walk-in patients can also affect the objective function value. For example, assume we have the optimal answer for the second question, which provides the optimal number of admitted walk-in patients, a bad arrangement of those patient can easily increase patient waiting time, provider idle time as well as the provider overtime. Hence, to find best answers for question 3 and 4, we might need to apply the techniques, which have been tested for solving job shop scheduling problem.

As we can see, none of four questions can be easily addressed. When they are combined together, there is no doubt that both advanced modeling skills and leading solution approaches should be adopted to find the right answer. The existing modeling approach for clinic scheduling optimization might not be applicable or hard to be applied to the walk-in patient admission optimization problem. In this sense, it is in urgent need of developing models and tools that could be used to make optimal walk-in patient admission decisions. As a result, this study will also develop models and solution approaches for solving the proposed walk-in admission optimization problem, which will fill the gap between the existing studies and the needs from clinics.

At last, it should be indicated that the contributions of this study will be as follows:

1) Our study will be the first quantitative research investigating the dynamic walk-in patient admission optimization problem in regards of mitigating the negative impacts of patient no-shows and short-notice appointment cancellations. Hence, this study can fill the gap between the existing study and needs from clinics.

2) Two different types of modeling approaches, i.e., Markov Decision Process (MDP) and Stochastic Mixed Integer Programming (SMIP) will be piloted on modeling the walk-in patient admission optimization problem. Although both modeling approaches have been studies and applied by many researches, there is no general framework which can be used to model the newly proposed problem. Hence, our models will be the first MDP model and SMIP model for the dynamic walk-in patient admission optimization problem.

3) Advanced solution approaches will be developed to solve the newly proposed problem. For the MDP model, the large size of state space can hinder the efficiency of solution procedure. To make it worse, it might be impossible to locate the optimal solution for some large size problems. In addition, there is no standard solution procedure for MDP models, i.e., the MDP models in the existing literature all have their own dedicate solution procedures. Hence, as one major contribution of this study, we propose a simulation based genetic algorithm to find the heuristic optimal walk-in patient admission policy.

For the SMIP model, since the walk-in patient admission decision need to be made in real time or within short time (less than 1 minute), it can be expected that the solution approaches should be quite efficient. As it is well known, the stochastic characteristic and the integer decision variables can make the SMIP problems very difficult to solve. Our developed solution approach, which is based on the sample average approximation (SAA) method, will overcome these challenges and provide the optimal admission decision for clinics.

4) It is reported that the number of walk-in clinics is expected to increase from 1400 in 2012 to 2800 in 2015 with a saving of $800 million per year for the entire healthcare industry in the U.S. (Sussman, 2013). With our developed models and solution approaches, the results of this study can be applied to any clinics which provide services to walk-in patients. It can be

expected that our study can help these clinics to make the best walk-in patient admission decision, make sure that the admitted walk-in patients are served at the right time and thus lead to the substantial cost savings.

2. LITERATURE REVIEW

2.1. Patient no-shows and short-notice appointment cancellations

Many studies have identified patient no-shows as one of the most challenging problems associated with the healthcare industry in terms of operational planning (Cayirli and Veral, 2003; Gupta and Denton, 2008). It is claimed that clinic no-show rate varies among clinical departments, regions, and nations, with the range being between 5%-60% (Woodcock, 2003; Zeng *et al*; 2013). In addition, it can be shown that patient non-attendance disrupts the clinical flow operations in multiple ways. Not only it causes waste and disrupts clinical flows, but also it limits clinic access of other patient. This in turn leads to poor health outcomes, reduced revenues and lower clinical staff productivity (Zeng et al., 2013).

With the understanding of the side effect of patient no-shows, tremendous research has been conducted to discover the factors that affect patient no-show behavior. Among these studies, Lehmann et al. (2007) indicate that patients who have higher no-show risks are younger, born earlier in the year, and belonging to the ethnicity other than European. In addition, the authors indicate that follow-up patients are likely to be no-shows as compared to the new patients. The findings of the increased tendency of the patients belonging to the minority groups are also confirmed by Cohen et al. (2008). The authors divide the existing data in two groups (i.e., adults and children). The study indicates that among the group pertaining the children group, ethnicity and geographic considerations are significant. Additionally, the parents' perceptions on the physician also play a role in the determination of the no-show risk among this group. Among the adult group, it is indicated that gender, geographic consideration, time of the appointment, and the amount of time elapsed between the date of scheduling the appointment and actual appointment all play significant roles. The role of ethnicity is also confirmed by the study of

Dreiher et al. (2008). In addition, the authors indicate that the ethnic origin and time of the year might play a role in determining the no-show risk. The effect of the time of the year (i.e., seasons) might be linked with the weather conditions. Since the higher rate of no-show risks is associated with spring and winter, it might be indicated that weather conditions might also play a role in determining the no-show risk. In addition, González-Arévalo et al. (2009) point out that the patients who live far away from the hospital and the patients who have poor healthcare conditions and deteriorated mental status are more likely to become no-shows. Meanwhile, forgetting appointment time/date, hospitalization or death of the patient, not being notified about the appointment, or the change of appointment date/time are the main factors for the patients to miss the appointments. Ciechanowski et al. (2006) demonstrate that among the diabetes patients, the fearful and secure type of attachment in a non-depressed patient less likely leads to a no-show compared with that in a depressed patient. Similarly, Norris *et al*. (2014) indicate that among the main factors that affect the no-show rates, prior attendance history, age, and distance play the most significant role. Another interesting finding is that the clinical practices that aim to increase the utilization contribute to the increase of no-show rate.

Among the common factors mentioned above, patient forgetfulness, patient dissatisfaction due to extensive delays in accessing to the healthcare services, time conflicts, transportation related issues, patient's emotions, patient's perceptions regarding the merits of the healthcare system can all be considered to be the factors that contribute to the no-show rates (Goldman et al., 1982; Bean and Talaga, 1992; Garuda et al., 1998; Lacy et al., 2004; Zeng et al., 2013).

With the understanding of the significant factors related to patient no-show behaviors, various non-scheduling (as compare to the scheduling approaches mentioned above) intervention strategies such as sending out reminders, providing the transportation service, or charging the no-show fee are offered in the literature. For example, Woods (2010) indicates that the telephone reminders decrease the no-show rate by 50% (from 8% to 4%). Similarly, the notification letters as a form of intervention also decreases the no-show risks for the patient by 29%. These types of interventions can reduce the no-show problem to some extent (Macharia *et al.* 1992, Guy *et al.* 2012), but the elimination of patient no-shows through implementing these interventions is unlikely. Some clinics still report high non-attendance rates after implementing those measures, which are as much as 20% or even higher (Hashim *et al.* 2001, Geraghty *et al.* 2007; Liu *et al.,* 2013). In addition, Henry *et al.* (2012) in a study conducted in the VA hospital setting, indicate that intervention methods based for decreasing the no-show risk might work for some patient groups, while they might not be effective in some other. To cite an instance, the authors indicate that patients, who are not homeless, not in depression, and had five or more appointments in the last 6 months are associated with lower risk of no-show risk based on the automatic telephone reminder intervention. However, for the patients not having those attributes, the no-show risk is not affected by the notification schemes based on telephone reminder.

In addition to patient no-shows, short notice appointment cancelations also should be considered as impediments of clinic efficiency and patient accessibility to care. Usually short-notice appointment cancellations disrupt the clinic operations flow. Given the short time notice, it would be difficult to notify another patient about opening of the time slot of the cancelled appointment, and in most of the cases, the short notice appointment cancellations are somehow equivalent to patient no-shows with regard to the associated outcomes.

There are some studies in the literature that consider the appointment cancellation. Although the number is not as high as the no-show studies, they deserve special attention. As previously discussed, appointment cancellations especially with the short notice might disrupt the operations and lead to wastes and inefficiencies. In that regard, Olson *et al.* (1994) indicate that higher body mass index might be associated with the higher appointment cancellation among the female population. Additionally, Buckley *et al.* (2009) discuss the role of approaches that might be followed by the radiologist to reduce the appointment cancellation rates. The patients when referred to radiology department might fear for embarrassing situation. Parikh *et al.* (2010) indicate that live telephone reminders initiated by the staff members or auto reminders might help with decreasing the appointment cancellation risk.

To sum up, patient no-shows and short-notice appointment cancellations have significant adverse effect on the clinic efficiency and patient accessibility to care. They can disrupt the clinic flows, increase operation cost and decease the patient satisfaction and quality of care. Although non-scheduling intervention can reduce the no-show rate and appointment cancellation rate to some extent, it is unlikely that patient no-show and appointment cancellations could be completely eliminated. Hence, it is of great importance to apply scheduling interventions with the combination of non-scheduling intervention to further reduce the adverse effect of patient non-attendance. Note that the scheduling interventions are usually adopted through the appointment scheduling optimization process. In the following, we will review the existing literatures on the appointment scheduling optimization problems and the corresponding scheduling interventions.

2.2. Clinic appointment scheduling optimization

Appointment scheduling was introduced in outpatient clinics in the 1950s to reduce patient waiting time in clinics (Bailey 1952, Welch and Bailey 1952). Currently, most outpatient clinics in the U.S. use this approach to control the access to care services that they provide. The downside is that appointment scheduling may result in long lead time and long waiting lists for appointments (Murray and Tantau 1999, Pinto et al. 2002). The long lead time for appointments leads to more patient no-shows and late cancellations (Bean and Talaga 1995, Lacy et al. 2004, Lee et al. 2005), which waste clinical resources and thus increase healthcare costs. Meanwhile, several studies show that the long waiting lists for appointments prevent patients in acute conditions from seeing their own physicians in a timely manner, which undermines the accessibility to healthcare (Pinto et al. 2002, Murray and Berwick 2003). It should be noted that the U.S. is not alone in this regard – similar problems on appointment scheduling exist in the healthcare systems of other developed countries such as England (Kerssens et al. 2004; Schoen et al. 2004). Hence, the appointment scheduling optimization problem has receiving consistently attention from researchers and practitioners.

Queuing theory and simulation are the major quantitative methods used to evaluate and optimize appointment schedules. Among the two, simulation is more relevant to our work so we will survey the relevant studies in details. In early studies, simulation is primarily used to compare alternative appointment scheduling templates (or more commonly referred to as ASRs in the operations research literature) in outpatient clinics with respect to key system performance measures such as patient waiting time and provider idle time (Bailey, 1952; Fetter and Thompson, 1966; Vissers, 1979). In more recent studies, simulation experiments are conducted to help identify the most appropriate ASR with various environmental characteristics (Klassen and Rohleder, 1996; Rohleder and Klassen, 2000; Ho and Lau, 1992; Ho and Lau, 1999; Cayirli, Veral and Rosen, 2006). For example, Klassen and Rohleder (1996 and 2000) compare various ASRs under different distributions of patient service time, and illustrate that the optimal ASR depends on the mean and variance of the service time. Ho and Lau (1992 and 1996) compare various ASRs with different specifications of patient service time as well as no-show rate and the length of clinic sessions. Cayirli et al. (2006) further incorporate patient heterogeneity into the assessment of ASRs. Overall, these papers show that ASRs have significant effect on operational performance and that patient characteristics are important compounding factors, including walk-in rate, no-show rate, and arrival punctuality.

In recent years, optimization models have been developed for designing optimal ASRs in outpatient appointment scheduling. For example, Vanden Bosch et al. (1999) propose an efficient heuristic search algorithm to design an optimal ASR under the assumptions of independent service times following the identical Erlang distribution and punctual patient arrivals. Bosch and Dietz (2000 and 2001) extend the model by considering general phase-type distributed service times as well as patient no-shows. Kaandorp and Koole (2007) propose a stochastic optimization model with a multimodular objective function and presented a local search method based on the multimodularity of the objective function. Rohleder and Klassen (2000) apply simulation optimization for optimal ASR design with more flexible clinical settings. It is worth noting that most of the recent studies incorporate patient no-show uncertainty in their models.

2.3. Open access scheduling

In the past two decades, open access scheduling has been extensively studied. Murray and Tanau (1999) first propose the concept of open access scheduling to overcome the problem of high no-show rates in outpatient clinics. In the study, a successful case of open access scheduling in a clinic in the U.S. is demonstrated. Gupta et al. (2006) conduct an empirical study of clinics within Minneapolis metropolitan area that applies open access scheduling. It is pointed out that the factors, which include different practice styles of doctors, differences in panel compositions, and patient preferences, could hinder the successful sustaining of supply-demand balance. Several performance measures are proposed to help management for monitoring and evaluating the implementation of open access scheduling. There are also other publications that report the successful implementations of open access scheduling, which all indicate that open access scheduling is capable of reducing healthcare cost while improving the access to care, clinic resource utilization and patient satisfaction (Kennedy and Hsu, 2003; Murray et al., 2003; O'Hare and Corlett, 2004; Mallard et al., 2004; Bundy et al., 2005; Parente et al., 2005; O'Connor et al., 2006; Cameron S et al., 2010). In addition, Rose et al. (2011) conduct a systematic review on the performance of open access scheduling, which shows the benefits of reducing patient waiting time and no-show rate as well. Generally, the critical parameters for open access scheduling systems are determined based on experts' experiences rather than analytical methods. For instance, the percentage of open access appointments may range from 30% to 80% depending on the scheduler's experience (Herriott, 1999; Kennedy and Hsu, 2003; Murray and Tantau, 2000).

Besides the empirical studies, mathematical modeling approaches have also been widely applied to analyze the open access scheduling systems. Green et al. (2007) study the relationship between the panel size and the probability of "working overtime" or "extra work" for a provider in an open access clinic. The "extra work" is measured by the expected number of extra patients that a provider has to see in the open access clinic. Kopach et al. (2007) conduct a simulation study to evaluate the effects of open access scheduling on the continuity of care. It is concluded that the increasing fraction of open access patients have an adverse effect on the continuity of care, but the adverse effect could be mitigated by providers working as a team. Qu et al. (2007) develop a closed-form approach to quantitatively determine the optimal percentage of open access appointments to match daily provider capacity to demand. It shows that the optimal percentage of open access appointments mainly depends on the ratio of average demand for open access appointments to provider capacity and the ratio of the show-up rates for traditional and open access appointments. Liu et al. (2010) propose a dynamic programming model to study the heuristic policies of patient appointment scheduling by taking patient no-shows and cancellations into account. The results suggest that open access scheduling works best when the patient load is relatively low. Robinson et al. (2010) conduct a comparison study between traditional patient scheduling methods and open access scheduling. It is claimed that open access scheduling is significantly better than traditional methods in terms of patient waiting, provider idle and provider overtime. Lee and Yih (2010) conduct a simulation study to investigate the impact of open access configuration considering clinic setting conditions including demand variability, no-show rate, and the percentage of same-day appointments. The performance of different open access configurations is analyzed in terms of patient waiting time, patient rejection rate, and clinic utilization. Furthermore, Dobson et al. (2011) develop a stochastic model to evaluate the performance of open access scheduling in a primary care practice. It is found that encouraging routing patients to call for same-day appointment is a key element for the success of open access scheduling. Qu et al. (2011) propose a hybrid policy for open access scheduling, which consider two time horizons instead of one for the short-notice appointments. It is shown that the hybrid policy is no worse than the single time horizon policy in terms of the expectation and variance of the number of patients seen. Balasubramanian et al. (2012) propose a two stage stochastic integer programming model to maximize timely access and patient-physician continuity simultaneously for open access clinics. Qu et al. (2012) propose a mean-variance model to optimize the ratio of traditional versus open access appointments for open access scheduling systems. In addition, Patrick (2012) proposes a Markov decision model for determining optimal outpatient scheduling. In his study, open access scheduling is compared to the short booking window concept, and the latter appears to be more effective in term of cost minimization.

2.4. Overbooking

The overbooking strategy has long been studied in appointment scheduling optimization problems with the consideration of patient no-show and/or appointment cancellation. One outstanding study conducted by LaGanga and Lawrence (2007) examines the patient no-show problem and proposes overbooking strategy as the way to reduce the adverse effect of patient no-shows. In this study, a new utility function is developed, which can characterize the trade-offs between benefits and cost. The authors show that overbooking can increase patient access and productivity of providers, while increasing the patient waiting time and provider overtime significantly. The authors also show that the relative values that a clinic assigns to serving additional patients, minimizing patient waiting time, and minimizing provider overtime will determine whether overbooking is warranted. At last, it is concluded in the study that the overbooking strategy produce more utility for clinics with a larger patient population size, higher patient no-show rate, and lower service time variability. LaGanga and Lawrence (2012) extend their previous work and propose the flexible appointment scheduling model, which can mitigate the negative effect of patient no-show by overbooking policy. Their results show that the near-optimal overbooking appointment scheduling, which balances the extra benefits obtained from serving additional patient and the resulted cost associated with the possible patient waiting time and provider overtime, can be searched with a fast and effective solution procedure.

In addition, Daggy *et al*. (2010), based on a VA hospital setting, compare the performance of two scheduling approaches. The first approach does not consider the no-shows, while the second approach employs overbooking in the presence of no-shows. The results indicate that incorporating no-show probabilities in the clinic booking improves the patient flows and clinic efficiency. More recently, Samorani and LaGanga (2013) suggest a data mining approach to capture the no-show probability of individual patients and develop the optimal overbooking policy based on appointment characteristics on the appointment day. Furthermore, Giachetti (2008) conducts a simulation study, which suggests double book the habitual no-show patients whenever they make appointments, while Kros et al. (2009) conduct empirical study, which claims a saving of 95,000$ per semester after implementing the overbooking process instructed by a novel overbooking model.

To add more, Kim and Giachetti (2006) develop a stochastic mathematical appointment overbooking model (SMOM) to optimize the number of appointment that should be accept in regard to maximize the expected total profit. The SMOM approach is then compared with two other approaches, namely “no overbooking” and “naive statistical overbooking approach” (NSOA). The NSOA simply adds the mean number of no-shows minus the mean number of walk-ins to the number of appointment to accept. Their result shows that by comparing to the “no overbooking” approach, the SMOM can increase the expected total profit by 43.72%, while the NSOA can increase the corresponding profit by 29.66%. Similarly, Muthuraman and Lawley (2008) develop a stochastic overbooking model for an outpatient clinic with patient no-shows with the consideration of patient waiting time, provider overtime and the revenue collected from patients. In addition, Zeng et al. (2010) also propose a clinic scheduling model, which applies overbooking to compensate the negative impact of patient no-shows, in order to maximize the expected profit including revenue from patients and cost related with the patient waiting times and provider idle time. It is proved that the objective function, which is the expected profit, has different characteristics for patients with heterogeneous no-show probabilities and patients with homogeneous no-show probabilities. To be more specific, the objective function is multimodular for patients with homogeneous no-show probabilities. However, this property does not apply when patients hold different no-show probabilities. To extend the existing work, Huang and Zuniga (2012) develop a dynamic overbooking scheduling policy with the consideration of predicted no-show probability for individual patients. This policy is targeting at improving the patient access to care as well as reducing the patient waiting, while minimizing the clinic's cost. Their results show that the proposed dynamic overbooking scheduling policy outperforms the naive overbooking strategy, which overbook patients evenly throughout a clinic session.

2.5. Walk-in patient admission

Generally, the walk-in patients are referred to the patients who come to clinics and request for medical service without an appointment. It is shown that people persist in presenting to clinic and requesting immediate healthcare service, although there is no indication that a walk-in service is available in the clinic (Crismani and Galletly, 2011). However, the walk-in patients do not receive much attention from the researchers and healthcare practioners for the following reasons: 1) clinics do not want to admit walk-ins if they are running at their full capacity (i.e., sufficient patients with appointment); 2) providers do not feel the obligations to serve walk-in patients, since they do not have an appointment; 3) in the primary care clinic setting, it is speculated that the quality of care might be reduced, if the providers are not familiar with the walk-in patients (i.e., a provider is not the primary care physician whom the walk-in patients usually see). Among the limited literature related to walk-in patients, a few empirical studies show that the walk-in patients make an important piece of the entire patient population and walk-in patient friendly clinics are preferred by many consumers (Crismani and Galletly, 2011; Chmiel et al., 2011; Gail, 2007; Su and Shih, 2003;). With this prevision of walk-in patients, it can be concluded that quantitative methods will in urgent need for the scheduling optimization problem with the consideration of the walk-in patient admissions.

Although quantitative methods have been applied to analyzing and optimizing appointment scheduling systems (Cayirli and Veral, 2003; Gupta and Denton, 2008), most studies focus on outpatient clinics that do not accept walk-in patients. Only a few recent studies consider the appointment scheduling systems of the clinics accepting walk-in patients (Kim and Giachetti, 2006; LaGanga and Lawrence, 2008; Cayirli et al., 2012). For example, Kim and Giachetti (2006) develop a stochastic mathematical overbooking model with the consideration of probability distributions of patient no-shows and walk-ins. Their model determines the optimal number of appointments to be scheduled in order to maximize the total expected profits. LaGanga and Lawrence (2008) extend their heuristic algorithm in (LaGanga and Lawrence, 2012) to optimize appointment schedules in clinics accepting walk-in patients. Their results show that a modest number of scheduled appointments can significantly improve service quality and clinic performance compared to an all-walk-in clinic. Cayirli et al. (2012) propose a “Dome” appointment rule in which the disruptive effects of no-shows and walk-ins are considered. In the “Dome” appointment rule, appointment times are determined based on the mean and standard deviation of service times that are adjusted by the no-show rate and the walk-in rate. In these studies considering walk-in patients, it is assumed that a clinic admits all walk-in patients, and the same cost structure is assumed for scheduled patients and walk-in patients. However, in reality, many clinics actually have the flexibility of admitting walk-in patients selectively. For instance, a few primary care clinics that we worked with have the policy of seeing walk-in patients if possible, and their 5-month statistics show that 20% – 50% of walk-in patients were seen.

In addition, studies have shown that there are sufficient service requests (demands) from walk-in patients. For example, Howard et al. (2008) show that about one quarter of patients with their own primary care physician use the walk-in patient service in a six month period. For another example, a study in VA hospital reports that the investigated primary care clinic in a medical center encounters 45-72 walk-in patients per day on average (Phatak et al. 2011). Similarly, Huarng (2003) conducts a study, which shows that the ratio of walk-in patients could be as high as 55% in some outpatient clinic. Hence, admitting all arrived walk-in patients can make clinics running over capacity, which can dramatically increase the operation cost and lower the quality of care. In such clinics, a policy is needed to determine whether to admit a walk-in patient or not and in which slot a walk-in patient should be seen. This need motivates us to investigate walk-in patient admission policies in outpatient clinics.

3. TWO-STAGE STOCHASTIC PATIENT OVERBOOKING OPTIMIZATION MODEL

3.1. Introduction

Faced with the challenges of increased cost, limited capacity, and expanding demands for healthcare services, the outpatient clinics are running at dual objectives including stabilizing revenue and improving the healthcare access. It is the clear that the access is mostly determined by the appointment scheduling system. Traditionally, a patient, who needs an appointment, would call the clinic and be immediately booked with an appointment in a few days, weeks or months later, which is often referred as the lead time for an appointment. When there is sufficient patient demand, the clinics are most likely to operation at or close to their full capacity, which often means that no appointment is available in the near feature, i.e. the appointment lead time is long. However, ill patients usually cannot wait such a long time for an appointment. They will either enter the clinic as without an appointment or use the expensive Emergency Department. In addition, long lead time can cause the serious patient no-show problem, since the patient may have recovered, moved, forgot, or even died during this period. It is well-known that overbooking is an effective strategy for improving clinic accessibility and stabilizing revenue when a significant portion of patients choose not to show up for their appointments. In order to find the optimal overbooking strategy, a clinic needs to determine, given the stochastic characteristic of patient arrival pattern, patient non-attendance rate, and service time, (1) the optimal number of appointment slot to be overbooked; (2) where the overbooked appointment slots should be located.

In this study, we develop a two-stage mix-integer stochastic programming model to solve the clinic scheduling and overbooking optimization problem, which determines the optimal strategy for assigning appointment seeking patients to a fixed number of appointment slots with fixed length (typically 15, or 30 minutes). The model will also optimize the service start time of each patient show-up for their appointment, given different scenarios, which are subject to the stochastic patient arrival pattern, patient non-attendance rate, and patient service time. The objective is to minimize the operational cost, which is measured in terms of patient waiting time, provider idle time and provider overtime. Note that without overbooking, the resource of clinic will be under-utilized due to patient no-show, i.e., high provider idle time. However, too much overbooking can easily lead to the overloading problem, which causes high patient waiting time and provider overtime. Hence it is important to only overbook a moderate number of the appointment slot. More importantly, clinics need to overbook the right appointment slot, given the optimal number appointments to be overbooked. A wrong decision may reduce patient no-show but worsen the overloading problem, i.e., further decrease provider idle time and increase patient waiting time. For example, a clinic overbooks the first appointment, while only the patient in the last appointment is expected to be no-show. Clearly, by doing this, it will delay the process of all patients starting from the second appointment, if both patients in the first slot show up for their appointment.

3.2. Terminology and problem formulation

3.2.1. Terminology

Index:

i : Index of patients with appointment.

j : Index of providers.

ω : Scenario index.

Patients with appointment:

$p_{ij}^{1} = \{t_{ij}^{A1}, t_{ij}^{S1}, t_{ij}^{E1}, t_{ij}^{L1}, t_{ij}^{W1}, I_{ij}^{C1}, I_{ij}^{N1}\}$: Denote the first patient scheduled in the i^{th} appointment slot of provider j.

$p_{ij}^{2} = \{t_{ij}^{A2}, t_{ij}^{S2}, t_{ij}^{E2}, t_{ij}^{L2}, t_{ij}^{W2}\ I_{ij}^{C2}, I_{ij}^{N2}\}$: Denote the second patient scheduled in the i^{th} appointment slot of provider j.

Parameters:

I_{ij}^{C1}, I_{ij}^{C2} : The cancellation indicator of the corresponding patient, with "1" indicating cancellation.

I_{ij}^{N1}, I_{ij}^{N2} : The no-show indicator of the corresponding patient, with "1" indicating no-show.

T_{ij}^{S} : The expected starting time of the i^{th} appointment of provider j.

T_{j}^{E} : The expected service ending time of provider j.

$dummy\_t_{ij}^{A1}$, $dummy\_t_{ij}^{A2}$: Dummy arrival time of the corresponding patient, which is generated from given arrival time distribution.

$dummy\_t_{ij}^{L1}$, $dummy\_t_{ij}^{L2}$: Dummy service length of the corresponding patient, which is generated from a given service length distribution.

c^{Wait} : The cost coefficient related to patient waiting time.

c^{Idle} : The cost coefficient related to provider idle time.

$c^{Overtime}$: The cost coefficient related to provide over time.

Variables:

t_{ij}^{A1}, t_{ij}^{A2} : The arrival time of the corresponding patient.

t_{ij}^{S1}, t_{ij}^{S2} : The service starting time of the corresponding patient.

t_{ij}^{E1}, t_{ij}^{E2} : The service ending time of the corresponding patient.

t_{ij}^{L1}, t_{ij}^{L2} : The service length of the corresponding patient.

t_{ij}^{W1}, t_{ij}^{W2} : The waiting time of the corresponding patient.

I_{ij}^{A} : Appointment indicator; $I_{ij}^{A} = 1$, if at least 1 patient is scheduled in the ith appointment slot of provider j .

I_{ij}^{D} : Double-booking indicator; $I_{ij}^{D} = 1$, if the ith appointment slot of provider j is double-booked.

t_{j}^{I} : The idle time of provider j.

t_{j}^{O} : The overtime of provider j.

3.2.2. Problem formulation

For the purpose of solving the clinical overbooking optimization problem, we develop a two-stage stochastic mixed-integer programming (SMIP) model. For an introduction to stochastic integer programming, we refer to Birge and Louveaux (2001). In order to develop this SMIP model, we make the following assumptions.

1) Each clinical session is evenly divided into appointment slots. One or two patient appointment can be scheduled in one appointment slot. If two patients are booked in the same appointment slot, it will be referred as double-booking or overbooking.

2) Once an appointment is made, it cannot be modified unless it is cancelled by the patient.

3) Providers only see their own patients, i.e. patients scheduled for “provider A” will not be served by other providers in the clinic.

4) All patients with an appointment, who are not no-show or do not cancel their appointments, must be served by their corresponding provider within the clinic session, even if the provider has to work overtime.

5) Patients with earlier appointments, who are not no-show or do not cancel their appointments, will not be served later than patients with later appointments.

6) Providers do not serve any patients before the expected start time of the first appointment, i.e., the starting point of the clinic session.

Note that all the above assumptions are very basic and commonly made in the outpatient appointment scheduling literature (Cayirli and Veral, 2003, Gupta and Denton, 2008). More assumptions other assumes made in the existing literature on clinic overbooking optimization. For example, Kim and Giachetti (2006) propose a stochastic mathematical overbooking model (SMOM), which can determine the optimal number of appointment to be scheduled in advance to maximize the total expected profit. However, the model does not address the appointment allocation problem, i.e., which appointment slots should be used for the overbooking. In addition, the model requires a small variance for the service time distribution, in order to approximate the total service time as the mean service time of each appointment multiplied by the total number of appointment. For another example, Zacharia and Pinedo (2014) develop an appointment scheduling model with consideration of no-shows and overbooking. Their study investigates the single provider system and assumes punctual arrivals of patients at the start point of the appointment. In addition, Muthuraman and Lawley (2008) develop a stochastic overbooking model for outpatient clinical scheduling with no-shows. Their model only considers the single provider system. In addition, they assume that the service times follow the exponential distribution and patient arrives at the beginning of each appointment, i.e., the possibility of patient arrivals in the middle of appointment slot is ignored. To add more, LaGanga and Lawrence (2012) construct a flexible appointment scheduling model to mitigate the adverse effect of patient no-shows through overbooking strategy. Their study assumes the constant service time and punctual patient arrivals, if not no-show, at the starting point of the corresponding appointment.

By comparing with the existing clinic overbooking literature, our model has no constraints for patient arrival patterns. Hence, the patient arrival time can be assumed to follow any distributions that generate meaningful (non-negative) arrival time and service length, while some of the existing clinic overbooking studies assume the patients have to arrive punctually if they are not no-show or do not cancel their appointment (Zacharia and Pinedo, 2014; Muthuraman and Lawley, 2008; LaGanga and Lawrence, 2012). In addition, there are also no limitations about the service time distribution in our model. Hence, the model can be used for clinics with various service time distributions. Note that, some of the existing clinic overbooking studies rely on a specific service time distribution, or assume constant service time, which is often not realistic (Kim and Giachetti, 2006; Muthuraman and Lawley, 2008; LaGanga and Lawrence, 2012). Furthermore, our model does not have restrictions on the service start times of patients, except for the first one (see assumption No. 6), while some of the existing studies assume that provides don't see patient before the expected appointment start time (LaGanga and Lawrence, 2009). Therefore, our model have much less restrictions on the patient arrival patterns, service length distribution, as well as the service start time, as compared to the existing outpatient appointment overbooking studies. Meanwhile, our model contributes to the literature in the following ways. First, our model simultaneously can determine the optimal number of patients to be scheduled in the each appointment slot, and the patient sequence including the service starting time and ending time, in presence of patient arrivals on the day of service. Instead of finding a heuristic solution using simulation models, our model can derive the optimal solution or near optimal solution with tight optimality gaps. Secondly, unlike the queueing and dynamic programming models, our model requires no specific assumptions regarding the distributions of service time, no-show probability, appointment cancellation probability, the patient arrival time, or other model parameters. Third, our model addresses the problem with consideration with multiple providers, while most of the existing studies only consider overbooking problem with single provider.

The objective of the problem is to minimize the expectation of the weighted sum of patient waiting time, provider idle time and provider overtime in a clinic session, as shown in Eq. 3.1 (Cayirli and Veral, 2003). The patient waiting time is measured as the time difference between the actual appointment start time and expected appointments start time. In case that the actual appointment starts before the scheduled expected appointment start time, the waiting time will be defined as zero. The provider overtime is the time that provider worked after scheduled working hour. As for the provider idle time, it is defined as the amount of time that a provider is not seeing any patient during the scheduled working hour.

Min

$$\mathrm{E}_{\omega}\left(c^{Wait}\cdot\sum_{j}\sum_{i}\left(t_{ij}^{W1}(\omega)+t_{ij}^{W2}(\omega)\right)+\sum_{j}\left(c^{Overtime}\cdot t_{j}^{O}(\omega)+c^{Idle}\cdot t_{j}^{I}(\omega)\right)\right) \tag{3.1}$$

S.T.

$$t_{ij}^{L1}(\omega)=dummy\_t_{ij}^{L1}(\omega)\cdot I_{ij}^{A}\cdot\left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right),\forall i,j. \tag{3.2}$$

$$t_{ij}^{L2}(\omega)=dummy\_t_{ij}^{L2}(\omega)\cdot I_{ij}^{D}\cdot\left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right),\forall i,j. \tag{3.3}$$

$$t_{ij}^{A1}(\omega)=dummy\_t_{ij}^{A1}(\omega)\cdot I_{ij}^{A}\cdot\left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right),\forall i,j. \tag{3.4}$$

$$t_{ij}^{A2}(\omega)=dummy\_t_{ij}^{A2}(\omega)\cdot I_{ij}^{D}\cdot\left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right),\forall i,j. \tag{3.5}$$

$$I_{ij}^{A}\geq I_{ij}^{D},\forall i,j. \tag{3.6}$$

$$t_{ij}^{E1}(\omega)=t_{ij}^{S1}(\omega)+t_{ij}^{L1}(\omega),\forall i,j. \tag{3.7}$$

$$t_{ij}^{E2}(\omega)=t_{ij}^{S2}(\omega)+t_{ij}^{L2}(\omega),\forall i,j. \tag{3.8}$$

$$t_{ij}^{S1}(\omega)+M\cdot\left(1-I_{ij}^{A}\cdot\left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right)\right)\geq t_{i'j}^{E1}(\omega),\forall i,i',j,i>1,and\ i'<i. \tag{3.9}$$

$$t_{ij}^{S2}(\omega)+M\cdot\left(1-I_{ij}^{A}\cdot\left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right)\right)\geq t_{i'j}^{E2}(\omega),\forall i,i',j,i>1,and\ i'<i \tag{3.10}$$

$$t_{ij}^{S1}(\omega)+M\cdot\left(1-I_{ij}^{A}\cdot\left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right)\right)\geq t_{i'j}^{E2}(\omega),\forall i,i',j,i>1,and\ i'<i \tag{3.11}$$

$$t_{ij}^{S2}(\omega)+M\cdot\left(1-I_{ij}^{A}\cdot\left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right)\right)\geq t_{i'j}^{E1}(\omega),\forall i,i',j,i>1,and\ i'<i \tag{3.12}$$

$$t_{ij}^{S2}(\omega)+M\cdot\left(1-I_{ij}^{A}\cdot\left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right)\right)\geq t_{ij}^{E1}(\omega),\forall i,j. \tag{3.13}$$

$$t_{ij}^{W1}(\omega)\geq t_{ij}^{S1}(\omega)-T_{ij}^{S},\forall i,j. \tag{3.14}$$

$$t_{ij}^{W1}(\omega)\geq 0,\forall i,j. \tag{3.15}$$

$$t_{ij}^{W2}(\omega)\geq t_{ij}^{S2}(\omega)-T_{ij}^{S},\forall i,j. \tag{3.16}$$

$$t_{ij}^{W2}(\omega)\geq 0,\forall i,j. \tag{3.17}$$

$$t_{j}^{O}(\omega)\geq t_{ij}^{E1}(\omega)-T_{j}^{E},\forall i,j. \tag{3.18}$$

$$t_{j}^{O}(\omega)\geq t_{ij}^{E2}(\omega)-T_{j}^{E},\forall i,j. \tag{3.19}$$

$$t_{j}^{O}(\omega)\geq 0,\forall j. \tag{3.20}$$

$$t_{j}^{I}(\omega)\geq T_{j}^{E}-T_{1j}^{S}+t_{j}^{O}(\omega)-\sum_{i}\left(t_{ij}^{L1}(\omega)+t_{ij}^{L2}(\omega)\right),\forall j. \tag{3.21}$$

$$t_{j}^{I}(\omega)\geq 0,\forall j. \tag{3.22}$$

$$t_{ij}^{S1}(\omega) \geq t_{ij}^{A1}(\omega), \forall i, j. \tag{3.23}$$

$$t_{ij}^{S2}(\omega) \geq t_{ij}^{A2}(\omega), \forall i, j. \tag{3.24}$$

$$t_{ij}^{S1}(\omega) \geq T_{1j}^{S}, \forall i, j. \tag{3.25}$$

$$t_{ij}^{S2}(\omega) \geq T_{1j}^{S}, \forall i, j. \tag{3.26}$$

Beside the objective function, there are also 25 constraints, as defined by Eqs. 3.2 – 3.26. To be more specific, Eqs. 3.2 – 3.3 are the service length constraint, which make sure the service length equal to zero if the corresponding patient is no-show or cancels the appointment, or there is no such patient, i.e., $I_{ij}^{A} = 0$ or $I_{ij}^{D} = 0$; otherwise, the service length will be random drawn from a given distribution. Similarly, Eqs. 3.4 – 3.5 define the patient arrival time. If the corresponding patient is no-show or cancels the appointment, or there is no such patient, then the patient arrival time will be equal to zero; otherwise the patient arrival time will be drawn from a given distribution. Eq. 3.6 is the double-booking constraint. It defines that if an appointment slot is not single booked, then it cannot be double-booked. Meanwhile, Eqs. 3.7 – 3.8 define the relationship among the service start time, service ending time and service length. Eqs. 3.9 – 3.13 are the service commitment constraints. These constraints prevent a provider serving two or more patients at the same time based on the assumption that patients scheduled in early appointment slot should receive the service early, if they are not no-show or cancel the appointment. Eqs. 3.14 – 3.17 are the patient waiting time constraints. The patient waiting time is defined as the real service start time minus the expected service start time or zero, whichever is greater. Furthermore, Eqs. 3.18 – 3.20 define the provider overtime, which should equal to the ending time of the last patient less the expected ending time of the clinic session, or zero, whichever is greater. In addition, Eqs. 3.21 – 3.22 define the provider idle time, which equals the length of clinic length plus the provider overtime and minus the sum of service time for each patient seen in the clinic session. At last, Eqs. 3.23 – 3.26 are the service start time constraint, which define the service start time of a patient should be earlier than the arrival time of the corresponding patient, as well as the expected service start time of the first appointment.

The decision variables in this model can be divided into two stages. In the first stage, the decision variables are the appointment indicator and the double-booking indicator, i.e., I_{ij}^{A} and I_{ij}^{D}, which decide the number of patients who have been scheduled in each appointment slot for each provider. These patients with appointment are subject to a random arrival process and service process. The random arrival process is controlled by the no-show rate, cancellation rate, and arrival time distribution. Similarly, the service times of the patients are controlled by the service time distribution. Hence, the decision variables in the second stage are the service start times and ending times for the patients to minimize the weight sum of patient waiting time, provider idle time and provider overtime, for each scenario ω. Note that for a different scenario, there will be a different optimal second stage decision which includes the service starting times and ending times of patients, and thus a different second-stage objective value, which is defined by $c^{Wait} \cdot \sum_{j} \sum_{i} \left(t_{ij}^{W1}(\omega) + t_{ij}^{W2}(\omega) \right) + \sum_{j} \left(c^{Overtime} \cdot t_{j}^{O}(\omega) + c^{Idle} \cdot t_{j}^{I}(\omega) \right)$. As we can see in Eq. 3.1, the first-stage objective value is just the expectation of the second-stage objective value.

3.3. Solution approach

In this two-stage decision making problem, it is relatively easy to evaluate the objective function for a given first-stage decision under certain scenario. However, it could be extremely difficult to evaluate the expectation of the recourse function for a given first-stage decision. More specifically, the recourse problem is of different sizes under different scenarios, and the service times are defined with as continuous random variables, which may follow a variety of distributions (e.g., lognormal, exponential, uniform, triangle and etc.) depending on the patient group and service type (Cayirli and Veral, 2003; Bailey, 1952; Klassen and Rohleder, 1996; Cayirli et al., 2006). In addition, with the consideration of random patient arrival pattern, patient no-show and appointment cancellation, it is unlikely to develop an analytical formulation to estimate the expectation of the recourse function. As a matter of fact, for any two-stage stochastic mixed integer programing models, with continuous random variables, the expected recourse functions, in general, cannot be analytically derived as a function of the first-stage decision. Hence, we apply sample average approximation (SAA) approach to estimate the expected value of recourse function. To be more specific, the objective value function, shown in Eq. 3.1 will be estimated as

$\frac{1}{n}\sum_{\omega\in\Omega_s}\left(c^{Wait}\cdot\sum_j\sum_i\left(t_{ij}^{W1}(\omega)+t_{ij}^{W2}(\omega)\right)+\sum_j\left(c^{Overtime}\cdot t_j^{O}(\omega)+c^{Idle}\cdot t_j^{I}(\omega)\right)\right)$, where n is the sample size, and Ω_s is a set of sample scenarios randomly drawn from the entire scenario space. In this way, the model can be solved with the commercial mixed integer programming (MIP) optimization solvers, such as CPLEX and GAMS. For the state-of-the-art research on applying sample average approximation to SMIPs, we refer to Kleywegt et al. (2002), and Shapiro and Homem-de-Mello (2000). It is obvious that a large scenario sample size will lead to a better approximation of the expected value of recourse function and improve the solution quality. However, the computational time for find the optimal solution by using sample average approximation will be increase exponentially with the increase of scenario sample size drawn from the entire scenario space. A large sample size may lead to extremely large computation time (unacceptable) or even make it impossible to locate the optimal solution. Hence, it is important to select a reasonable sample size, which balances the computational time and solution quality.

3.4. Model verification

For model verification, we consider a single case, which has 2 providers and 3 30-minute appointment slots for each provider. It is assumed that the first stage decisions are given, which schedules only 1 patient for each appointment slot, i.e., $I_{ij}^{A}=1, I_{ij}^{D}=0, \mathit{for}\ i=1,2,3\ \mathit{and}\ j=1,2$. We assume that the clinic opens at time 0, while the providers start to work 30 minutes after the opening of the clinic. Hence, the expected service starting times for the 1st, 2nd and 3rd appointments of each provider are at 30, 60, and 90 minutes, respectively. In addition, the expected ending time for this clinic session is 120 minutes, since each appointment is 30 minutes. In total, 10 different scenarios are considered. For the scenarios 1 – 5, we assume that the patient no-show rate and cancellation rate are 0, and patient arrives randomly within 30 minutes (uniform distribution) before the expected appointment start time. It is also assumed that the service length is uniformly distributed from 20 to 35 minutes. For the scenarios 6 – 10, we assume that the patient no-show rate and cancellation rate are both 0.2, and patient can arrives with a lead time, which is exponentially distributed with mean equal to 30 minutes. Hence the arrival time will be equal to expected service starting time of the patient less the lead time. Note that if the arrival time is less than 0, it will be adjusted to 0. In addition, we also assume the service length to follow the exponential distribution with the mean equal to 25 minutes. Note that for all 10 scenarios, the cost coefficients related to patient waiting time, provider idle time and provider overtime are all equal to 1. For each scenario, the optimal service starting time and ending time obtained through the model are compared with the corresponding values obtained from manual calculation. In addition, the achieved optimal objective value from the model is compared with the corresponding optimal objective value achieved from manual calculation. The comparison results are shown in Tables 3.1 and 3.2. As we can see, for all scenarios, the results obtained from the model, including the service starting time and ending time for each patient, as well as the objective values, are the same as the result obtained through manual calculation. Note that for some scenarios in Table 3.1, the service length and the corresponding patient arrival time are equal zero. The reason for this is that these patients either are no-show or cancel their appointments, as indicated by the no-show/cancellation indicator. In addition, the service start time for patients who are no-shows or cancel their appointments, is also defined as the expected start time for the first appointment, which is 30 minutes in our case. Note that it will not influence the service starting time and ending time of other patients, by defining the arrival time, service length and service start time for patients who do not actually receive service from providers in this way.

Table 3.1: Comparison result for scenarios 1 – 5

		Scenario1		Scenario2		Scenario3		Scenario4		Scenario5	
		Pvd1	Pvd2	Pvd1	Pvd2	Pvd1	Pvd2	Pvd1	Pvd2	Pvd1	Pvd2
Arrival time (min)	Appt1	0	27	13	10	3	11	24	28	3	24
	Appt2	49	49	47	44	56	46	60	49	44	60
	Appt3	86	61	63	64	81	88	60	60	67	75
No-show/ Cancellation Indicator	Appt1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	Appt2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	Appt3	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Service starting time (min)	Appt1	30*	30*	30*	30*	30*	30*	30*	30*	30*	30*
	Appt2	61*	62*	63*	59*	59*	51*	60*	51*	51*	60*
	Appt3	88*	92*	97*	91*	93*	88*	90*	71*	78*	92*
Service length (min)	Appt1	31	32	33	29	29	21	24	21	21	27
	Appt2	27	30	34	32	34	20	30	20	27	32
	Appt3	26	29	34	21	25	25	32	29	24	22
Service ending time (min)	Appt1	61*	62*	63*	59*	59*	51*	54*	51*	51*	57*
	Appt2	88*	92*	97*	91*	93*	71*	90*	71*	78*	92*
	Appt3	114*	121*	131*	112*	118*	113*	122*	100*	102*	114*
Objective value (min)	Wait	5*		11*		3*		0*		2*	
	Idle	6*		8*		26*		26*		27*	
	Overtime	1*		11*		0*		2*		0*	
	total	12*		30*		29*		28*		29*	

* indicates that the results from model and manual calculation are equal

Table 3.2: Comparison result for scenarios 6 – 10

		Scenario6		Scenario7		Scenario8		Scenario9		Scenario10	
		Pvd1	Pvd2	Pvd1	Pvd2	Pvd1	Pvd2	Pvd1	Pvd2	Pvd1	Pvd2
Arrival time (min)	Appt1	24	0	0	0	7	0	0	0	0	7
	Appt2	21	0	59	45	43	51	45	0	0	23
	Appt3	61	47	84	74	0	88	0	0	0	85
No-show/ Cancellation Indicator	Appt1	0/0	1/0	1/0	0/0	0/0	0/1	1/0	0/1	1/0	0/0
	Appt2	0/0	0/1	0/0	0/0	0/0	0/0	0/0	0/1	0/1	0/0
	Appt3	0/0	0/0	0/0	0/0	0/1	0/0	1/0	1/0	0/1	0/0
Service starting time (min)	Appt1	30^*	30^*	30^*	30^*	30^*	30^*	30^*	30^*	30^*	30^*
	Appt2	68^*	30^*	59^*	45^*	46^*	51^*	45^*	30^*	30^*	60^*
	Appt3	93^*	47^*	84^*	82^*	30^*	88^*	30^*	30^*	30^*	98^*
Service length (min)	Appt1	38	0	0	5	16	0	0	0	0	30
	Appt2	25	0	25	37	37	22	20	0	0	38
	Appt3	22	39	71	35	0	43	0	0	0	33
Service ending time (min)	Appt1	68^*	30^*	30^*	35^*	46^*	30^*	30^*	30^*	30^*	60^*
	Appt2	93^*	30^*	84^*	82^*	83^*	73^*	65^*	30^*	30^*	98^*
	Appt3	115^*	86^*	155^*	117^*	30^*	131^*	30^*	30^*	30^*	131^*
Objective value (min)	Wait	11^*		0^*		0^*		0^*		8^*	
	Idle	56^*		42^*		73^*		160^*		90^*	
	Overtime	0^*		35^*		11^*		0^*		11^*	
	total	67^*		77^*		84^*		160^*		109^*	

* indicate that the results from model and manual calculation are equal

In order to further verify our model with consideration of double-booking, we develop another 10 different scenarios, namely, scenario 11 – 20. As compared to the first 10 scenarios, scenarios 11 – 15 have exactly the setting as scenarios 1 – 5, except for the first-stage decision values have been changed to $I_{ij}^{A} = 1, I_{ij}^{D} = 1, for\ i = 1,2,3\ and\ j = 1,2$. Similarly, scenarios 16 – 20 have exactly the setting as scenarios 6 – 10, except for the same change of first-stage decision values. Note that the change of first-stage decision indicate the all appointment slots have been booked with two patients, i.e. double-booking. For the purpose of model verification, we also compare the optimal solution obtained from the model to the optimal solution obtained by manual calculation. The comparison results are shown in Tables 3.3 and 3.4. As we can see, for all the scenarios the optimal solution obtained from our model are exactly the same as the model obtained from manual calculation. To sum up, our model can derive the optimal start and ending times, and optimal objective value accurately for any given first-stage decision, service length distribution, patient arrive pattern, patient no-show rate and appointment cancellation rate. Hence, the validation of this model is verified. Note that although we fix the first decision variables and apply them as constant for model verification, the model will be able to find optimal first-stage decision variables as long as the first-stage decision variables are longer fixed. This is because for each given first-stage decision variables, the accurate value of objective function can be estimated, as shown in the model verification. The optimal first-stage decision will be the one that leads to the minimal value of the objective function. However, the solution space will be increased and more computing time will be needed for finding the optimal first-stage decision variable. In the following, a numerical analysis will be conducted, which uses a more realistic clinic settings, including the number of appointment slot, number of provider, patient arrival pattern, service length, patient no-show and appointment cancellation rate.

Table 3.3: Comparison result for scenarios 11 – 15

		Scenario11				Scenario12				Scenario13				Scenario14				Scenario15			
		Pvd1		Pvd2		Pvd1		Pvd2		Pvd1		Pvd2		Pvd1		Pvd2		Pvd1		Pvd2	
Arrival time (min)	Appt1	3	24	3	17	28	30	14	16	5	27	0	22	17	24	8	25	2	9	5	19
	Appt2	34	39	37	42	34	54	36	53	47	53	31	57	45	49	34	51	35	45	40	52
	Appt3	70	86	81	83	72	74	70	80	65	82	63	72	72	84	61	88	66	88	70	87
No-show/ Cancellation Indicator	Appt1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	Appt2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	Appt3	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Service starting time (min)	Appt1	30*	61*	30*	59*	30*	59*	30*	60*	30*	55*	30*	57*	30*	50*	30*	55*	30*	64*	30*	57*
	Appt2	86*	115*	89*	112*	85*	108*	94*	125*	86*	116*	87*	118*	77*	102*	87*	115*	87*	110*	89*	124*
	Appt3	139*	160*	143*	168*	143*	167*	152*	174*	147*	174*	149*	179*	129*	159*	148*	175*	139*	162*	156*	182*
Service length (min)	Appt1	31	25	29	30	29	26	30	34	25	31	27	30	20	27	25	32	34	23	27	32
	Appt2	29	24	23	31	23	35	31	27	30	31	31	31	25	27	28	33	23	29	35	32
	Appt3	21	23	25	33	24	34	22	25	27	25	30	25	30	34	27	26	23	28	26	28
Service ending time (min)	Appt1	61*	86*	59*	89*	59*	85*	60*	94*	55*	86*	57*	87*	50*	77*	55*	87*	64*	87*	57*	89*
	Appt2	115*	139*	112*	143*	108*	143*	125*	152*	116*	147*	118*	149*	102*	129*	115*	148*	110*	139*	124*	156*
	Appt3	160*	183*	168*	201*	167*	201*	174*	199*	174*	199*	179*	204*	159*	193*	175*	201*	162*	190*	182*	210*
Objective value (min)	Wait	472*				507*				508*				437*				510*			
	Idle	0*				0*				0*				0*				0*			
	OT	144*				160*				163*				154*				160*			
	total	616*				667*				671*				591*				670*			

* indicate that results from model and manual calculation are equal

Table 3.4: Comparison result for scenarios 16 – 20

		Scenario16				Scenario17				Scenario18				Scenario19				Scenario20			
		Pvd1		Pvd2		Pvd1		Pvd2		Pvd1		Pvd2		Pvd1		Pvd2		Pvd1		Pvd2	
Arrival time (min)	Appt1	0	0	27	19	0	0	0	9	0	16	26	23	0	11	6	0	27	0	21	0
	Appt2	30	0	0	0	47	35	0	0	14	48	40	33	0	0	50	0	10	6	0	34
	Appt3	87	82	0	62	81	0	6	88	0	69	76	0	0	0	83	0	0	51	61	58
No-show/ Cancellation Indicator	Appt1	1/0	0/0	0/0	0/0	0/1	0/0	0/1	0/0	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	Appt2	0/0	1/0	1/0	1/0	0/0	0/0	0/1	1/0	0/0	0/0	0/0	0/0	0/0	0/1	0/0	1/0	0/0	0/0	0/0	0/1
	Appt3	0/0	0/0	1/0	0/0	0/0	1/0	0/0	0/0	0/1	0/0	0/0	1/0	1/0	0/1	0/0	0/1	1/0	0/0	0/0	0/0
Service starting time (min)	Appt1	30*	30*	30*	56*	30*	30*	30*	30*	30*	30*	30*	46*	30*	80*	30*	35*	30*	93*	30*	30*
	Appt2	42*	30*	30*	30*	55*	94*	30*	30*	90*	148*	66*	81*	130*	30*	50*	30*	104*	139*	51*	114*
	Appt3	87*	96*	30*	62*	131*	30*	41*	88*	30*	158*	106*	30*	30*	30*	97*	30*	30*	161*	122*	135*
Service length (min)	Appt1	0	12	26	2	0	25	0	11	0	60	16	20	50	50	5	8	63	11	21	0
	Appt2	18	0	0	0	39	37	0	0	58	10	15	25	14	0	47	0	35	22	63	8
	Appt3	9	5	0	27	48	0	39	54	0	24	85	0	0	0	3	0	0	6	13	55
Service ending time (min)	Appt1	30*	42*	56*	58*	30*	55*	30*	41*	30*	90*	46*	66*	80*	130*	35*	43*	93*	104*	51*	30*
	Appt2	60*	30*	30*	30*	94*	131*	30*	30*	148*	158*	81*	106*	144*	0*	97*	30*	139*	161*	114*	122*
	Appt3	96*	101*	30*	89*	179*	30*	80*	142*	30*	182*	191*	30*	30*	0*	100*	30*	30*	167*	135*	190*
Objective value (min)	Wait	32*				75*				245*				132*				388*			
	Idle	81*				8*				0*				27*				0*			
	OT	0*				81*				133*				24*				117*			
	total	113*				164*				378*				183*				505*			

* indicate that results from model and manual calculation are equal

3.5. Numerical analysis

In this section, we conduct a numerical analysis based on the proposed model to illustrate how the overbooking strategy should be applied to overcome the adverse effect of patient no-show and short notice cancellation. In order to show that how different patient characteristics could affect the optimal overbooking strategy, we consider different patient arrival patterns, server length distributions, patient no-show rates, appointment cancellation rates, and cost coefficients.

3.5.1. Data collection and study design

For the numerical analysis, we first construct a base case (case 0) to illustrate the effectiveness of the proposed SMIP model to optimize the overbooking strategy. For the base case, we consider a clinic session of 4 hours, which are evenly divided into 8 30-minute appointment slots. It is assumed that two providers work at the same time in the clinic session. The parameters, such as patient arrival time, service time distribution, no-show rate and cancellation rate, used in the base model are chosen based on the data in the literature, as well as the data collected in an outpatient clinic in a local hospital. In the following, we summarize the related literature, as well as the parameter choices in our base case.

First of all, regarding the non-attendance rate (which include no-shows and cancellations), the literature reports the following,

1) Johnson et al. (2007) indicate that the no-show rate vary from 3% to 42%, with an average of 17%.

2) George and Rubin (2003) report that the non-attendance rate (no-shows and cancellations) in U.S. primary care clinics range from 5% to 55%.

3) Al-Shammari (1992) and Hermoni et al. (1990) report non-attendance rates of 29.5% and 36%, respectively.

4) Moore et al. (2001) suggest that no-shows and cancelled appointments combined amount 31.1% of appointments.

In the numerical analysis, we consider three levels of patient non-attendance rate for patients, namely, low non-attendance rate (no-show: 3%, cancellation: 2%), medium non-attendance rate (no-show: 17%, cancellation: 13%), and high non-attendance rate (no-show: 42%, cancellation: 13%). Note that in the case of medium non-attendance rate, we consider the mean non-attendance rate of 30%, which is the average of the lower bound (5%) and upper bound (55%) of the non-attendance rate in U.S. primary care clinics. In the base case, the medium attendance rate will be used.

Secondly, regarding the service time distribution, the literature reports the following:

1) LaGanga and Lawrence (2012) assume a constant patient service time, which equals to the length of an appointment slot in their clinic overbooking study.

2) Qu et al. (2013) assume the patient service time follows the lognormal distribution in their outpatient appointment optimization study based on a Women's clinic.

3) Jing et al. (2014) assume the patient service time follows the Gamma distribution in their patient flow simulation study based on a local VA medical center.

As we can see, there are various assumptions for the patient service time distribution, which are related to the studied clinics and patient groups. In the numerical study, we can consider three different service time distributions, which are Gamma (2.9898, 9.10383) minutes, Lognormal(3.0479, 0.71566) minutes, and constant 27.22 minutes. Parameters for the Gamma distribution are estimated based on the service time data collected from a local VA medical center, while the parameters for the lognormal distribution are calculated by using the same mean and double variance as the Gamma distribution. The constant 27.22 minutes are chosen based on the mean service time drawn from the Gamma distribution. In the base case, we use the above Gamma distribution as the service time for all arrived patients.

Thirdly, regarding the patient arrival patterns, the literature reports the following:

1) There are many studies that use the exponential distribution to model the interval time of patients, although the parameter chosen for the exponential distribution may vary based on the clinics characteristic and patient population (Alexopoulos et al., 2008).

2) Fontantesi et al. (2002) point out that the assumption of exponentially distributed patient interval time is not realistic for many clinics. They further indicate that patient arrivals tend to be "clumped" due to the common busy schedule, traffic light time, and availability of parking space. According to their study, most patients arrive 15 minutes earlier or 10 minutes late for their appointment. On average, the patients arrive 3 – 4 minutes before their scheduled appointment time.

3) Parmessar (2010) apply the appointment driven arrival in his simulation study for appointment optimization. The appointment driven arrivals assume that the patients should be arriving within a certain time interval, which is based on the schedule appointment time. For example, if a patient has an appointment at time t_0, then the patient will arrive at a random time drawn from the interval $[t_0 - a, t_0 + b]$, where a and b are positive constant that determine the width of the interval.

As we can see, there are various assumptions about the patient arrival patterns. In our numerical analysis, we tentatively consider three different types of patient arrival pattern, which are driven by the scheduled appointment time. For the first patient arrival pattern, we assume patients (except for no-show and appointment cancellation) arrive within 2 hours before the scheduled appointment and we assume the arrival lead time allow the Uniform distribution, i.e.

Uniform (0,120) minutes. Note that the actual patient arrival time of will equal to scheduled appointment time less the lead time, or zero, whichever is larger. For the second patient arrival pattern, we assume the arrival lead time follows the exponential distribution, with the mean equal to 4 minutes, as reported in the above mentioned study conducted by Fontantesi et al. (2002). For the third arrival pattern, we assume that patients (except for no-show and appointment cancellation) arrive punctually for their appointments, i.e., the lead time will hold as constant 0. For the base case, we use the second patient arrival pattern, i.e., arrival lead time equal to exponential(4) minutes.

The cost coefficients are chosen based on the hourly wages of all occupation and primary providers in United States. According to the Bureaus of Labor Statistics (BLS, 2013), the 10th, 50th and 90th percentiles of national hourly wage in 2012 are $8.7, $16.71, and $41.74, respectively, over all U.S. industry sectors. The average hourly wage of Family and General Practitioners is $86.95 in 2012. In addition, by considering the compensation for providers to work overtime, the hourly wage for providers working overtime is assumed to be 1.5 times of the regular hourly wage. Thus, in the case study, three sets of the cost coefficients for patient waiting time, provider idle time, and provider overtime are considered. The three ratios are 1:10:15, 1:5.2:7.8 and 1:2.1:3.1 corresponding to the 10th, 50th and 90th percentiles of national hourly wage, respectively. In the base case, we consider the ratio that corresponds to the 50% of the percentile of national hourly wage, i.e., 1:5.2:7.8.

As shown in Table 3.5, the parameters used for the base case are summarized. Beside the base case, another 8 cases are also developed in order to investigate how parameter selection would influence the resulted optimal overbooking strategy. As compare to the base case, each of the other eight cases represents a certain extreme condition by changing only one or a few parameters from the base case.

1) Cases 1 & 2 represent the situations of high attendance rate and low attendance rate, respectively, by altering the no-show rate and cancellation rate, simultaneously.

2) Cases 3 & 4 represent the situations high variance and low variance of patient service time by altering the patient service time distribution.

3) Cases 5 & 6 represent the situations of high variance and low variance of patient arrival time, respectively, by altering the patient arrival lead time distribution.

4) Cases 7 & 8 illustrate the situation of provider seeing low-income patients and high-income patients, respectively, by altering the cost coefficients, i.e. c^{Wait}, c^{Idle} and $c^{Overtime}$.

The altered parameters for Cases 1-8 are shown in Table 3.6. Note that, in each case, except for the altered parameters, all the remaining parameters are the same as those in the base case. For example, in Case 8, the cost coefficients are changed to 1:2.1:3.1, which correspond to the 90th percentile of national hourly wage. However, all other parameters remain the same as Base case.

Table 3.5: Parameters used for the base case

Parameters	Rate/Distribution	Parameters	Rate/Distribution
Session length	4 hours	Service time distributions	Gamma(2.9898, 9.10383) minutes
Number of appointment for each provider	8	Patient arrival lead time distribution	Exponential(4) minutes
Length of each appointment	30 minutes	c^{Wait}	1
No-show rate	17%	c^{Idle}	5.2
Cancellation rate	13%	$c^{Overtime}$	7.8

Table 3.6: Parameter adjustments for Cases 1-8 compared with Base case

Case number	parameters	Rate Distribution
Case 1	No-show rate	3%
	Cancellation rate	2%
Case 2	No-show rate	42%
	Cancellation rate	13%
Case 3	Service time distributions	Lognormal(3.0479, 0.71566) min
Case 4	Service time distributions	27.22 min
Case 5	Patient arrival lead time distribution	Uniform (0,120) min
Case 6	Patient arrival lead time distribution	0 min
Case 7	c^{Wait}	1
	c^{Idle}	10
	$c^{Overtime}$	15
Case 8	c^{Wait}	1
	c^{Idle}	2.1
	$c^{Overtime}$	3.1

3.5.2. Numerical analysis result

The proposed two-stage mixed integer stochastic linear programming model is solved through the use of CPLEX solver for each case as presented in above. The solver is run on a personal computer with an Intel 2.67GHz i5 dual-core processor and 2.9GB RAM. It takes less than 1 minute to find the optimal solution for each of the 9 cases (base case plus the other eight cases). The optimal overbooking strategy found for the nine cases are shown in Table 3.7, and the descriptive performance statistics of each overbooking strategy are also presented.

Table 3.7: Optimal overbooking strategy for Cases 0 – 8

Slot Index	Case 0		Case 1		Case2		Case3		Case4		Case5		Case6		Case7		Case8	
	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
1	S	D	S	S	D	D	D	D	D	S	S	S	D	D	D	D	S	S
2	D	S	S	S	D	D	S	S	S	D	S	S	S	S	S	D	S	S
3	S	S	S	S	D	D	S	D	S	S	D	D	D	S	D	S	S	S
4	D	S	S	S	D	D	D	S	D	S	S	S	S	S	S	S	S	S
5	S	S	S	S	D	D	S	S	S	D	S	S	S	S	S	S	S	S
6	S	S	S	S	D	D	S	S	S	S	D	D	S	S	S	S	S	S
7	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
8	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
# of overbooking	3		0		12		4		4		4		3		4		0	

"S" means single booking, "D" means double-booking.

Clearly, the optimal overbooking strategies are different among the 9 cases. For instance, Cases 0, 1 and 2, have medium, high, and low patient attendance rate, respectively. The corresponding optimal overbooking strategies show that for high patient attendance rate (95%), no overbooking is needed; for medium patient attendance rate (70%), three out of sixteen appointment slots (18.75%) should be double-booked; for low patient attendance rate, twelve out of sixteen appointment slots (75%) should be double-booked. It can be seen that with the decreasing patient attendance rate, the optimal number of double-booked appointment should increase. Similarly, Cases 0, 3 and 4, present medium, high, and low patient service time variance. By compare the number of double-booked appointment slots among the three cases, no significant impact of service time variance on the optimal number of double-booked appointment slots is found. In other word, the service time variance doesn't influence the optimal number of double-booked appointment slots. In addition, Cases 0, 5 and 6 represent medium, high and low patient arrival time variability. Similarly, no significant impact of patient arrival pattern can be found on the optimal number of double-booked appointment slots. At last, Cases 0, 7 and 8 represent the situation of medium, low, and high income patients. As we can see, the optimal number of double-booking has been reduced significantly for high income patients (Case 8). This is because the double-booking may significantly increase patient waiting time and the waiting cost for high income patients (low cost coefficient ratio) is more valuable as compare to the waiting cost of low income patients (high cost coefficient ratio). Therefore, high income patients are less willing to accept longer waiting time caused by double-booking. On the other hand, a high cost coefficient ratio indicates the gap of time values between the patients and the provider is large. This will lead to more double-booking, because the patients' time is less valuable and it is better to make more double-booking to avoid provider idle rather than patient waiting.

Table 3.8: Summary of descriptive performance statistics for Cases 0 – 8

		Case0	Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case8
Objective	mean	981.8	781.4	1168.9	1156.7	681.9	806.3	1093.9	1797.8	480.5
	std of mean	9.96	6.80	7.87	11.12	6.45	9.34	9.61	12.93	5.45
Patient waiting/min	mean	169.1	117.3	205.2	221.5	122.2	103.8	198.4	254.2	53.0
	std of mean	6.26	2.73	2.52	5.38	2.28	3.70	4.28	5.21	2.01
Provider idle /min	mean	132.9	96.7	160.1	132.8	100.4	110.0	134.6	114.9	181.0
	std of mean	3.08	2.32	3.11	3.40	2.60	3.27	3.30	3.13	3.24
Provider overtime/min	mean	15.6	20.7	16.8	31.4	4.8	16.7	25.0	26.3	15.3
	std of mean	1.17	1.185	0.835	2.11	0.42	1.515	1.5	1.68	1.067

In Table 3.8 the mean, as well as the standard deviation of performance metrics including patient waiting time, provider idle time, provider overtime, and objective function value, are shown for each case. The results reveal a few interesting phenomena which are commonly seen in practice. For instance, Case 1 and Case 0 have the same clinic settings except for the attendance rate, where it is higher for Case 1. The statistics indicate that Case 1 has a lower objective value compared with Case 0, which supports the general concept that high attendance rates are preferred in clinics. This concept can also be revealed by comparing Case 2 with Case 0, where the attendance rate is higher for Case 0. For another instance, Case 3 and Case 4 have the same clinic setting except for the service time distribution, where the variance is higher for Case 3. The statistics indicate that Case 3 has a higher objective value, as well as the patient waiting time, provider idle time, and provider overtime. This supports the general concept that clinics want to standardize their procedure and reduce the service time variability. In addition, Case 5 and Case 6 have the same clinic setting except for the patient arrival pattern, where the patients tend to arrive earlier for Case 5. The statistics indicate that Case 5 has a lower objective value, which supports the general concept that clinics want patient to arrive early for their appointments. To add more, Case 7 and Case 8 also have the same clinic settings except for the cost coefficient ratio, where Case 7 represents the scenario of low-income patients by using a high ratio. The statistics indicate that Case 7 have a higher objective value compared with Case 8. The implication is that high-income patients are preferred by the clinics.

In Figure 3.1, we compare the average waiting time of patients in single booked slots against the waiting time of patients in double-booked slots. As we can see, for all cases except for Case 1 and Case 8, the waiting time of patients in double-booked appointment slots are significantly higher. In addition, for some cases, such as Cases 0, 2, 4, and 5, the waiting time of patients in the doubled booked slots is more than twice as much as the waiting time of patients in the single booked slots. The implication here is that although double-booking can resolve the patient no-show problem it can dramatically increase the waiting time of patients who are scheduled in the double-booked appointment slots. Note that there are no double-booked slots for Case 1 and Case 8. Hence, the waiting time of patients in the doubled booked slots is set to zero for both case.

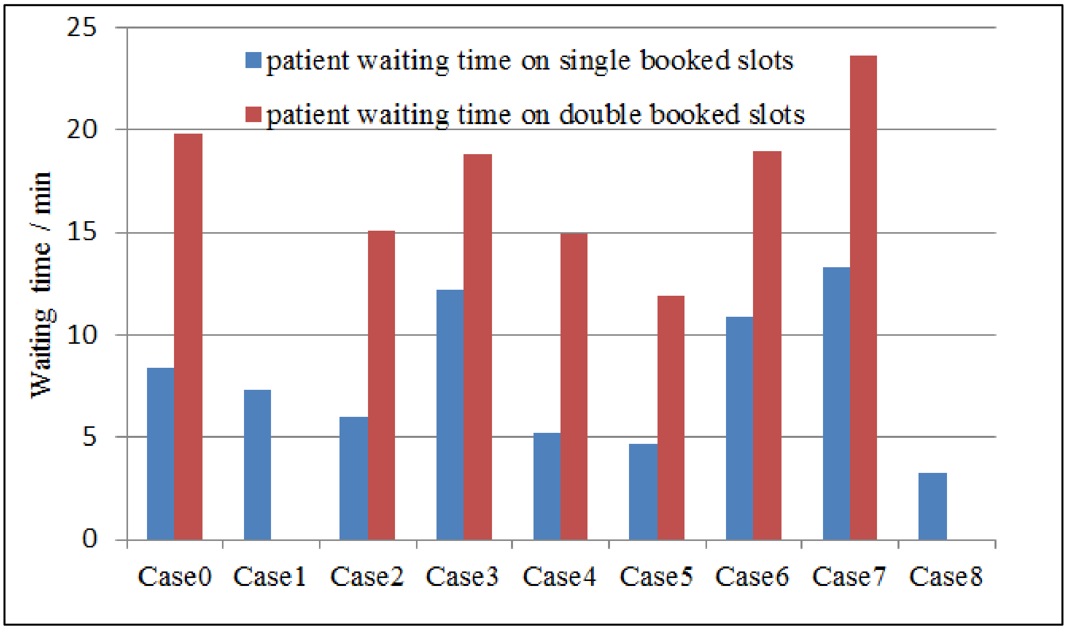

Fig. 3.1: Average patient waiting time – double-booked slots vs. single booked slots

In Figure 3.2, the average patient waiting time with respect to appointment slots is shown. Clearly, the average waiting times of patients are not equal for different slots. As we can see, these waiting times appear to form a “hump”, whereas the waiting times for slots 2, 3, 4, 5 and 6 are high and the waiting times for slots 1, 7 and 8 are low. It can be concluded that, for patient with appointment in the middle of a clinic session, such as appointments 3 and 4, their average waiting times are higher than the waiting times of those patients who have their appointment at the beginning and ending of a clinic session, such as appointments 1 and 8.

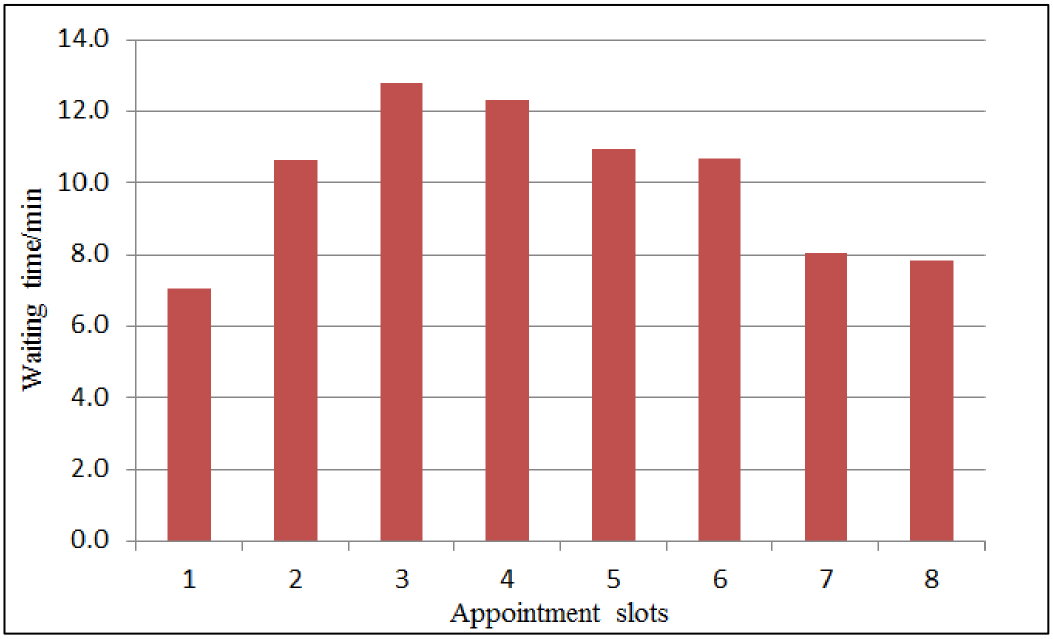

Fig. 3.2: Average waiting time over appointment slots

Through the numerical study, we show that our model can be used to effectively find out the optimal double-booking strategy for cases of different patient arrival pattern, no-show rate, appointment cancellation rate and service time distribution. In addition, we show that with the increase of patient no-show and appointment cancellation rate, the number of double-booked appointment slot should also increase. However, for patients with high income, the double-booking is not a good strategy, since it can dramatically increase the patient waiting time, which is of high value for high income patients. In the following, a sensitivity analysis is conducted on patient no-show rate and appointment cancellation rate to quantitatively investigate their effect on the objective function value as well as the optimal number of appointment slot to be double-booked.

3.5.3. Adverse effect of patient no-show and appointment cancellation

To further investigate how the no-show rate and appointment cancellation rate could affect the optimal number of double-booked appointment slots and the objective function value, which is the weighted sum of patient waiting time, provider idle time and provider overtime, we consider a case with punctual patient arrival (if not no-show or appointment cancellation), and constant service time 30 minutes. Note that all other parameters, except for no-show and appointment cancellation rate, are set the same as the base case. In studying the effect of no-show rate, the appointment cancellation rate is fixed as zero, vice versa. This way, the coupling effect of patient arrival time variability and service time variability can be eliminated. Since patient no-show and appointment cancellation make no difference in our model, we only discuss the effect of patient no-show in the following. The effect of appointment cancellation is expected to be the same as the effect of patient no-show.

In Figure 3.3, it shows the effect of no-show rate on the objective function value. As we can see, the blue line shows that the objective function value increase approximately linearly with the increase of no-show rate, when no double-booking is considered. The slop is approximately 2500, i.e., the objective value increases 250 with 0.1 increment of no-show rate. In addition, the red line shows that objective function value also increases with the increasing of no-show rate, when double-booking is considered. For no-show rate equal or less than 0.2, the red line and blue line coincide with each other. After that the red line increases with a much lower rate as compare to the blue line. Note that the gap between the blue line and the red line is the reduced cost brought by double-booking. Clearly, double-booking cannot reduce the objective value when no-show rate is low (less than 0.2 in our case).

In Figure 3.4, it shows the effect of no-show rate on the optimal double-booking rate. Note that the double-booking rate is measured as the number of doubled appointment slots divided by the total number of appointment slots. As we can see, the optimal double-booking rate does not change (constant zero) when the no-show rate is less or equal to 0.2. However, with the further increase of patient no-show rate, the optimal double-booking rate is expected to increase approximately linearly until the maximum double-booking rate (100%) is reached.

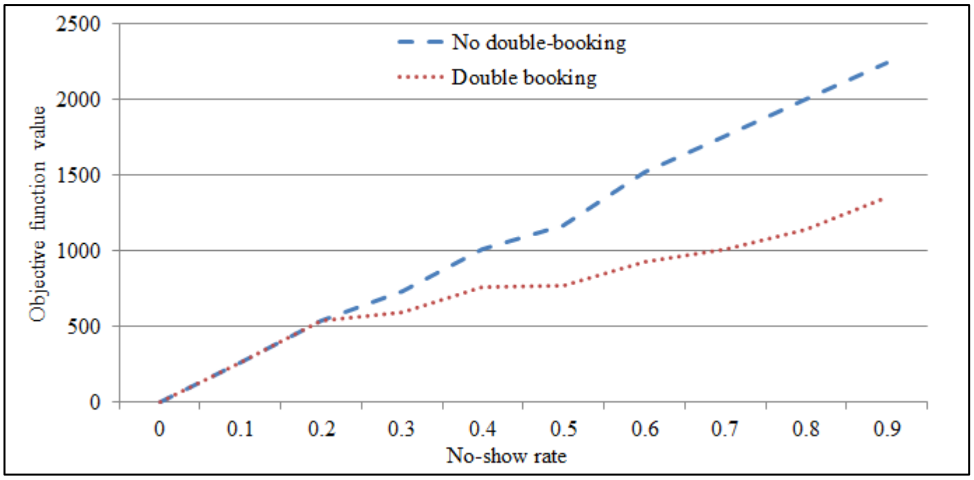

Fig. 3.3: Effect of no-show rate on the objective function value

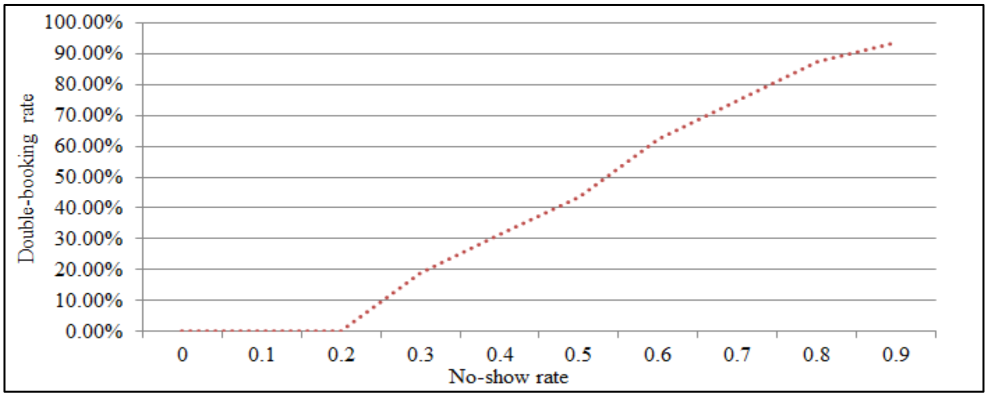

Fig. 3.4: Effect of no-show rate on the double-booking rate

3.5.4. Effect of non-homogenous rate of patient no-show rate and appointment cancellation rate

Note that all the above discussions are based on the homogenous patient no-show rate and appointment cancellation rate. In this section, we investigate the optimal overbooking strategy under the non-homogenous rate of patient no-show and appointment cancellation. To be more specific, we intend to study whether an appointment slot should be double-booked if the first patient booked in this slot has particularly high no-show rate or appointment cancellation rate. For this purpose, we consider 8 different cases, namely, Cases A1 – A8, all with punctual patient arrival (if not no-show or appointment cancellation), and constant service time 30 minutes. For Case A1, we assume all patients, but the patients booked in slot 1 (low attendance rate: no-show rate is 42%, and cancellation rate is 13%), have medium attendance rate (same as base case), i.e. no-show rate is 17%, and cancellation is 13%. Similarly, for Case A2, we assume all patients, but the patients booked in slot 2 (low attendance rate: no-show rate is 42%, and cancellation rate is 13%), have high attendance rate, i.e. no-show rate is 3%, and cancellation is 2%. Note that similar assumption will be made for Cases A3 – A8.

Table 3.9: Optimal double-booking strategies

Slot Index	Case A1		Case A2		Case A3		Case A4		Case A5		Case A6		Case A7		Case A8	
	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
1	**D**	**D**	S	S	S	S	S	S	S	S	S	S	S	S	S	S
2	S	S	**D**	**D**	S	S	S	S	S	S	S	S	S	S	S	S
3	S	S	S	S	**D**	**D**	S	S	S	S	S	S	S	S	S	S
4	S	S	S	S	S	S	**D**	**D**	S	S	S	S	S	S	S	S
5	S	S	S	S	S	S	S	S	**D**	**D**	S	S	S	S	S	S
6	S	S	S	S	S	S	S	S	S	S	**D**	**D**	S	S	S	S
7	S	S	S	S	S	S	S	S	S	S	S	S	**D**	**D**	S	S
8	S	S	S	S	S	S	S	S	S	S	S	S	S	S	**D**	**D**

“S” means single booking, “D” means double-booking.

In Table 3.9, the optimal double-booking strategies for Cases A1 – A8 are shown. Clearly, the results show that clinic should double-book the slots if the patients scheduled in the slots are expected to have high non-attendance rates. For example, in Case A2, the patients scheduled in slots 2 are expected with high non-attendance rate. As a result, the optimal double-booking strategy suggest slot 2 (neither slot 1 nor slot 3) to be double-booked. This finding supports the common double-booking practice adopted by the clinics, i.e., book another patient in the same slot if the current patient in the slot has high non-attendance rate.

3.6. Model extension

3.6.1. Problem description

As we can see, the above mentioned patient overbooking model cannot handle the situation, which providers cooperate with each other in their work. In many clinics, they assign patient with two providers, one primary provider and one secondary provider. In case that the primary provider is too busy to see a patient, the secondary provider can take this patient over if the patient is willing to accept it. In order to enable our overbooking model with this kind of provider cooperation, we relax the constraint "providers can only see their own patients" in the above discussed model by assuming that one provider can take the patient who is overbooked in a slot from other providers. Note that for any single slot of a provider, at most one overbooked patient in the slot can be taken by other providers, the provider still holds full responsibility the other remaining patient in this slot. In addition, we also consider patient's preference by introducing the "preference" indicator, in order to prevent patient being seen by some providers that this patient doesn't like. In the following, we present the formulation of this modified patient overbooking model.

3.6.2. Terminology

Index:

i : Index of patients with appointment.

j : Index of providers.

ω : Scenario index.

Patients with appointment:

$p_{ij}^{1} = \{t_{ij}^{A1}, t_{ij}^{S1}, t_{ij}^{E1}, t_{ij}^{L1}, t_{ij}^{W1}, I_{ij}^{C1}, I_{ij}^{N1}\}$: Denote the first patient scheduled in the *i*th appointment slot of provider *j*.

$p_{ij}^{2} = \{t_{ij}^{A2}, t_{ij}^{S2}, t_{ij}^{E2}, t_{ij}^{L2}, t_{ij}^{W2}\ I_{ij}^{C2}, I_{ij}^{N2}\}$: Denote the second patient scheduled in the *i*th appointment slot of provider *j*.

Parameters:

I_{ij}^{C1}, I_{ij}^{C2} : The cancellation indicator of the corresponding patient, with “1” indicating cancellation.

I_{ij}^{N1}, I_{ij}^{N2} : The no-show indicator of the corresponding patient, with “1” indicating no-show.

T_{ij}^{S} : The expected starting time of the *i*th appointment of provider *j*.

T_{j}^{E} : The expected service ending time of provider *j*.

$dummy\_t_{ij}^{A1}$, $dummy\_t_{ij}^{A2}$: Dummy arrival time of the corresponding patient, which is generated from given arrival time distribution.

$dummy\_t_{ij}^{L1}$, $dummy\_t_{ij}^{L2}$: dummy service length of the corresponding patient, which is generated from given service length distribution.

c^{Wait} : The cost coefficient related to patient waiting time.

c^{Idle} : The cost coefficient related to provider idle time.

$c^{Overtime}$: The cost coefficient related to provide over time.

Variables:

t_{ij}^{A1}, t_{ij}^{A2} : The arrival time of the corresponding patient.

t_{ij}^{S1}, t_{ij}^{S2} : The service starting time of the corresponding patient.

t_{ij}^{E1}, t_{ij}^{E2} : The service ending time of the corresponding patient.

t_{ij}^{L1}, t_{ij}^{L2} : The service length of the corresponding patient.

t_{ij}^{W1}, t_{ij}^{W2} : The waiting time of the corresponding patient.

I_{ij}^{A} : Appointment indicator; $I_{ij}^{A} = 1$, if at least 1 patient is scheduled in the *i*th appointment slot of provider *j* .

I_{ij}^{D} : Double-booking indicator; $I_{ij}^{D} = 1$, if the *i*th appointment slot of provider *j* is double-booked.

t_{j}^{I} : The idle time of provider *j.*

t_{j}^{O} : The overtime of provider *j.*

$I_{ijj'}^{assign}$: Assignment index for the second patient scheduled in the *i*th appointment slot of provider *j.* $I_{ijj'}^{assign} = 1$, if this patient is seen by the provider *j'.*

$I_{ijj'}^{prefernce}$: Preference index for the second patient scheduled in the *i*th appointment slot of provider *j.* $I_{ijj'}^{prefernce} = 0$, if this patient prefer not to be seen by provider *j'.*

3.6.3. Formulation

Min

$$\mathrm{E}_{\omega}\left(c^{Wait} \cdot \sum_{j} \sum_{i} \left(t_{ij}^{W1}(\omega) + t_{ij}^{W2}(\omega) \right) + \sum_{j} \left(c^{Overtime} \cdot t_{j}^{O}(\omega) + c^{Idle} \cdot t_{j}^{I}(\omega) \right) \right) \tag{3.27}$$

S.T.

$$t_{ij}^{L1}(\omega) = dummy\_t_{ij}^{L1}(\omega) \cdot I_{ij}^{A} \cdot \left(1 - I_{ij}^{C1}(\omega)\right) \cdot \left(1 - I_{ij}^{N1}(\omega)\right), \forall i, j. \tag{3.28}$$

$$t_{ij}^{L2}(\omega) = dummy\_t_{ij}^{L2}(\omega) \cdot I_{ij}^{D} \cdot \left(1 - I_{ij}^{C2}(\omega)\right) \cdot \left(1 - I_{ij}^{N2}(\omega)\right), \forall i, j. \tag{3.39}$$

$$t_{ij}^{A1}(\omega) = dummy\_t_{ij}^{A1}(\omega) \cdot I_{ij}^{A} \cdot \left(1 - I_{ij}^{C1}(\omega)\right) \cdot \left(1 - I_{ij}^{N1}(\omega)\right), \forall i, j. \tag{3.40}$$

$$t_{ij}^{A2}(\omega)=dummy\_t_{ij}^{A2}(\omega)\cdot I_{ij}^{D}\cdot\left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right),\forall i,j. \qquad (3.41)$$

$$I_{ij}^{A}\geq I_{ij}^{D},\forall i,j. \qquad (3.42)$$

$$t_{ij}^{E1}(\omega)=t_{ij}^{S1}(\omega)+t_{ij}^{L1}(\omega),\forall i,j. \qquad (3.43)$$

$$t_{ij}^{E2}(\omega)=t_{ij}^{S2}(\omega)+t_{ij}^{L2}(\omega),\forall i,j. \qquad (3.44)$$

$$t_{ij}^{S1}(\omega)\geq t_{(i-1)j}^{E1}(\omega),\forall i,j,i>1. \qquad (3.45)$$

$$t_{ij}^{S2}(\omega)+M\cdot\left(1-I_{ijj'}^{assign}(\omega)\right)\geq t_{ij'}^{E1}(\omega),\forall i,j,j' \qquad (3.46)$$

$$t_{ij'}^{S1}(\omega)+M\cdot\left(1-I_{(i-1)jj'}^{assign}\right)\geq t_{(i-1)j}^{E2}(\omega),\forall i,j,j',i>1 \qquad (3.47)$$

$$t_{ij}^{W1}(\omega)+M\cdot I_{ij}^{A}\cdot\left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right)\geq t_{ij}^{S1}(\omega)-T_{ij}^{S},\forall i,j. \qquad (3.48)$$

$$t_{ij}^{W1}(\omega)\geq 0,\forall i,j. \qquad (3.49)$$

$$t_{ij}^{W2}(\omega)+I_{ij}^{D}\cdot\left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right)\geq t_{ij}^{S2}(\omega)-T_{ij}^{S},\forall i,j. \qquad (3.50)$$

$$t_{ij}^{W2}(\omega)\geq 0,\forall i,j. \qquad (3.51)$$

$$t_{j}^{O}(\omega)+M\cdot I_{ij}^{A}\cdot\left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right)\geq t_{ij}^{E1}(\omega)-T_{j}^{E},\forall i,j. \qquad (3.52)$$

$$t_{j}^{O}(\omega)+M\cdot\left(1-I_{ij'j}^{assign}(\omega)\right)\geq t_{ij'}^{E2}(\omega)-T_{j}^{E},\forall i,j,j'. \qquad (3.53)$$

$$t_{j}^{O}(\omega)\geq 0,\forall j. \qquad (3.54)$$

$$t_{j}^{I}(\omega)\geq T_{j}^{E}-T_{1j}^{S}+t_{j}^{O}(\omega)-\sum_{i}\left(t_{ij}^{L1}(\omega)+\sum_{j'}dummy\_t_{ij}^{A2}(\omega)*I_{ij'j}^{assign}(\omega)\right),\forall j. \qquad (3.55)$$

$$t_{j}^{I}(\omega)\geq 0,\forall j. \qquad (3.56)$$

$$t_{ij}^{S1}(\omega)\geq t_{ij}^{A1}(\omega),\forall i,j. \qquad (3.57)$$

$$t_{ij}^{S2}(\omega)\geq t_{ij}^{A2}(\omega),\forall i,j. \qquad (3.58)$$

$$t_{ij}^{S1}(\omega) \geq T_{1j}^{S}, \forall i, j. \tag{3.59}$$

$$t_{ij}^{S2}(\omega) \geq T_{1j}^{S}, \forall i, j. \tag{3.60}$$

$$\sum_{j} I_{ij'j}^{assign}(\omega) = I_{ij'}^{D} \cdot \left(1 - I_{ij'}^{C2}(\omega)\right) \cdot \left(1 - I_{ij'}^{N2}(\omega)\right), \forall i, j'. \tag{3.61}$$

$$\sum_{j'} I_{ij'j}^{assign}(\omega) \leq 1, \forall i, j. \tag{3.62}$$

$$I_{ijj'}^{assign}(\omega) \leq I_{ijj'}^{prefernce}(\omega), \forall i, j, j'. \tag{3.63}$$

$$\sum_{j'} I_{ij'j}^{assign}(\omega) \leq 1, \forall i, j. \tag{3.64}$$

$$\sum_{j} I_{ij}^{D} \leq 1, \forall i. \tag{3.65}$$

$$I_{ij}^{D} + I_{(i-1)j}^{D} \leq 1, \forall i, j, i > 1 \tag{3.66}$$

$$I_{1j}^{D} = 0, \forall j. \tag{3.67}$$

$$I_{8j}^{D} = 0, \forall j. \tag{3.68}$$

Beside the objective function, there are also 31 constraints, as defined by Eqs. 3.28 – 3.68. To be more specific, Eqs. 3.28 – 3.29 are the service length constraint, which make sure the service length equal to zero if the corresponding patient is no-show or cancels the appointment, or there is no such patient i.e. $I_{ij}^{A} = 0$ or $I_{ij}^{D} = 0$; otherwise, the service length will be random drawn from a given distribution. Similarly, Eqs. 3.30 – 3.31 define the patient arrival time. If the corresponding patient is no-show or cancels the appointment, or there is no such patient, then the patient arrival time will be equal to zero; otherwise the patient arrival time will be drawn from a given distribution. Eq. 3.32 describes the double-booking constraint. It defines that if an appointment slot is not single booked, then it cannot be double-booked. Eqs. 3.33 – 3.34 define the relationship among the service start time, service ending time and service length. Also, Eqs. 3.35 – 3.37 are the service commitment constraints. These constraints prevent a provider serving two or more patients at the same time based on the assumption that patients scheduled in early appointment slot should receive the service early, if they are not no-show or cancel the appointment. Eqs. 3.38 – 3.41 are the patient waiting time constraints. The patient waiting time is defined as the real service start time minus the expected service start time or zero, whichever is greater. Furthermore, Eqs. 3.42 – 3.44 define the provider overtime, which should equal to the ending time of the last patient less the expected ending time of the clinic session, or zero, whichever is greater. Eqs. 3.45 – 3.46 define the provider idle time, which equals the length of clinic length plus the provider overtime and minus the sum of service time for each patient seen in the clinic session. Eqs. 3.47 – 3.50 are the service start time constraint, which define the service start time of a patient should be earlier than the arrival time of the corresponding patient, as well as the expected service start time of the first appointment. Eqs. 3.51 – 3.54 are the assignment constraints for the overbooked patients. At last, Eqs. 3.55 – 3.58 define a few heuristic overbooking rules.

Similar to our first patient overbooking model, the decision variables in this model can be divided two stages. In the first stage, the decision variables will be the appointment indicator and the double-booking indicator, i.e. I_{ij}^{A} and I_{ij}^{D}, which decide the number of patients that have been scheduled in each appointment slot for each provider. These patients with appointment are subject to random arrival process and service process. The random arrival process is controlled by the no-show rate, cancellation rate, and arrival time distribution. Similarly, the service times of the patients are controlled by the service time distribution. Hence, the decision variables in the second stage are the service start times and ending times for the patients to minimize the weight sum of patient waiting time, provider idle time and provider overtime, for each scenario ω. Note that for different scenarios, there will be different optimal second stage decisions, i.e. the service starting times and ending times of patients, and thus a different second-stage objective value, which is defined by $c^{Wait} \cdot \sum_j \sum_i \left(t_{ij}^{W1}(\omega) + t_{ij}^{W2}(\omega) \right) + \sum_j \left(c^{Overtime} \cdot t_j^O(\omega) + c^{Idle} \cdot t_j^I(\omega) \right)$. As we can see in Eq. 3.27, the first-stage objective value is just the expectation of the second-stage objective value.

3.6.4. Numerical analysis

In the similar manner, the model is code in GAMS and can be solved with CPLEX. In order to demonstrate performance of the model, we applied the model to the same cases (case0 – case 8) which were used to study the overbooking booking model discussed earlier. The solver is run on a personal computer with an Intel 2.67GHz i5 dual-core processor and 2.9GB RAM. It takes around 30 minutes to find the optimal solution for each of the 9 cases (base case plus the other eight cases). The optimal overbooking strategy found for the nine cases are shown in Table 3.10, and the descriptive performance statistics of each overbooking strategy are also presented.

Table 3.10: Optimal overbooking strategy for Case 0 – 8

Slot Index	Case 0		Case 1		Case2		Case3		Case4		Case5		Case6		Case7		Case8	
	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
1	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
2	**D**	S	**D**	S	S	**D**	S	**D**	S	**D**	**D**	S	S	**D**	**D**	S	**D**	S
3	S	**D**	S	S	**D**	S	**D**	S	**D**	S	S	**D**	**D**	S	S	**D**	S	**D**
4	**D**	S	S	S	S	**D**	S	**D**	S	**D**	**D**	S	S	**D**	**D**	S	S	S
5	S	S	S	S	**D**	S	**D**	S	**D**	S	S	**D**	**D**	S	S	**D**	**D**	S
6	S	**D**	S	S	S	**D**	S	**D**	S	**D**	**D**	S	S	S	**D**	S	S	**D**
7	S	S	S	S	**D**	S	S	S	S	S	S	**D**	S	**D**	S	S	S	S
8	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
# of overbooking	4		1		6		5		5		5		5		5		4	

Clearly, the optimal overbooking strategies are different among the 9 cases. For instance, Cases 0, 1 and 2, have medium, high, and low patient attendance rate, respectively. The corresponding optimal overbooking strategies show that for high patient attendance rate (95%), one overbooking is needed; for medium patient attendance rate (70%), one out of four appointment slots (25%) should be double-booked; for low patient attendance rate, three out of eight appointment slots (37.5%) should be double-booked. It can be seen that with the decreasing patient attendance rate, the optimal number of double-booked appointment should increase. Note that our heuristic rules define that: (1) the first and last appointment slot cannot be overbooked; (2) providers don't overbook appointment slots at the same time (e.g. if provider A overbooked his/her appointment slot #3, then provider B cannot overbook his/her appointment slot #3); (3) There should be no consecutive overbookings (e.g. if provider A overbooked his/her appointment slot #2, then he/her cannot over booked appointment slot #2). With the consideration of heuristic rules, Case 2 has actually achieved the maximum possible overbooking level. Hence, the optimal overbooking number could increase further if the heuristic rules are removed. Similarly, Case 0, 3 and 4, present medium, high, and low patient service time variance. By compare the number of double-booked appointment slots among the three cases, no significant impact of service time variance on the optimal number of double-booked appointment slots is found. In other word, the service time variance doesn't influence the optimal number of double-booked appointment slots. In addition, Cases 0, 5 and 6 represent medium, high and low patient arrival time variability. Similarly, no significant impact of patient arrival pattern can be found on the optimal number of double-booked appointment slots. At last, Cases 0, 7 and 8 represent the situation of medium, low, and high income patients. Similarly, no significant impact of patient arrival pattern can be found on the optimal number of double-booked appointment slots. As compare to the previous overbooking model (no cooperation), in which the optimal overbooking number have been significantly dropped (in case 8) due to the cost of waiting, the optimal overbooking number for case 8 haven't decreased, since the cooperation mechanism can help to reduce patient waiting time.

In Table 3.11 the mean, the standard deviations, and "the percent change" of performance metrics including patient waiting time, provider idle time, provider overtime, and objective function value, are shown for each case. The results reveal a few interesting phenomena which are commonly seen in practice. For instance, Case 1 and Case 0 have the same clinic settings except for the attendance rate, where it is higher for Case 1. The statistics indicate that Case 1 has a lower objective value compared with Case 0, which supports the general concept that high attendance rates are preferred in clinics. This concept can also be revealed by comparing Case 2 with Case 0, where the attendance rate is higher for Case 0. For another instance, Case 3 and Case 4 have the same clinic setting except for the service time distribution, where the variance is higher for Case 3. The statistics indicate that Case 3 has a higher objective value, as well as the patient waiting time, provider idle time, and provider overtime. This supports the general concept that clinics want to standardize their procedure and reduce the service time variability. In addition, Case 5 and Case 6 have the same clinic setting except for the patient arrival pattern, where the patients tend to arrive earlier for Case 5. The statistics indicate that Case 5 has a lower objective value, which supports the general concept that clinics want patient to arrive early for their appointments. To add more, Case 7 and Case 8 also have the same clinic settings except for the cost coefficient ratio, where Case 7 represents the scenario of low-income patients by using a high ratio. The statistics indicate that Case 7 have a higher objective value compared with Case 8. The implication is that high-income patients are preferred by the clinics. The "percentage change" measures the change of these performance measures by comparing with the previous model (no cooperation). As we can see, for all cases except for case 2, a lower (better) objective value is achieved. This indicates that enable the cooperation among providers can decrease the overall cost of the clinic session in terms of patient waiting time, provider idle time and provider over time. For Case 2, the new model end up with a higher objective value, since the heuristic rule prevented clinic overbooking more patient in a clinic session. In Case 2 (low attendance rate) the most significant decrease in patient waiting time are achieved. This indicates that the cooperation mechanism can most effectively reduce patient waiting time when patient non-attendance rate is high. On the contrary, in Case 1(high attendance rate) and Case 2 (high patient income) the patient waiting time has significantly increased. This means that the cooperation mechanism doesn’t help to reduce patient waiting time when the non-attendance rate is low or patient have high income. In addition, for all cases (except for case 2) the new model lead to a reduced provider idle time. This further proved that the cooperation mechanism can reduce provider idle time and balance the workload among providers. Similarly, the special situation of Case 2 is caused by the heuristic rule. As for the overtime, the cooperation mechanism has various effect among cases. In Case 2, 6 and 7, the new model leads to a reduced overtime, while in other cases, the new model leads to an increased overtime.

Table 3.11: Summary of descriptive performance statistics for Cases 0 - 8

		Case0	Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case8
Objective	mean	898.2	773.6	1225.0	1147.7	621.6	632.3	950.7	1514.3	467.9
	std of mean	9.95	7.65	7.90	11.94	5.40	7.05	8.47	11.51	5.54
	percent change	-8.5%	-1.0%	4.8%	-0.8%	-8.8%	-21.6%	-13.1%	-15.8%	-2.6%
Patient waiting/min	mean	150.7	196.4	56.1	242.3	108.4	126.4	184.3	196.1	165.0
	std of mean	4.83	4.96	0.79	6.60	1.31	4.67	4.83	3.79	5.95
	percent change	-10.9%	67.4%	-72.7%	9.4%	-11.3%	21.8%	-7.1%	-22.9%	211.3%
Provider idle/min	mean	114.1	64.4	212.0	109.8	89.9	71.5	112.0	96.5	108.0
	std of mean	2.86	1.43	3.45	3.50	1.63	1.47	2.30	3.46	1.14
	percent change	-14.1%	-33.4%	32.4%	-17.3%	-10.5%	-35.0%	-16.8%	-16.0%	-40.3%
Provider overtime/min	mean	19.8	31.0	8.5	42.8	5.8	17.2	23.6	23.6	24.6
	std of mean	0.6	1.7	0.4	3.0	0.3	0.6	1.9	0.6	0.8
	percent change	26.9%	50.1%	-49.4%	36.5%	20.8%	3.0%	-5.6%	-10.3%	60.8%

4. SINGLE-PROVIDER WALK-IN PATIENT ADMISSION OPTIMIZATION MODEL

4.1. Problem definition

This study addresses the admission problem of walk-in patients in clinics with high no-show rates, high late-cancellation rates and many walk-in patients. In such clinics, overbooking and admitting walk-in patients are usually adopted to reduce the negative impact of patient no-shows and late cancellations and improve the operations efficiency and the accessibility of the clinics. Thus, the objectives of the walk-in patient admission are to optimize the efficiency and the accessibility by admitting a certain number of walk-in patients and assigning them to proper appointment slots. In this study, the operations efficiency of a clinic is measured by provider idle time and overtime, and the accessibility to the clinics is indicated by patient waiting time.

Patients are classified into two categories in this study: elective patients versus walk-in patients. The elective patients are defined as the group of patients who arrive before/by their appointment times in the current clinic session. At the end of a clinic session, all elective patients still waiting must be seen during overtime, which is a common practice. The walk-in patients are patients who show up without appointments in the current clinic session. In addition, in this study, a patient who arrives late for his/her appointment (i.e., arrives later than the beginning of the slot scheduled for his/her appointment) is treated as a walk-in patient. The walk-in patients could be admitted or rejected to wait for services available in the current session. The rejected walk-in patients could be scheduled with appointments in later clinic sessions or referred to other outpatient facilities such as an emergency department. Similarly, the admitted walk-in patients who could not be served by the end current clinic session will be rescheduled in later clinic sessions or referred to other outpatient facilities. Fig. 4.1 illustrates the flow of two categories of patients and how patient flow is affected by the walk-in patient admission decisions.

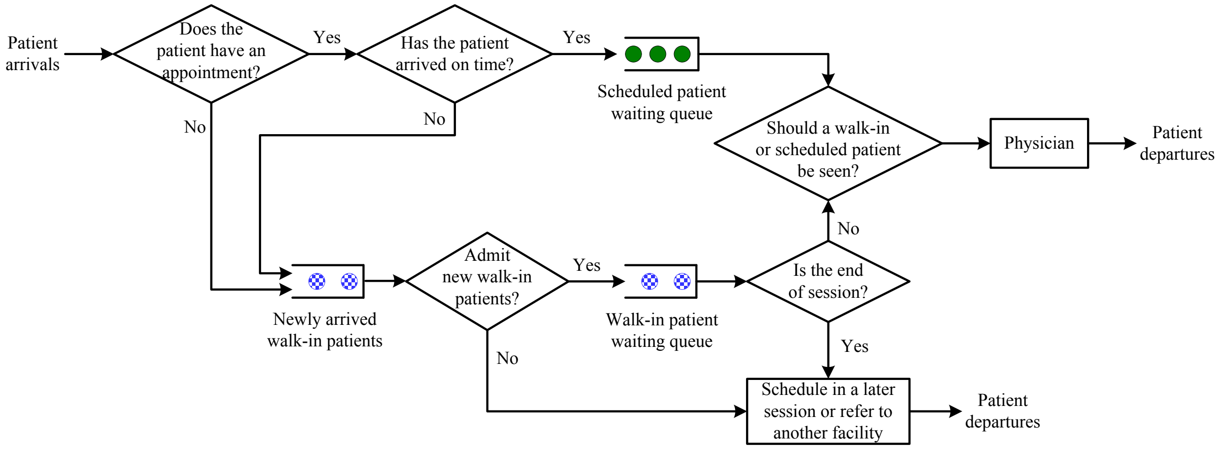
Patient arrivals
Does the patient have an appointment?
Yes
No
Has the patient arrived on time?
Yes
No
Scheduled patient waiting queue
Should a walk-in or scheduled patient be seen?
Physician
Patient departures
Newly arrived walk-in patients
Admit new walk-in patients?
Yes
No
Walk-in patient waiting queue
Is the end of session?
No
Yes
Schedule in a later session or refer to another facility
Patient departures

Fig. 4.1: Patient flow and walk-in patient admission decisions in a clinic

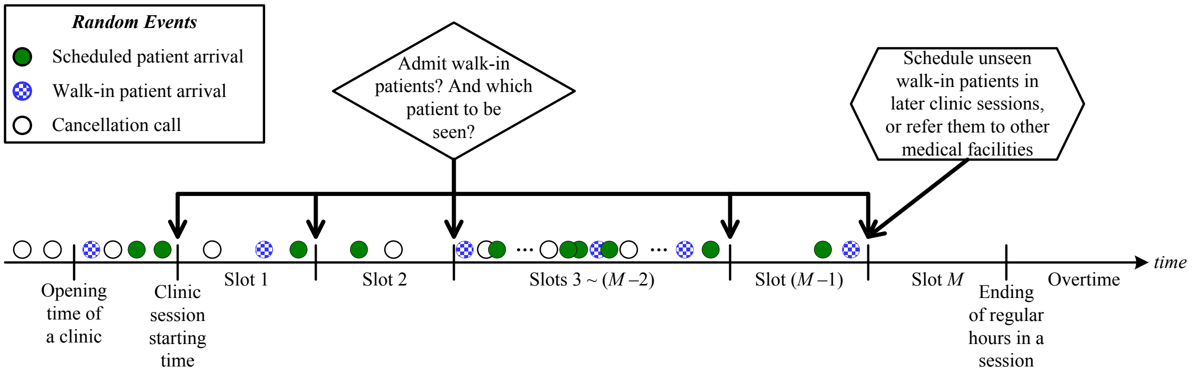
Random Events
Scheduled patient arrival
Walk-in patient arrival
Cancellation call
Admit walk-in patients? And which patient to be seen?
Schedule unseen walk-in patients in later clinic sessions, or refer them to other medical facilities
time
Opening time of a clinic
Clinic session starting time
Slot 1
Slot 2
Slots 3 ~ (M –2)
Slot (M –1)
Slot M
Ending of regular hours in a session
Overtime

Fig. 4.2: Patient arrival process and walk-in patient admission process in a clinic session

As a complement to patient flow, shown in Fig. 4.1, Fig. 4.2 illustrates the walk-in patient admission process and the patient arrival process in a clinic session. As we can see, it is assumed that the walk-in patient admission decisions are made at discrete time points, which are the beginnings of M equal-length appointment slots in a clinic session. Here M denotes the total number of appointment slots in a clinic session, which is typically 4 hours. The walk-in patient admission decisions include whether or not to admit the walk-in patients who arrive during the previous slot to wait for service, and whether a walk-in patient should be seen in the next slot. In addition, at the beginning of the Mth slot, rescheduling and referrals need to be done for the currently admitted walk-in patients, who cannot be serviced by the end of current clinic session. It should be noted that patient arrivals and short-notice appointment cancellations can happens at any time between clinic opening and the beginning of the Mth slot. In addition, it is assumed that the short-notice appointment cancellations can also happen shortly before the clinic opening.

Due to high no-show rates and high late-cancellation rates, overbooking is also considered in this study. For the overbooking policy, the most popular practice is to schedule at most two appointments in one slot. Therefore, a double-booking policy is assumed in this study. In addition, it is assumed that the schedule of patient appointments in the current clinic session is known at the beginning of the session.

We assume that patients independently arrive for or cancel their appointments, that the arrivals of patients with appointments follow a given probability distribution, and that the probability of canceling an appointment is known. A Poisson process with a constant rate λ_w is used to approximate the arrival process of patients without appointments (Fetter and Thompson 1966, Ashton et al. 2005, Kopach et al. 2007). Meanwhile, we assume that the service time per patient is constant because most physicians, especially primary care physicians, have fairly consistent control of the service time for each patient, the variation of service times among patients is limited (LaGanga and Lawrence 2007, Gupta and Denton 2008). Thus, the length of an appointment slot is set to equal the average of service time per patient.

In this study, it is assumed that a provider (e.g., a physician or a nurse practitioner) only sees his/her own patients and new patients. This is common for most outpatient clinics, especially in primary care clinics, to ensure continuity of care, which is an important indicator of the quality of care (Cayirli and Veral 2003). Medical research suggests that continuity of care can improve both the clinical and process quality of patient care, especially early in the patient-provider relationship and for patients with worse health. This is because scheduling patients with the same healthcare provider improves patients' responses to recommended treatments and follow-up care (Saultz and Lochner 2005, Pandhi and Saultz 2006, Rodriguez et al. 2007). Under this assumption, the walk-in patient admission to one provider is independent of the walk-in patient admission to other providers. Therefore, this study determines the walk-in patient admission policy in the context of a single-provider system. Some clinics group several (two to four) providers as a provider team in an effort to balance continuity of care with scheduling flexibility, and this will be considered in our future study.

4.2. MDP model

Under the assumption that the walk-in patient admission decisions are made at the beginning of each of *M* appointment slots, a finite-horizon MDP model is developed in this study to capture the decision process of admitting walk-in patients in a clinic session with equal-length appointment slots. The decision horizon of the MDP model consists of *M* decision stages. Let *i* denote the index of appointment slots, and *m* and *m'* the index of decision stages. In the MDP model, decision stage *m* is the beginning of slot *i*, where *i* = *m*. For readers' convenience, Table 4.1 summarizes the definitions of the indices and parameters used in the MDP model.

4.2.1. States and actions

The state of the MDP model at decision stage *m* is denoted by

$$\begin{aligned}\mathbf{s}^m &= (\mathbf{x}^m, \mathbf{y}^m, \mathbf{z}^m, n^m, w^m) \\ &= (x_1^m, x_2^m, \ldots, x_M^m; y_1^m, y_2^m, \ldots, y_M^m; z_1^m, z_2^m, \ldots, z_M^m; n^m, w^m)\end{aligned}, \text{ for } m = 1, 2, \ldots, M, \tag{4.1}$$

where x_i^m denotes the number of elective patients with appointments in slot *i* who wait to be seen at decision stage *m*, n^m denotes the number of walk-in patients who wait to be seen at decision stage *m*, w^m denotes the number of walk-in patients who arrive between decision stages

(m–1) and m, y_i^m indicates whether a patient with an appointment in slot i could arrive between decision stages m and (m+1), and z_i^m indicates whether more than one patient with an appointment in slot i may arrive between decision stages m and (m+1). Define the initial state of the MDP model as

$$\mathbf{s}^0 = (\mathbf{x}^0, \mathbf{y}^0, \mathbf{z}^0, n^0, w^0) \ = (0,0,\ldots,0; y_1^0, y_2^0, \ldots, y_M^0; z_1^0, z_2^0, \ldots, z_M^0; 0,0)\,, \tag{4.2}$$

$$y_i^0 = \begin{cases} 1, & \text{if } b_i \geq 1 \\ 0, & \text{otherwise} \end{cases}, \tag{4.3}$$

$$z_i^0 = \begin{cases} 1, & \text{if } b_i \geq 2 \\ 0, & \text{otherwise} \end{cases}, \tag{4.4}$$

where b_i is the number of appointments scheduled in slot i of the clinic session. Since the double-booking policy specifies that at most two appointments could be scheduled in one slot, the state space, denoted by S, is

$$S = \{\, (\mathbf{x},\mathbf{y},\mathbf{z},n,w) \mid x_i = 0,\ 1 \text{ or } 2,\ y_i = 0 \text{ or } 1,\ z_i = 0 \text{ or } 1,\ z_i \leq y_i, \text{ for } i=1,\ldots,M;\ n \in \mathbf{Z}^+;\ w \in \mathbf{Z}^+\}, \tag{4.5}$$

where $\mathbf{Z}^+$ denotes the set of nonnegative integer numbers.

At each decision stage, two decisions are made based on the current state $\mathbf{s}^m$. One decision is whether to admit the walk-in patients who arrive between decision stages (m–1) and m to wait in the clinic for services, and the other decision is whether a elective or walk-in patient should be seen between decision stages m and (m+1). Thus, an action at decision stage m could be represented by

$$\mathbf{a}^m = (d_a^m, d_s^m, d_w^m), \text{ for } m = 1, 2, \ldots, M, \tag{4.6}$$

where d_a^m, d_s^m, and d_w^m are the binary decision variables of action $\mathbf{a}^m$, which are defined respectively as

$$d_a^m = \begin{cases} 1, & \text{if walk-in patientsarriving between decision stages } (m-1) \text{ and } m \\ & \quad \text{are admitted to wait in the clinic for services} \\ 0, & \text{otherwise} \end{cases}, \tag{4.7}$$

$$d_s^m = \begin{cases} 1, & \text{if a scheduled patient will be seen between decision stages } m \text{ and } m+1 \\ 0, & \text{otherwise} \end{cases}, \tag{4.8}$$

$$d_w^m = \begin{cases} 1, & \text{if a walk-in patient will be seen between decision stages } m \text{ and } m+1 \\ 0, & \text{otherwise} \end{cases}. \tag{4.9}$$

Since the length of an appointment slot equals the constant service time per patient, a provider could see at most one patient during one slot. Thus any valid action must satisfy $d_s^m + d_w^m \leq 1$. Therefore, the action set of the MDP model, denoted by A, consists of only six actions $\{(0,0,0), (0,1,0), (0,0,1), (1,0,0), (1,1,0), (1,0,1)\}$. Denote $A(\mathbf{s}^m)$ the set of all valid actions in state $\mathbf{s}^m \in S$. The notation for the states, actions, random variables and rewards in the MDP model is summarized in Table 4.2.

Table 4.1: Indices and parameters in the MDP model

	Indices
i	Index of appointment slots
m and m'	Index of decision stages
	Parameters
b_i	Number of appointments scheduled in slot i of the clinic session
c_i	Provider idle cost per slot
c_o	Provider overtime cost per slot
c_s	Waiting cost per elective patient per slot after their appointment times
c_w	Waiting cost per walk-in patient per slot
M	Total number of appointment slots and total number of decision stages
$p_{i\|m}^{m'}$	Conditional probability that a patient with an appointment in slot i arrives between decision stages m' and $(m'+1)$ given that (s)he has not arrived or cancelled his/her appointment at decision stage m
p_c	Probability that an appointment is cancelled between two adjacent decision stages
r_s	Gain from seeing one elective patient
r_w	Gain from seeing one walk-in patient
T	Length of an appointment slot
λ_w	Arrival rate of patients without appointments

Table 4.2: States, actions, random variables and rewards in the MDP model

	States
$\mathbf{s}^m$	State of the MDP model at decision stage m, for $m = 1, \ldots, M$, where $\mathbf{s}^m = (\mathbf{x}^m, \mathbf{y}^m, \mathbf{z}^m, n^m, w^m)$
$\mathbf{s}^0$	Initial state of the MDP model, where $\mathbf{s}^0 = (\mathbf{x}^0, \mathbf{y}^0, \mathbf{z}^0, n^0, w^0)$
S	Set of all possible $\mathbf{s}^m$
$\mathbf{x}^m$	Vector representing the numbers of elective patients waiting for service in state $\mathbf{s}^m$, where $\mathbf{x}^m = (x_1^m, x_2^m, \ldots, x_M^m)$
x_i^m	Number of elective patients with appointments in slot i who wait to be served in state $\mathbf{s}^m$
$\mathbf{y}^m$	Vector indicating whether a patient with an appointment in each slot could arrive between decision stages m and $(m+1)$, where $\mathbf{y}^m = (y_1^m, y_2^m, \ldots, y_M^m)$
y_i^m	Indicator whether a patient with an appointment in slot i could arrive between decision stages m and $(m+1)$
$\mathbf{z}^m$	Vector indicating whether more than one patient with an appointment in each slot could arrive between decision stages m and $m+1$, where $\mathbf{z}^m = (z_1^m, z_2^m, \ldots, z_M^m)$
z_i^m	Indicator whether more than one patient with an appointment in slot i may arrive between decision stages m and $(m+1)$
n^m	Number of walk-in patients who wait to be seen at decision stage m
w^m	Number of walk-in patients arriving between decision stages $(m-1)$ and m
	Actions
$\mathbf{a}^m$	Action taken at decision stage m, for $m = 1, 2, \ldots, M$, where $\mathbf{a}^m = (d_a^m, d_s^m, d_w^m)$ and d_a^m, d_s^m and d_w^m are the binary decision variables of action $\mathbf{a}^m$
$A(\mathbf{s}^m)$	Set of all valid actions in state $\mathbf{s}^m$
d_a^m	= 1, if walk-in patients who arrive between decision stages $(m-1)$ and m are admitted; = 0, otherwise
$\bar{d}_a^m$	Optimal values for the decision variable d_a^m in $\pi^*(\mathbf{s}^m)$
d_s^m	= 1, if a elective patient will be seen between decision stages m and $(m+1)$; = 0, otherwise
$\bar{d}_s^m$	Optimal value for the decision variable d_s^m in $\pi^*(\mathbf{s}^m)$
d_w^m	= 1, if a walk-in patient will be seen between decision stages m and $(m+1)$; = 0, otherwise
$\bar{d}_w^m$	Optimal value for the decision variable d_w^m in $\pi^*(\mathbf{s}^m)$
$\pi^*(\mathbf{s}^m)$	Optimal action for a state $\mathbf{s}^m$ at decision stage m
	Random variables
$\tilde{A}_i^m$	Number of patients with appointments in slot i who arrive between decision stages m and $(m+1)$, for $i = 1, \ldots, M$; $m = 0, 1, \ldots, (M-1)$
$\tilde{B}^m$	Number of patients without appointments who arrive between decision stages m and $(m+1)$, for $m = 0, 1, \ldots, (M-1)$
$\tilde{C}_i^m$	Number of appointments in slot i which are cancelled between decision stages m and $(m+1)$, for $i = 1, \ldots, M$; $m = 0, 1, \ldots, (M-1)$
$\tilde{U}_i^m$	= 1, if $x_i^m > 0$ and $x_j^m = 0 \; \forall j < i$; = 0, otherwise
	Rewards
$R^m(\mathbf{s}^m,\mathbf{a}^m)$	Immediate net reward of a valid action $\mathbf{a}^m$ in state $\mathbf{s}^m$
$V(\mathbf{s}^m,\mathbf{a}^m)$	Expected total net reward obtained in the $(M-m+1)$ remaining decision stages of an action $\mathbf{a}^m$ in state $\mathbf{s}^m$
$V^*(\mathbf{s}^m)$	Maximum of $V(\mathbf{s}^m,\mathbf{a}^m)$.

4.2.2. State transitions

The stochastic events considered in the proposed MDP model are random patient arrivals and random appointment cancellations during a clinic session. Let $\tilde{A}_i^m$ denote the number of patients with appointments in slot *i* who arrive between decision stages *m* and (*m*+1), $\tilde{B}^m$ the number of patients without appointments who arrive between decision stages *m* and (*m*+1), and $\tilde{C}_i^m$ the number of appointments in slot *i* which are cancelled between decision stages *m* and (*m*+1), where $i = 1, 2,\ldots, M$ and $m = 0, 1, 2,\ldots, (M–1)$. Due to the assumption of the double-booking policy, $\tilde{A}_i^m$ and $\tilde{C}_i^m$ could equal 0, 1 or 2. In the MDP model, decision stage 0, which corresponds to the initial state, represents the beginning of the clinic session (see Fig. 4.2). Since patients independently arrive for their appointments with given probabilities, the random variable $\tilde{A}_i^m$ follows the binomial distribution with $(y_i^m+z_i^m)$ trials and probability $p_{i|m}^m$, where $p_{i|m}^m$ is the conditional probability that a patient with an appointment in slot *i* arrives between decision stages *m* and (*m*+1) given that he/she has not arrived or cancelled his/her appointment at decision stage *m*. Similarly, since patients independently cancel their appointments with a known constant probability, the random variable $\tilde{C}_i^m$ follows the binomial distribution with $(y_i^m+z_i^m)$ trials and probability p_c, where p_c is the probability that an appointment in slot *i* is cancelled between two adjacent decision stages. Due to the approximate Poisson arrival process of patients without appointments, the random variable $\tilde{B}^m$ is Poisson distributed with rate $\lambda_w T$, where λ_w is the constant arrival rate, and *T* is the length of an appointment slot.

Given a state $\mathbf{s}^m \in S$ at decision stage *m*, the transition from the state $\mathbf{s}^m$ to a next state $\mathbf{s}^{m+1}$ depends on the action taken at decision stage *m* and the stochastic events occurring between decision stages *m* and (*m*+1). Once an action $\mathbf{a}^m = (d_a^m, d_s^m, d_w^m)$ is taken in the state $\mathbf{s}^m$, the state $\mathbf{s}^{m+1}$ to which the process transitions satisfies

$$x_i^{m+1} = x_i^m - \widetilde{U}_i^m d_s^m \text{, for } i = 1, 2, \ldots, m, \tag{4.10}$$

$$x_i^{m+1} = x_i^m + \widetilde{A}_i^m - \widetilde{U}_i^m d_s^m \text{, for } i = m+1, m+2, \ldots, M, \tag{4.11}$$

$$y_i^{m+1} = \begin{cases} 0, & \text{if } z_i^m = 0 \text{ and } \widetilde{A}_i^m + \widetilde{C}_i^m = 1 \\ 0, & \text{if } z_i^m = 1 \text{ and } \widetilde{A}_i^m + \widetilde{C}_i^m = 2, \quad \text{for } i = 1, 2, \ldots, M, \\ y_i^m, & \text{otherwise} \end{cases} \tag{4.12}$$

$$z_i^{m+1} = \begin{cases} 0, & \text{if } \widetilde{A}_i^m \geq 1 \text{ or } \widetilde{C}_i^m \geq 1 \\ z_i^m, & \text{otherwise} \end{cases}, \quad \text{for } i = 1, 2, \ldots, M, \tag{4.13}$$

$$n^{m+1} = n^m + w^m d_a^m - d_w^m, \tag{4.14}$$

$$w^{m+1} = \widetilde{B}^m + \sum_{i=1}^{m} \widetilde{A}_i^m , \tag{4.15}$$

$$\widetilde{U}_i^m = \begin{cases} 1, & \text{if } x_i^m > 0 \text{ and } x_j^m = 0 \; \forall j < i \\ 0, & \text{otherwise} \end{cases}, \quad \text{for } i = 1, 2, \ldots, M. \tag{4.16}$$

4.2.3. Rewards and costs

The reward of each action is the gain from seeing an elective or walk-in patient, and the cost associated with an action in a given state includes patient waiting costs and provider idle and overtime costs. Similar to other appointment scheduling studies in the literature, we assume linear relationships between the reward and the number of patients seen (LaGanga and Lawrence 2007), and between patient waiting cost and patient waiting time, between provider idle cost and provider idle time, between provider overtime cost and provider overtime (Cayirli and Veral 2003, Muthuraman and Lawley 2008).

At each decision stage, the cost associated with an action consists of patient waiting cost and provider idle cost. Thus, the immediate net reward of a valid action $\mathbf{a}^m$ in state $\mathbf{s}^m \in S$, denoted by $R^m(\mathbf{s}^m, \mathbf{a}^m)$, is

$$R^{m}(\mathbf{s}^{m},\mathbf{a}^{m}) = r_s d_s^m + r_w d_w^m - c_s \max\left(0, \sum_{i=1}^{m} x_i^m - d_s^m\right) \\ - c_w\left(n^m + w^m d_a^m - d_w^m\right) - c_i\left(1 - d_s^m - d_w^m\right), \ \text{for m} = 1, \ldots, \text{M - 1} \quad , \tag{4.17}$$

where r_s is the gain from seeing one elective patient, r_w the gain from seeing one walk-in patient, c_s the waiting cost per elective patient per slot after their appointment times, c_w the waiting cost per walk-in patient per slot, and c_i the provider idle cost per slot. Since the extra preparation work may be needed in order to see each walk-in patient, it is assumed that $r_s \geq r_w$ in this study. Meanwhile, since walk-in patients usually tolerate longer waiting time than elective patients (Cayirli and Veral 2003), it is assumed that $c_s > c_w$. The cost of waiting prior to appointment time is ignored in the immediate net reward because waiting due to early arrival is voluntary. At decision stage M, since all walk-in patients who could not be served in the current clinic session are dismissed, the waiting cost of walk-in patients is not associated with any action. Thus, the immediate net reward of a valid action $\mathbf{a}^M$ in state $\mathbf{s}^M \in S$ is

$$R^{M}(\mathbf{s}^{M},\mathbf{a}^{M}) = r_s d_s^M + r_w d_w^M - c_s \max\left(0, \sum_{i=1}^{M} x_i^M - d_s^M\right) - c_i\left(1 - d_s^M - d_w^M\right). \tag{4.18}$$

Since all elective patients waiting for service at the end of a clinic session must be seen during overtime, the provider overtime cost and the waiting cost of elective patients should be considered at the end of the session. Then the net reward at the end of the session of a valid action $\mathbf{a}^M$ in state $\mathbf{s}^M \in S$, denoted by $R^{M+1}(\mathbf{s}^M,\mathbf{a}^M)$, is

$$R^{M+1}(\mathbf{s}^{M},\mathbf{a}^{M}) = \begin{cases} (r_s - c_o)\left(\sum_{i=1}^{M} x_i^M - d_s^M\right) - \frac{c_s}{2}\left(\sum_{i=1}^{M} x_i^M - d_s^M\right)\left(\sum_{i=1}^{M} x_i^M - d_s^M - 1\right), & \text{if } \left(\sum_{i=1}^{M} x_i^M - d_s^M\right) > 0 \\ 0, \ \text{otherwise} \end{cases} \quad , \tag{4.19}$$

where c_o is the provider overtime cost per slot. The cost coefficient of provider overtime depends on provider overtime payment and provider willingness to work overtime. Usually, providers are not willing to do so unless the overtime payment is significantly higher than the regular figure. Therefore, in this study, we assume $c_o > r_w$ to prevent the average overtime from being too long because high frequency of long overtime decreases job satisfaction of providers and other staff members.

When the process is in state $\mathbf{s}^m$ at decision stage m, the action maximizing the expected total net reward obtained in the (M–m+1) remaining stages should be taken. Let $V(\mathbf{s}^m,\mathbf{a}^m)$ denote the expected total net reward obtained in the (M–m+1) remaining stages of an action $\mathbf{a}^m$ in state $\mathbf{s}^m$, and $V^*(\mathbf{s}^m)$ the maximum of $V(\mathbf{s}^m,\mathbf{a}^m)$. Then, $V^*(\mathbf{s}^m)$ and $V(\mathbf{s}^m,\mathbf{a}^m)$ could be determined by

$$V^*(\mathbf{s}^m) = \max_{\mathbf{a}^m \in A(\mathbf{s}^m)} V(\mathbf{s}^m,\mathbf{a}^m)\text{, for } m = 1,\ldots, M, \tag{4.20}$$

$$V(\mathbf{s}^m,\mathbf{a}^m) = R^m(\mathbf{s}^m,\mathbf{a}^m) + E[V^*(\mathbf{s}^{m+1} | \mathbf{s}^m,\mathbf{a}^m)]\text{, for } m = 1,\ldots, M–1, \tag{4.21}$$

$$\text{and } V(\mathbf{s}^M,\mathbf{a}^M) = R^M(\mathbf{s}^M,\mathbf{a}^M) + R^{M+1}(\mathbf{s}^M,\mathbf{a}^M)\text{,} \tag{4.22}$$

where $E[\bullet]$ denotes the expectation of a random variable. Thus, the optimal action for a state $\mathbf{s}^m$ at decision stage m, denoted by $\pi^*(\mathbf{s}^m)$, is $\pi^*(\mathbf{s}^m) = \arg\max_{\mathbf{a}^m \in A(\mathbf{s}^m)} V(\mathbf{s}^m,\mathbf{a}^m)$. Denote $\bar{d}_a^m$, $\bar{d}_s^m$ and $\bar{d}_w^m$ the optimal values for the decision variables ${d_a}^m$, ${d_s}^m$ and ${d_w}^m$ in $\pi^*(\mathbf{s}^m)$.

4.3. Properties of the MDP model

In Section 3, we propose the MDP model to capture the walk-in patient admission process in a clinic session. In this section, we analyze the properties of the proposed MDP model. For any possible state $\mathbf{s}^m$ of the MDP model, the set of valid actions $A(\mathbf{s}^m)$ must satisfy Equations 4.23 – 4.25 in Proposition 1.

Proposition 1. Any valid action $\mathbf{a}^m$ in a state $\mathbf{s}^m \in S$ must satisfy

$${d_s}^m = 0 \text{ if } \textstyle\sum_{i=1}^{M} x_i^m = 0\text{,} \tag{4.23}$$

$${d_w}^m = 0 \text{ if } n^m + w^m {d_a}^m = 0, \tag{4.24}$$

$$\text{and } d_s^{\,m} + d_w^{\,m} \leq 1. \qquad (4.25)$$

Proof. If no elective patients are waiting to be seen at decision stage *m*, the provider could not see a elective patient between decision stages *m* and (*m*+1). This means that if $\sum_{i=1}^{M} x_i^m = 0$, $d_s^{\,m}$ must equal 0. Similarly, if neither walk-in patients are waiting to be seen at decision stage *m* nor walk-in patients arrive between decision stages (*m*–1) and *m*, the provider could not see a walk-in patient between decision stages *m* and (*m*+1). This means that if $n^m + w^m d_a^{\,m} = 0$, $d_w^{\,m}$ must equal 0. Finally, since the length of an appointment slot equals the constant service time per patient, a provider could see at most one patient between two adjacent decision stages, i.e., $d_s^{\,m} + d_w^{\,m} \leq 1$. Therefore, any valid action $\mathbf{a}^m$ in state $\mathbf{s}^m$ must satisfy conditions (4.23) – (4.25).

From Equations 4.23 – 4.25, we can infer the corollary that determines the optimal actions for states $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m = 0$ and $n^m + w^m = 0$.

Corollary 1. For a state $\mathbf{s}^t \in S$ satisfying $\sum_{i=1}^{M} x_i^m = 0$ and $n^m + w^m = 0$, the valid action set $A(\mathbf{s}^m)$ is {(0,0,0), (1,0,0)} and both actions are optimal.

Proof. According to Equations 4.23 – 4.25 in Proposition 1, $d_s^{\,m} = 0$ and $d_w^{\,m} = 0$ for a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m = 0$ and $n^m + w^m = 0$. That means that (0,0,0) and (1,0,0) are the only valid actions.

Since $n^m \geq 0$ and $w^m \geq 0$ for any state $\mathbf{s}^m \in S$, $n^m + w^m = 0$ implies $n^m = 0$ and $w^m = 0$. When $w^m = 0$, the decision for $d_a^{\,m}$ does not affect the immediate net reward according to Equation 4.17, and the next state to which the process transitions only depends on state $\mathbf{s}^m$ according to Equations 4.10 – 4.16. Therefore, taking either (0,0,0) or (1,0,0) results in the identical expected total net reward in the (*M*–*m*+1) remaining stages. Thus, both (0,0,0) and (1,0,0) are optimal.

Next, Propositions 2 and 3 reveal the other two properties of the proposed MDP model. Proposition 2 establishes the condition for an optimal action at the states satisfying $\sum_{i=1}^{M} x_i^m > 0$ or n^m>0, while Proposition 3 provides the condition for an optimal action at the states satisfying $\sum_{i=1}^{m} x_i^m > 0$ (see Appendix for the proofs of Propositions 2 and 3. According to Propositions 1 and 2, we obtain the conditions for the optimal actions in three state subsets specified in Corollaries 2, 3, and 4, respectively. Corollaries 2 and 3 could be directly derived from Proposition 2, and the proof of Corollary 4 could be found in Appendix.

Proposition 2. If $\sum_{i=1}^{M} x_i^m > 0$ or n^m>0 in a state $\mathbf{s}^m \in S$, the optimal action $\pi^*(\mathbf{s}^m)$ satisfies $\bar{d}_s^m + \bar{d}_w^m = 1$.

Corollary 2. If $\sum_{i=1}^{M} x_i^m > 0$ and $n^m + w^m = 0$ in a state $\mathbf{s}^m \in S$, the optimal action $\pi^*(\mathbf{s}^m)$ satisfies $\bar{d}_s^m = 1$ and $\bar{d}_w^m = 0$.

Corollary 3. If $\sum_{i=1}^{M} x_i^m = 0$ and n^m>0 in a state $\mathbf{s}^m \in S$, the optimal action $\pi^*(\mathbf{s}^m)$ satisfies $\bar{d}_s^m = 0$ and $\bar{d}_w^m = 1$.

Corollary 4. If $\sum_{i=1}^{M} x_i^m = 0$, $n^m = 0$ and $w^m = 1$ in a state $\mathbf{s}^m \in S$, the optimal action $\pi^*(\mathbf{s}^m)$ is (1,0,1).

Proposition 3. If $\sum_{i=1}^{m} x_i^m > 0$ in a state $\mathbf{s}^m \in S$, the optimal action $\pi^*(\mathbf{s}^m)$ satisfies $\bar{d}_s^m = 1$ and $\bar{d}_w^m = 0$.

4.4. Optimal and heuristic rules for walk-in patient admission

In the previous section, we derive the conditions for the optimal actions in several state subsets of the MDP model representing the walk-in patient admission process in a clinic session. According to these conditions, in this section, we propose optimal and heuristic walk-in-patient admission rules, which could be used to form better walk-in patient admission policies in targeted outpatient clinics.

First, we derive the optimal rules for the states $\mathbf{s}^m \in S$ satisfying $w^m = 0$. For such states, the decision for d_a^m does not affect either the immediate net reward or the next state that the process transitions to. According to Corollaries 1 – 3 and Proposition 3, we divide the states $\mathbf{s}^m \in S$ satisfying $w^m = 0$ into five subsets, and then analyze the optimal actions and admission rules for four of them, which are summarized in Table 4.3. For the remaining state subset S_{05}, the valid state set, $A(\mathbf{s}^m)$, consists of actions (0,0,1), (1,0,1), (0,1,0) and (1,1,0). Since $w^m = 0$, we know that $V(\mathbf{s}^{\mathrm{m}},(0,0,1))=V(\mathbf{s}^{\mathrm{m}},(1,0,1))$ and $V(\mathbf{s}^{\mathrm{m}},(0,1,0))= V(\mathbf{s}^{\mathrm{m}},(1,1,0))$ for any state $\mathbf{s}^m \in S_{05}$. Thus, either {(0,0,1), (1,0,1)} or {(0,1,0), (1,1,0)} should be optimal in a state $\mathbf{s}^m \in S_{05}$. The optimal actions in a state $\mathbf{s}^m \in S_{05}$ depends not only on the numbers of elective and walk-in patients waiting to be seen ($\sum_{i=1}^{M} x_i^m$ and n^m), but also on the number of new patient arrivals in time for their appointments, which is a function of the no-show rate, the late-cancellation rate and the number of appointments in the remaining slots, i.e., $\sum_{i=m+1}^{M} \left(y_i^m + z_i^m\right)$.

Table 4.3: Optimal and heuristic admission rules for the states $s^m \in \mathrm{S}$ satisfying $w^m = 0$

State subset	Conditions for the state subset	Optimal actions	Admission rules
S01	$\sum_{i=1}^{M} x_i^m = 0$ and $n^m + w^m = 0$	(0,0,0) and (1,0,0)	Rule 1: $d_s^m = 0$, $\mathrm{d_w}^{\mathrm{m}} = 0$
S02	$\sum_{i=1}^{M} x_i^m > 0$ and $n^m + w^m = 0$	(0,1,0) and (1,1,0)	Rule 2: $d_s^m = 1$, $\mathrm{d_w}^{\mathrm{m}} = 0$
S03	$\sum_{i=1}^{M} x_i^m = 0$, $n^m > 0$ and $w^m = 0$	(0,0,1) and (1,0,1)	Rule 3: $d_s^m = 0$, $\mathrm{d_w}^{\mathrm{m}} = 1$
S04	$\sum_{i=1}^{m} x_i^m > 0$, $n^m > 0$ and $w^m = 0$	(0,1,0) and (1,1,0)	Rule 2: $d_s^m = 1$, $\mathrm{d_w}^{\mathrm{m}} = 0$
S05	$\sum_{i=1}^{M} x_i^m > 0$, $\sum_{i=1}^{m} x_i^m = 0$, $n^m > 0$ and $w^m = 0$		Rule 5a, Rule 5b or Rule 5c

We propose a heuristic rule for states in S_{05} depending on the number of remaining time slots and the expected total elective patients needed to be seen by the end of the clinic session.

Let $ESP(\mathbf{s}^m)$ denote the expected total number of elective patients needed to be seen by the end of the clinic session given a state $\mathbf{s}^m \in S$. $ESP(\mathbf{s}^m)$ could be estimated by

$$ESP(\mathbf{s}^m) = \sum_{i=1}^{M} x_i^m + \sum_{i=m+1}^{M} \sum_{m'=m}^{i-1} p_{i|m}^{m'} \left(y_i^m + z_i^m \right), \tag{4.26}$$

where $p_{i|m}^{m'}$ is the conditional probability that a patient with an appointment in slot i arrives between decision stages m' and $(m'+1)$ given that he/she has not arrived or cancelled his/her appointment until decision stage m. The proposed heuristic rule for a state $\mathbf{s}^m \in S_{05}$ is

Rule 5a: For a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m > 0$ and $\sum_{i=1}^{m} x_i^m = 0$, if $ESP(\mathbf{s}^m) \le M - m + 1$, then $d_s^m = 0$, $d_w^m = 1$; otherwise, $d_s^m = 1$, $d_w^m = 0$.

In this study, we compare Rule 5a to two simple rules for a state $\mathbf{s}^m \in S_{05}$:

Rule 5b: If $\sum_{i=1}^{M} x_i^m > 0$ and $n^m > 0$, then $d_s^m = 1$, $d_w^m = 0$.

Rule 5c: If $\sum_{i=1}^{m} x_i^m = 0$ and $n^m > 0$, then $d_s^m = 0$, $d_w^m = 1$.

Rule 5b implies that elective patients are always seen before walk-in patients. On the contrary, Rule 5c states that when all elective patients waiting to be seen have appointments in the later slots, a walk-in patient should be seen in the current slot.

Next, we discuss the rules for the states $\mathbf{s}^m \in S$ satisfying $w^m > 0$. According to Equations 4.5 and 4.7, when $w^m > 0$, the decision for d_a^m affects the waiting time of walk-in patients in the immediate net reward and the number of walk-in patients waiting to be seen in later states. Admitting walk-in patients to wait for service may reduce provider idle time, increase the number of patients seen and increase patient waiting time depending on the numbers of elective and walk-in patients waiting to be seen and the number of new patient arrivals in time for their appointments. We propose two heuristic rules for decision d_a^m, which consider the number of walk-in patients arriving between decision stages $(m-1)$ and m, the expected total elective patients needed to be seen, the number of walk-in patients waiting to be seen, and the number of remaining slots. The two heuristic rules, called *Tight Rule* and *Relaxed Rule* respectively, are

Rule A: (Tight Rule) For a state $\mathbf{s}^m \in S$, if $w^m + n^m + ESP(\mathbf{s}^m) > M - m + 1$, then $d_a^m = 0$; otherwise, $d_a^m = 1$.

Rule B: (Relaxed Rule) For a state $\mathbf{s}^m \in S$, if $n^m + ESP(\mathbf{s}^m) \geq M - m + 1$, then $d_a^m = 0$; otherwise, $d_a^m = 1$.

The Tight Rule rejects the new walk-in patients when the number of new walk-in patient arrivals is greater than the expected number of idle slots, i.e., $(M - m + 1) - ESP(\mathbf{s}^m) - n^m$, while the Relaxed Rule rejects the new walk-in patients only when the expected number of idle slots is less than or equal to 0. The performance of these two heuristic rules is compared in Section 6.

At a state $\mathbf{s}^m \in S$ satisfying $w^m > 0$, after making decision for d_a^m, the decision rules for d_s^m and d_w^m are similar to those for the states $\mathbf{s}^m \in S$ satisfying $w^m = 0$. According to Corollaries 2 – 4 and Proposition 3, we divide the states $\mathbf{s}^m \in S$ satisfying $w^m > 0$ into six subsets, and then analyze the potential optimal actions and the admission rules for each subset, which are summarized in Table 4.4. The optimal action and admission rule for subset S_{11} are obtained according to Corollary 4. For the states in the other subsets, S_{12}, S_{13}, S_{13}, S_{15}, and S_{16}, the two heuristic rules (Rule A and Rule B) are proposed for making decision for d_a^m, and the rules for d_s^m and d_w^m depends on the state properties of each subset and the decision for d_a^m. At a state $\mathbf{s}^m \in S_{12}$, if $d_a^m = 0$, the only feasible decision for d_s^m and d_w^m is $d_s^m = 0$ and $d_w^m = 0$. If $d_a^m = 1$, the optimal decision for d_s^m and d_w^m is $d_s^m = 0$ and $d_w^m = 1$ according to Propositions 1 and 2. $d_s^m = 0$ and $d_w^m = 1$ are optimal at a state $\mathbf{s}^m \in S_{13}$ according to Corollary 3, while $d_s^m = 1$ and $d_w^m = 0$ are optimal at a state $\mathbf{s}^m \in S_{14}$ according to Proposition 3. For the states in subsets S_{15} and S_{16}, the three heuristic rules (Rules 5a, 5b and 5c) are considered for making decision for d_s^m and d_w^m. This is because after making decision for d_a^m, the states in subsets S_{15} and S_{16} have similar properties to those in subset S_{05}.

Table 4.4: Optimal and heuristic admission rules for the states $s^m \in \mathrm{S}$ satisfying $w^m > 0$

State subset	Conditions for the state subset	Optimal actions	Admission rules
S11	$\sum_{i=1}^{M} x_i^m = 0$, $n^m = 0$ and $w^m = 1$	(1,0,1)	Rule 4: $d_a^m = 1$, $d_s^m = 0$, $d_w^m = 1$
S12	$\sum_{i=1}^{M} x_i^m = 0$, $n^m = 0$ and $w^m > 1$	(0,0,0) or (1,0,1)	Rule A and Rule B for d_a^m If $d_a^m = 0$, follow Rule 1 for d_s^m and d_w^m. If $d_a^m = 1$, follow Rule 3 for d_s^m and d_w^m.
S13	$\sum_{i=1}^{M} x_i^m = 0$ and $n^m > 0$	(0,0,1) or (1,0,1)	Rule A and Rule B for d_a^m Rule 3 for d_s^m and d_w^m
S14	$\sum_{i=1}^{m} x_i^m > 0$ and nm ≥ 0	(0,1,0) or (1,1,0)	Rule A and Rule B for d_a^m Rule 2 for d_s^m and d_w^m
S15	$\sum_{i=1}^{M} x_i^m > 0$, $\sum_{i=1}^{m} x_i^m = 0$ and $n^m = 0$		Rule A and Rule B for d_a^m If $d_a^t = 0$, follow Rule 2 for d_s^m and d_w^m. If $d_a^t = 1$, follow Rule 5a, 5b or 5c for d_s^m and d_w^m
S16	$\sum_{i=1}^{M} x_i^m > 0$, $\sum_{i=1}^{m} x_i^m = 0$ and $n^m > 0$		Rule A and Rule B for d_a^m Rule 5a, 5b or 5c for d_s^m and d_w^m

Table 4.5 summarizes the optimal and heuristic rules for walk-in patient admission. Among these rules, Rule 1 is the optimal admission rule for state subset S_{01}, and Rule 4 is the optimal admission rule for state subset S_{11}. Rule 2 is the optimal rule for d_s^m and d_w^m in state subsets S_{02}, S_{04}, and S_{14}, while Rule 3 is the optimal rule for d_s^m and d_w^m in state subsets S_{03}, and S_{13}. The other rules in Table 4.5 are heuristic rules, including Rules 5a, 5b and 5c that are heuristic rules for d_s^m and d_w^m in state subsets S_{05}, S_{15}, and S_{16}, and Rules A and B that are heuristic rules for d_a^m in state subsets S_{12}, S_{13}, S_{14}, S_{15}, and S_{16}

4.5. Comparison of heuristic walk-in patient admission rules

We derive the optimal rules and propose heuristic rules for walk-in patient admission in the previous section. In this section, the performances of the heuristic admission rules are compared over 36 scenarios representative of the possibilities, which consider different arrival patterns of patients with appointments, different arrival rates of patients without appointments, and different overbooking policies. After that, the walk-in admission policy adopting the best heuristic rules is compared to the policy admitting all walk-in patients and the policy rejecting all walk-in patients.

Table 4.5: Summary of the optimal and heuristic rules for walk-in patient admission

Walk-in patient admission rules	Definition of admission rule
Rule 1	$d_s^m = 0$, $d_w^m = 0$
Rule 2	$d_s^m = 1$, $d_w^m = 0$
Rule 3	$d_s^m = 0$, $d_w^m = 1$
Rule 4	$d_a^m = 1$, $d_s^m = 0$, $d_w^m = 1$
Rule 5a	For a state $s^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m > 0$ and $\sum_{i=1}^{m} x_i^m = 0$, if ESP(s^m) $\leq M-m+1$, then $d_s^m = 0$, $d_w^m = 1$; otherwise, $d_s^m = 1$, $d_w^m = 0$.
Rule 5b	If $\sum_{i=1}^{M} x_i^m > 0$ and $n^m > 0$, then $d_s^m = 1$, $d_w^m = 0$.
Rule 5c	If $\sum_{i=1}^{m} x_i^m > 0$ and $n^m > 0$, then $d_s^m = 0$, $d_w^m = 1$.
Rule A (Tight Rule)	For a state $s^m \in S$, if $w^m + n^m +$ ESP(s^m) $> M-m+1$, then $d_a^m = 0$; otherwise, $d_a^m = 1$.
Rule B (Relaxed Rule)	For a state $s^m \in S$, if $n^m +$ ESP(s^m) $\geq$ M–m+1, then $d_a^m = 0$; otherwise, $d_a^m = 1$.

4.5.1. Data collection and numerical scenarios

In most outpatient clinics, a common 4-hour clinic session is divided into 15-minute or 30-minute appointment slots (Giachetti et al. 2005, Green et al. 2007, Qu and Shi 2009). In the local clinic motivating this study, the length of a clinic session is 4 hours, which is divided into eight 30-minute slots. There are 3 FTE physicians in this clinic. It takes 25 – 30 minutes for a physician to see one patient, and physicians use a few remaining minutes to do the paperwork if the service time for a patient is less than 30 minutes. Therefore, 4-hour clinic sessions with eight 30-minute appointment slots and the service time of 30 minutes are assumed in all scenarios to compare the performance of heuristic rules.

Patients with appointments may arrive before or after their appointment time, cancel their appointments, or not show up for their appointments. According to the literature, the no-show and late-cancellation rate in an outpatient clinic could reach as high as 50 – 55% (George and Rubin 2003, Lee et al. 2005), and patients who arrive earlier than their appointment times are more than those who arrive late (Blanco White and Pike 1964, Fetter and Thompson 1966, Cayirli and Veral 2003). In the study of Blanco White and Pike (1964), the distribution of patients' unpunctuality (i.e., the difference between the appointment time and the arrival time of a patient) in an outpatient department is much more peaked than a normal distribution, and patients' unpunctuality ranges from 60 minutes early to 40 minutes late, with an average of 5 minutes early and standard deviation of 17 minutes. In the outpatient clinics or departments studied by Fetter and Thompson (1966), 59.9 – 86.5% patients with appointments arrive early or on time. The average early arrival time is 17.1 minutes and the average late arrival time is 16.6 minutes.

For the local clinic, the no-show rate, the late cancellation rate, and the walk-in patient arrival rate are obtained based on 6-month historical data of appointments scheduled and patient visits. As shown in Table 4.6, the average no-show rate is 5.87%, the average late-cancellation rate is 10.59%, and the walk-in patient arrival rate is 1.57 patients per clinic session for one physician. Meanwhile, the arrival pattern of patients with appointments is approximated based on one-month patient arrival data collected by time studies in the clinic. The arrivals of patients with appointments are classified into three groups: early arrivals, on-time arrivals, and late arrivals. The three groups of arrivals are defined as the arrivals more than one time slot earlier than the corresponding appointment time, the arrivals within one time slot before the corresponding appointment time, and the arrivals later than the corresponding appointment time, respectively. Among all patients who showed up for their appointments, the percentage ranges of early arrivals, on-time arrivals and late arrivals are 4.84 – 5.76%, 70.49 – 72.58%, and 20.96 – 24.59%, respectively. The percentages of three arrival groups vary among five physicians working in the clinic.

Table 4.6: Patient arrival pattern in the clinic motivating this study

Parameter	Average/Range
No-show rate	5.87% (299 appointments missed among 5089 elective appointments)
Late cancellation rate	10.59% (539 appointments cancelled late among 5089 elective appointments)
Arrival rate of patients without appointments	1.57 patients per provider per clinic session (984 walk-in patients over 628 provider-sessions)
Percentage of early arrivals *	4.84 – 5.76% of all arrivals of patients with appointments
Percentage of on-time arrivals *	70.49 – 72.58% of all arrivals of patients with appointments
Percentage of late arrivals *	20.96 – 24.59% of all arrivals of patients with appointments

* The percentages of early arrivals, on-time arrivals and late arrivals vary among providers. Therefore, the ranges of these percentages are listed.

For the comparison of heuristic walk-in patient admission rules, four arrival patterns of patients with appointments are considered, which are summarized in Table 4.7. Among the four arrival patterns, Pattern R1 approximates the arrival pattern of patients with appointments in the clinic studied. In addition, considering that only one patient could be seen during each appointment slot, three arrival rates of patients without appointments considered are less than 1 patient per slot. The three arrival rates are 0.2, 0.5, and 0.8 patients per slot, among which the lowest arrival rate captures the arrival rate of patients without appointments in the local clinic.

4.5.2. Overbooking policies and walk-in patient admission policies

To reduce the negative impact of high patient no-shows and late cancellations, clinics could adopt overbooking and/or admit some walk-in patients. In this study, we examine the combinations of three overbooking policies and eight walk-in patient admission policies. One walk-in patient admission policy consists of a set of optimal and heuristic walk-in patient admission rules for decisions d_a^m, d_s^m, and d_w^m.

Table 4.8 summarizes the three overbooking policies and the eight walk-in patient admission policies. In the three overbooking policies, the first one does not adopt overbooking due to a low demand for appointments. The second one allows the clinic to overbook one appointment slot, which is the overbooking policy used in the local clinic. The last overbooking policy allows overbooking half of the appointment slots (i.e., four appointment slots in our numerical scenarios).

Table 4.7: Patient arrival patterns for the comparison of the heuristic rules

Patient Type	Arrival Pattern	Description
Patients with appointments	Pattern R1	20% patients arriving 1 time slot early 60% patients arriving on time 5% patients arriving late Late-cancellation rate of 10% No-show rate of 5%
	Pattern R2	12.5% patients arriving 1 time slot early 37.5% patients arriving on time 5% patients arriving late Late-cancellation rate of 30% No-show rate of 15%
	Pattern R3	40% patients arriving 1 time slot early 40% patients arriving on time 5% patients arriving late Late-cancellation rate of 10% No-show rate of 5%
	Pattern R4	25% patients arriving 1 time slot early 25% patients arriving on time 5% patients arriving late Late-cancellation rate of 30% No-show rate of 15%
Patients without appointments	Low arrival rate (LA)	0.2 patient per appointment slot
	Medium arrival rate (MA)	0.5 patient per appointment slot
	High arrival rate (HA)	0.8 patient per appointment slot

In the eight walk-in patient admission policies, the first one does not admit any walk-in patients. The second one allows all walk-in patients waiting for services, and dismisses all unseen walk-in patients at the end of a clinic session. The second admission policy adopts Rules 1 – 3 and 5a to determine whether an elective patient or a walk-in patient should be seen in each slot. In the remaining six walk-in patient admission policies, Policies A3, A4, and A5 adopt Rule A for the admission decision (d_a^m), while Policies A6, A7, and A8 adopt Rule B. For decisions d_s^m and d_w^m in state subsets S_{05}, S_{15} and S_{16}, Policies A3 and A6 adopt Rule 5a, Policies A4 and A7 adopt Rule 5b, and Policies A5 and A8 adopt Rule 5c. For each of the other state subsets, the corresponding rule (Rule 1, Rule 2, Rule 3 or Rule 4) given in Tables 3 and 4 is adopted for decisions d_s^m and d_w^m in the six walk-in patient admission policies (Policies A3 –A8). For a more detailed description of overbooking policies and walk-in patient admission, please see the Appendix.

Table 4.8: Overbooking policies and walk-in patient admission policies

Policy Group	Index	Policy Name	Description
Overbooking	Policy O1	No overbooking	On average 20% slots unfilled due to low demand
Policy	Policy O2	1-slot overbooking	1 slot is overbooked for excess patient requests for appointment
	Policy O3	4-slot overbooking	4 slots are overbooked for excess patient requests for appointments
Walk-in patient	Policy A1	No walk-in admission	No walk-in patients are admitted.
admission policy	Policy A2	All walk-in admission	All walk-in patients are admitted.
	Policy A3	Tight rule with Rule 5a	Rule A for the admission decision (dam), and Rule 5a for the state subsets S05, S15, and S16, and the rules in Tables 4.3 and 4.4 for the other state subsets
	Policy A4	Tight rule with Rule 5b	Rule A for the admission decision (dam), and Rule 5b for the state subsets S05, S15, and S16, and the rules in Tables 4.3 and 4.4 for the other state subsets
	Policy A5	Tight rule with Rule 5c	Rule A for the admission decision (dam), and Rule 5c for the state subsets S05, S15, and S16, and the rules in Tables 4.3 and 4.4 for the other state subsets
	Policy A6	Relaxed rule with Rule 5a	Rule B for the admission decision (dam), and Rule 5a for the state subsets S05, S15, and S16, and the rules in Tables 4.3 and 4.4 for the other state subsets
	Policy A7	Relaxed rule with Rule 5b	Rule B for the admission decision (dam), and Rule 5b for the state subsets S05, S15, and S16, and the rules in Tables 4.3 and 4.4 for the other state subsets
	Policy A8	Relaxed rule with Rule 5c	Rule B for the admission decision (dam), and Rule 5c for the state subsets S05, S15, and S16, and the rules in Tables 4.3 and 4.4 for the other state subsets

4.5.3. Reward and cost coefficients

For the MDP model proposed in Section 3, a walk-in patient admission policy directly determines the expected total net reward, while an overbooking policy indirectly affects the expected total net reward by changing the initial state of the MDP model. For a given walk-in patient admission policy, the expected total net reward also depends on the reward from seeing a elective patient or a walk-in patient (r_s and r_w), provider idle cost and overtime cost per appointment slot (c_i and c_o), and patient waiting costs per appointment slot (c_s and c_w).

In this study, provider idle cost and overtime cost per slot are estimated based on the median annual compensation per primary care physician reported by the Bureau of Labor

Statistics (BLS). According to the BLS reports, the median annual compensation of primary care physicians was \$186,044 in 2008 (Bureau of Labor Statistics 2011). Since on average a provider would work 50 weeks per year and 40 hours per week, the hourly cost of hiring a primary care physician is about \$90. Thus, in this study, the provider idle cost per slot (half an hour) is \$45 (i.e., $c_i = 45$). Meanwhile, considering the compensation for providers' unwillingness to work overtime in this study, the provider overtime cost per slot is assumed to be about 50% higher (i.e., $c_o = 70$). The reward earned by seeing a patient is estimated based on the average payment to the provider per primary care visit from the Agency for Healthcare Research and Quality (AHRQ) and the median annual compensation of primary care physicians. The AHRQ news reports that the average payment per primary care visit is about \$100 during 2004 (Agency for Healthcare Research and Quality 2007). The reward per elective patient equals the average payment per visit minus the cost per half an hour of hiring a primary care physician (i.e., $r_s = 55$). Since nurses and providers usually need extra effort to take care of walk-in patients, a slightly lower reward (i.e., $r_w = 50$) is considered for seeing a walk-in patient in this study. The patient waiting cost per slot is estimated based on the average hourly wage of about \$17.43 per hour (Krueger 2009). Thus, the waiting cost per slot per elective patient of \$8.7 is used in this study (i.e., $c_s = 8.7$). Since walk-in patients arrive without appointments or late for their appointments are considered lower priority to the clinic, the waiting cost per slot per walk-in patient can be treated as a much lighter penalty to service performance than that per elective patient. In this study, it is assumed that the waiting cost (or penalty) per slot per elective patient is four times of that per walk-in patient (i.e., $c_w = 2.2$). These six cost coefficients are summarized in Table 4.9.

Table 4.9: Cost coefficients in the immediate net reward

Cost Coefficient	Notation	Value
Reward per elective patient	rs	55
Reward per walk-in patient	rw	50
Provider idle cost per appointment slot	ci	45
Provider overtime cost per appointment slot	co	70
Waiting cost per appointment slot per elective patient	cs	8.7
Waiting cost per appointment slot per walk-in patient	cw	2.2

4.5.4. Comparison of walk-in patient admission policies

For each patient arrival pattern, 5000 samples are randomly generated to estimate the expected total net reward of each of 24 combinations of the overbooking policies and the walk-in patient admission polices. To reduce the effect of patient arrival patterns, the policy combinations are compared in terms of the average of the differences between the expected total net reward of each combination and the maximum total net reward in each sample. The maximum total net reward for a sample is the total net reward of the best admission decisions, which is calculated after knowing all patient arrivals in the sample. Figures 3 – 5 demonstrate the performances of 21 combinations of walk-in patient admission policies A2 – A8 with overbooking policies O1 – O3. Since the performances of admission policy A1 are much worse than those of the other walk-in patient admission policies, its performances are not illustrated in Figures 4.3 – 4.5.

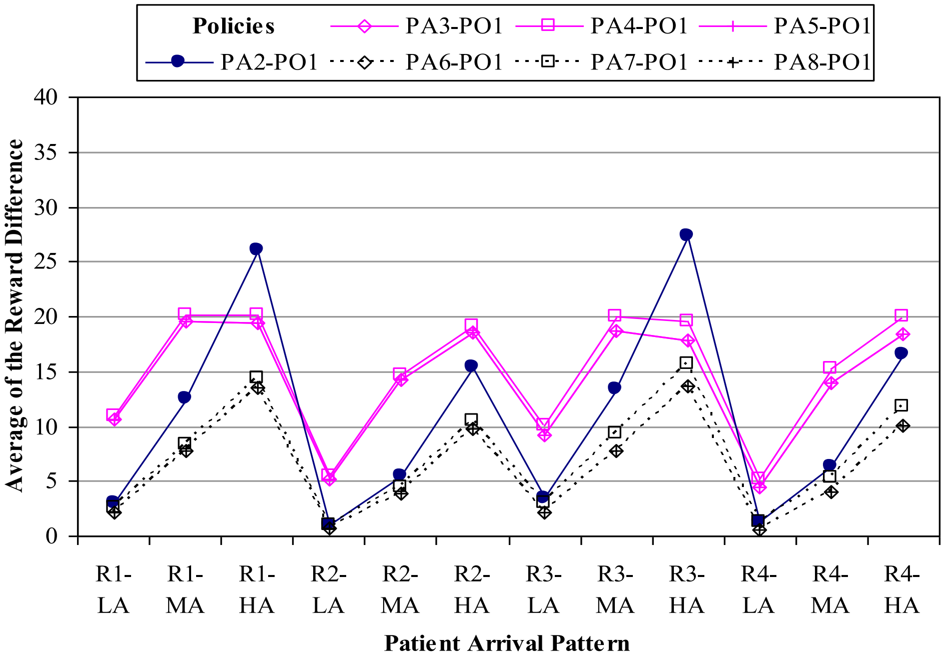

Fig. 4.3: Performance of the admission policies A2 – A8 under overbooking policy O1

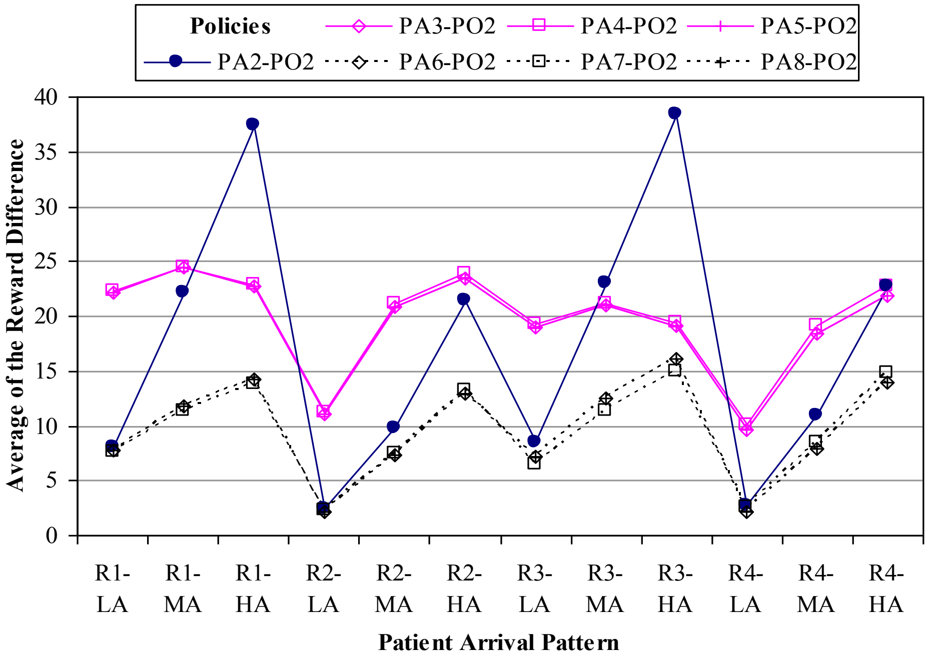

Fig. 4.4: Performance of the admission policies A2 – A8 under overbooking policy O2

Figures 4.3 – 4.5 show that the performances of admission policies A6, A7, and A8 are significantly better than those of admission policies A3, A4, and A5, which means that the

Relaxed Rule is better than the Tight Rule for the walk-in patient admission. Those figures also demonstrate that the performances of admission policies A3, A4, and A5 are similar, and that the performances of admission policies A6, A7, and A8 are similar, too. The further statistical analysis concludes that there are no significant difference in performance between admission policies A3 and A5, and no significant difference between admission policies A6 and A8. This result indicates that Rule 5c performs as well as Rule 5a. Meanwhile, the statistical analysis reveals that for some patient arrival patterns, admission policies A3 and A5 perform significantly better than admission policy A4, and admission policies A6 and A8 perform significantly better than admission policy A7. However, for the other patient arrival patterns, there are no significant performance differences among admission policies A3, A4 and A5, and no significant difference among admission policies A6, A7 and A8. These results imply that for the three heuristic rules, Rules 5a and 5c perform better than or as well as Rule 5b.

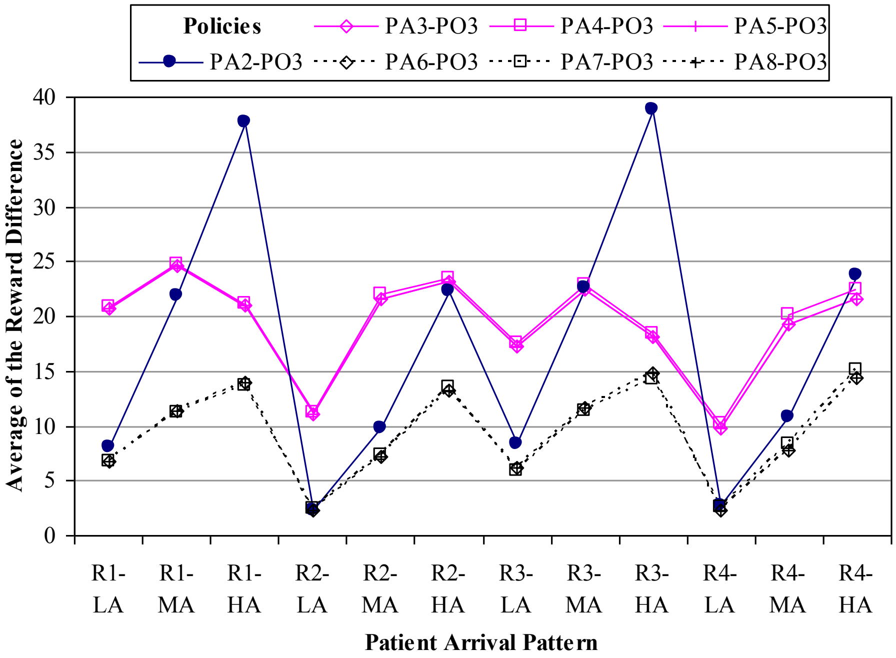

Fig. 4.5: Performance of the admission policies A2 – A8 under overbooking policy O3

Fig. 4.6 compares the performance of admission policies A1 and A2 with admission policy A6, which is the policy adopting the best heuristic admission rules. This figure demonstrates that the performance of admission policy A1 is much worse than that of admission policy A2, and deteriorates with the increase in the arrival rate of patients without appointments. This implies that admitting walk-in patients improves the expected total net reward more when the arrival rate of patients without appointments increases. In addition, Figures 4.3 – 4.6 reveal that the performance of admission policy A2 is close to that of the admission policy adopting the best heuristic admission rules in the patient arrival patterns with a low arrival rate of patients without appointments (20% of service rate). However, the performance of admission policy A2 decreases with the increase in the arrival rate of patients without appointments. Finally, the results in Table 4.10 show that the percentage of walk-in patients seen ranges from 16.18% – 84.21% when adopting admission policy A6 and overbooking policies O1 – O3.

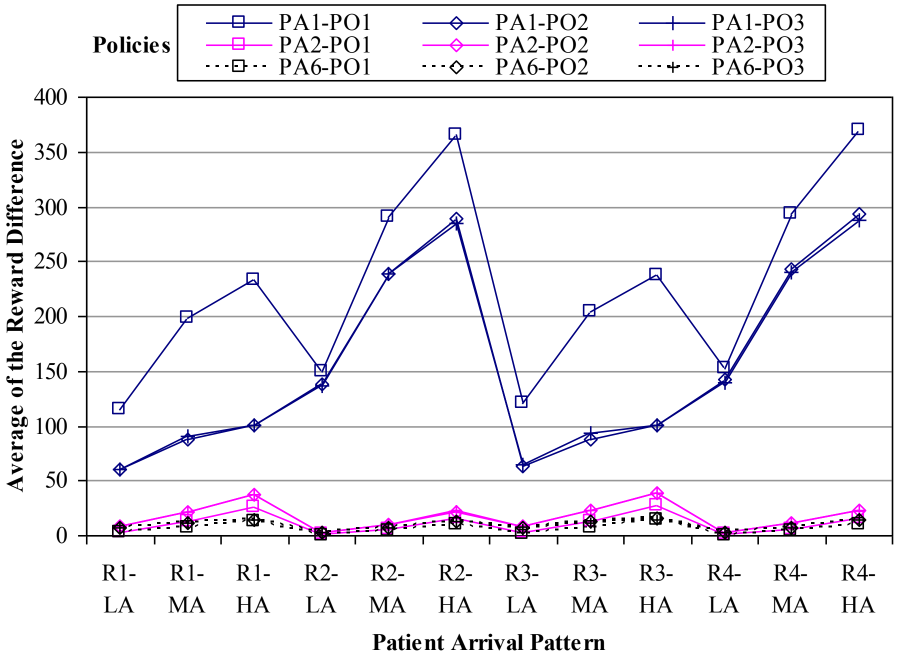

Fig. 4.6: Performance of the admission policies A1, A2, and A6 under overbooking policies O1 – O3

Table 4.10: Average number of walk-in patients seen per clinic session under admission policy A6

Arrival pattern of patients with appointments	Arrival pattern of patients without appointments	Average number of walk-in patients seen per clinic session			Average walk-in patient arrivals per clinic session		
		Policy PA6-PO1	Policy PA6-PO2	Policy PA6-PO3	Policy PA6-PO1	Policy PA6-PO2	Policy PA6-PO3
Pattern R1	Low arrival rate	1.2	0.6	0.6	1.9	2.0	2.0
	Medium arrival rate	2.1	0.9	1.0	4.3	4.4	4.4
	High arrival rate	2.5	1.1	1.1	6.7	6.8	6.8
Pattern R2	Low arrival rate	1.6	1.5	1.5	1.9	2.0	2.0
	Medium arrival rate	3.1	2.6	2.6	4.3	4.4	4.4
	High arrival rate	3.9	3.1	3.1	6.7	6.8	6.8
Pattern R3	Low arrival rate	1.3	0.7	0.7	1.9	2.0	2.0
	Medium arrival rate	2.2	1.0	1.0	4.3	4.4	4.4
	High arrival rate	2.6	1.1	1.1	6.7	6.8	6.8
Pattern R4	Low arrival rate	1.6	1.5	1.5	1.9	2.0	2.0
	Medium arrival rate	3.1	2.6	2.6	4.3	4.4	4.4
	High arrival rate	4.0	3.2	3.1	6.7	6.8	6.8

4.5.5. Discussion

The analytical results and the experimental results in this study provide insights for clinic managers to determine whether to admit a walk-in patient and in which slot a walk-in patient should be seen. This enables clinics to serve more patients and become more responsive and respectful toward their patients' needs. Patients can benefit from increased flexibility to accommodate their changing needs for long-term and short-term scheduling, and clinics can also benefit from it because the providers can be better utilized within reasonable working hours. The guidelines for the walk-in patient admission are summarized as follows:

1) A clinic should admit some or all walk-in patients because walk-in patient admission policies A2 – A8, in which all or some walk-in patients are admitted, perform much better than admission policy A1, in which all walk-in patients are rejected.

2) In a clinic with a walk-in patient arrival rate not greater than 20% of service rate, admitting all walk-in patients is a simple and good rule because in such cases admission policy A2, in which all walk-in patients are admitted, performs as well as the admission policy adopting the best heuristic rules.

3) In a clinic with a walk-in patient arrival rate greater than 20% of service rate, the Relaxed Rule should be adopted to determine whether to admit a walk-in patient. The Relaxed Rule allows admitting walk-in patients when the number of patients waiting for service plus the expected number of elective patients who will arrive on time or early is less than the total remaining slots; otherwise, walk-in patients should be rejected.

4) A walk-in patient should be seen only when there are no elective patients waiting for service, with appointments before or in the current slot.

5. MULTI-PROVIDER WALK-IN PATIENT ADMISSION OPTIMIZATION MODEL

5.1. Introduction

Under the setting of traditional appointment scheduling, patients make appointments weeks or months earlier by calling the clinic or right after their current visits. Usually, the appointments are not available in near term, since most clinics operate at their capacity. As a result, the patients need to wait several weeks or months for their clinic visits. In case of an urgent appointment, the patients may have to use the emergency department. It leads to a disruption of care continuity because they are not able to see their own providers in time. It also dramatically increases the care cost in an unnecessarily way since the cost of emergency department visits is much higher than that of primary clinic visits. Meanwhile, patient no-show rate and short-notice appointment cancellation rate are likely to increase due to the long waiting list for appointments. It is well-known that patient no-shows and short-notice appointment cancellations increase the volatility to the standard clinic process, which would eventually increase the healthcare expenditure and decrease clinic efficiency and patient accessibility.

It is also believed that clinics can reduce the adverse effects of patient no-show and short-notice appointment cancellation by admitting walk-in patients (Moore et al., 2001; Liu et al., 2010). The key concept is using walk-in patients to fill the empty appointments due to patient no-shows and short-notice appointment cancellations. In order to achieve the maximum profit and patient satisfaction, the clinics need to optimize their walk-in patient admission policy, since too many walk-in patient admissions lead to reduced patient satisfaction, while too fewer admissions can result in the loss of profit. Hence, it is worthwhile to develop optimization models/methods that can find the optimal walk-in patient admission policy where the optimal number of walk-in patients can be admitted at the right time. To solve walk-in patient admission optimization problem, we need to give answers to the following four questions: 1) When should the walk-in patient admission decisions be made? 2) At each decision point, how many walk-in patients should be admitted? 3) Which provider should serve the admitted walk-in patients? 4) When the admitted walk-in patient should be served?

In the existing outpatient clinic scheduling literature, the walk-in patient admission problem has not been sufficiently studied. Most of these studies ignore the walk-in patient by directly assuming that no walk-in patient be admitted. On the other hand, there are a few studies, which consider the walk-in patients by assuming the clinic admitting all walk-in patients. It is clearly that both assumptions have their disadvantages. If rejecting all walk-in patients, the clinic may miss the opportunity for seeing more patients, when there are available appointment slots due to patient no-show or appointment cancellation. If admitting all walk-in patients, there is a risk of overload due to the large number of the arrived walk-in patients. Hence, the best walk-in patient admission strategy should be neither rejecting all walk-in patients, nor admitting all walk-in patients. The optimal walk-in patient admitting decisions should be made based on the capacity of the clinic which is closely related to many factors, including, the current appointment schedule, patient no-show rate, appointment cancellation rate, and etc. As a result, we propose a two-stage stochastic mixed-integer programming (SMIP) model to investigate the walk-in patient admission optimization problem. The contributions of this study are listed as follows:

1) Our study is the first quantitative research investigating the dynamic walk-in patient admission optimization problem in regards of mitigating the negative impacts of patient no-shows and short-notice appointment cancellations. Hence, this study can fill the gap between the existing study and needs from clinics.

2) A novel solution approach, based on the Sample Average Approximation method, is developed to solve the SMIP, since the walk-in patient admission decision need to be made in real time or within short time.

3) The SMIP model can be applied for a broader range of clinics, since it does not have assumptions that often restrict the patient arrival pattern, service time distribution, and etc. Note that these assumptions are commonly used in other outpatient appointment scheduling studies. In addition, our model considers clinics with multiple providers.

5.2. Terminology and problem formulation

5.2.1. Terminology

Index:

i : Index of patients with appointment.

j : Index of providers.

k : Index of walk-in patients.

ω : Scenario index.

Patients with appointment:

$p_{ij}^{1} = \{t_{ij}^{A1}, t_{ij}^{S1}, t_{ij}^{E1}, t_{ij}^{L1}, t_{ij}^{W1}, I_{ij}^{C1}, I_{ij}^{N1}\}$: Denote the first patient scheduled in the i^{th} appointment slot of provider j.

$p_{ij}^{2} = \{t_{ij}^{A2}, t_{ij}^{S2}, t_{ij}^{E2}, t_{ij}^{L2}, t_{ij}^{W2}\ I_{ij}^{C2}, I_{ij}^{N2}\}$: Denote the second patient scheduled in the i^{th} appointment slot of provider j.

Walk-in patients:

$p_{k}^{w} = \{t_{kj}^{Aw}, t_{kj}^{Sw}, t_{kj}^{Ew}, t_{kj}^{Lw}, t_{kj}^{Ww}, I_{kj}^{Pw}, I_{kj}^{admit}\}$: Denote the k^{th} walk-in patient arrived at the clinic.

Parameters:

I_{ij}^{C1}, I_{ij}^{C2}: The cancellation indicator of the corresponding patient, with “1” indicating cancellation.

I_{ij}^{N1}, I_{ij}^{N2}: The no-show indicator of the corresponding patient, with “1” indicating no-show.

T_{ij}^{S}: The expected starting time of the *i*th appointment of provider *j*.

T_{j}^{E}: The expected service ending time of provider *j*.

$dummy\_t_{ij}^{A1}$, $dummy\_t_{ij}^{A2}$, $dummy\_t_{k}^{Aw}$: Dummy arrival time of the corresponding patient, which is generated from given arrival time distribution.

$dummy\_t_{ij}^{L1}$, $dummy\_t_{ij}^{L2}$, $dummy\_t_{k}^{Lw}$: dummy service length of the corresponding patient, which is generated from given service length distribution.

$c^{Wait.a}$: The cost coefficient related to waiting time of patients with appointment.

$c^{Wait.w}$: The cost coefficient related to waiting time of walk-in patients.

c^{Idle}: The cost coefficient related to provider idle time.

$c^{Overtime}$: The cost coefficient related to provide over time.

Variables:

t_{ij}^{A1}, t_{ij}^{A2}, t_{kj}^{Aw}: The arrival time of the corresponding patient.

t_{ij}^{S1}, t_{ij}^{S2}, t_{kj}^{Sw}: The service starting time of the corresponding patient.

t_{ij}^{E1}, t_{ij}^{E2}, t_{kj}^{Ew}: The service ending time of the corresponding patient.

t_{ij}^{L1}, t_{ij}^{L2}, t_{kj}^{Lw}: The service length of the corresponding patient.

t_{ij}^{W1}, t_{ij}^{W2}, t_{kj}^{Ww}: The waiting time of the corresponding patient.

I_{ij}^{A} : Appointment indicator; $I_{ij}^{A} = 1$, if at least 1 patient is scheduled in the i^{th} appointment slot of provider j .

I_{ij}^{D} : Double booking indicator; $I_{ij}^{D} = 1$, if the i^{th} appointment slot of provider j is double booked.

I_{kj}^{Pw} : Patient preference indicator; $I_{kj}^{Pw} = 1$, if the k^{th} walk-in patient is willing to be seen by provider j, otherwise $I_{kj}^{Pw} = 0$.

I_{kj}^{admit} : Admission indicator; $I_{kj}^{admit} = 1$, if the k^{th} walk-in patient is admitted by provider j.

t_j^I : The idle time of provider j.

t_j^O : The overtime of provider j.

d_{kij}^1 , d_{kij}^2 : Dummy binary variables.

5.2.2. Problem formulation

To solve the walk-in patient admission optimization problem, we develop a two-stage stochastic mixed-integer programming (SMIP) model, which is based on the SMIP model for optimizing the patient double booking strategy. For an introduction to stochastic integer programming, we refer to Birge and Louveaux (2001). In order to develop this stochastic mixed-integer programming (SMIP) model, we make the following assumptions.

1) Each clinical session is evenly divided into appointment slots; one or two patient appointment(s) can be scheduled in one appointment slot. If two patients are booked in the same appointment slot, it will be referred as double-booking or overbooking.

2) Once an appointment is made, it cannot be modified unless it is cancelled by the patient.

3) Providers only see their own patients, i.e., patients scheduled for “provider A” will not be served by other providers in the clinic.

4) All patients with an appointment, if they are not no-show or cancel their appointments, must be served by the corresponding providers within the clinic session, even if the providers have to work overtime.

5) Patients with earlier appointment, if they are not no-show or cancel their appointment, are not served later than the patients with later appointments.

6) Providers do not serve any patients before the expected start time of the first appointment, i.e., the starting point of the clinic session.

7) The admission decisions for walk-in patients are made upon arrivals of walk-in patients. At the end of the clinic session, admitted walk-in patients, who do not receive service in the session cannot be referred to other clinics nor scheduled for appointment in later clinic session. Instead, providers have to work overtime to serve all these admitted walk-in patients.

8) Walk-in patients’ preferences on providers are considered, e.g., if a walk-in patient only wants to see provider “A”, then only provider “A” can serve this patient if admitted. It is assumed that the arrived walk-in patients should have the least preference on one of the providers working in the clinic. Otherwise, no provider can serve the walk-in patients, if admitted.

Note that all these assumptions are made according to the common practice of outpatient clinics. In addition, by comparing with the existing outpatient clinic appointment scheduling literature, our model doesn’t have other assumptions, which put restrictions on the patient arrival patterns (e.g. punctual arrival and Poisson distribution), service time distributions (e.g. exponential and constant), service start time (e.g. provider doesn’t see patient before scheduled appointment start time). It is clear that, without these restrictions, our model can be applied to a broader range of clinics, which have various patient arrival patterns, and service time distributions.

The objective of the problem is to minimize the expectation of the weighted sum of patient waiting time, provider idle time and provider overtime in a clinic session, as shown in Eq. (1) (Cayirli and Veral, 2003). For patients with appointments, the waiting time is measured as the delayed time by comparing actual appointment start time with the expected appointments start time. In case that the actual appointment starts before the scheduled expected appointment start time, the waiting time will be defined as zero. As for the admitted walk-in patients, the waiting time is measured as the time difference between the actual appointment start time and the arrival time of the walk-in patients. The provider overtime is the time that provider worked after the scheduled working hour. As for the provider idle time, it is defined as the amount of time that a provider is not seeing any patient during the scheduled working hour.

Min

$$\mathrm{E}_{\omega}\left(c^{Wait.a} \cdot \sum_{j}\sum_{i}\left(t_{ij}^{W1}(\omega)+t_{ij}^{W2}(\omega)\right)+c^{Wait.w} \cdot \sum_{w} t_{wj}^{Ww}(\omega)+\sum_{j}\left(c^{Overtime} \cdot t_{j}^{O}(\omega)+c^{Idle} \cdot t_{j}^{I}(\omega)\right)\right) \quad (5.1)$$

S.T.

$$t_{ij}^{L1}(\omega)=dummy\_t_{ij}^{L1}(\omega)\cdot I_{ij}^{A} \cdot \left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right), \forall i, j. \quad (5.2)$$

$$t_{ij}^{L2}(\omega)=dummy\_t_{ij}^{L2}(\omega)\cdot I_{ij}^{D} \cdot \left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right), \forall i, j. \quad (5.3)$$

$$t_{ij}^{A1}(\omega)=dummy\_t_{ij}^{A1}(\omega)\cdot I_{ij}^{A} \cdot \left(1-I_{ij}^{C1}(\omega)\right)\cdot\left(1-I_{ij}^{N1}(\omega)\right), \forall i, j. \quad (5.4)$$

$$t_{ij}^{A2}(\omega)=dummy\_t_{ij}^{A2}(\omega)\cdot I_{ij}^{D} \cdot \left(1-I_{ij}^{C2}(\omega)\right)\cdot\left(1-I_{ij}^{N2}(\omega)\right), \forall i, j. \quad (5.5)$$

$$I_{ij}^{A} \geq I_{ij}^{D}, \forall i, j. \quad (5.6)$$

$$t_{ij}^{E1}(\omega) = t_{ij}^{S1}(\omega) + t_{ij}^{L1}(\omega), \forall i, j. \tag{5.7}$$

$$t_{ij}^{E2}(\omega) = t_{ij}^{S2}(\omega) + t_{ij}^{L2}(\omega), \forall i, j. \tag{5.8}$$

$$t_{ij}^{S1}(\omega) + M \cdot \left(1 - I_{ij}^{A} \cdot \left(1 - I_{ij}^{C1}(\omega)\right) \cdot \left(1 - I_{ij}^{N1}(\omega)\right)\right) \geq t_{i'j}^{E1}(\omega), \forall i, i', j, i > 1, and\ i' < i. \tag{5.9}$$

$$t_{ij}^{S2}(\omega) + M \cdot \left(1 - I_{ij}^{A} \cdot \left(1 - I_{ij}^{C2}(\omega)\right) \cdot \left(1 - I_{ij}^{N2}(\omega)\right)\right) \geq t_{i^{'}j}^{E2}(\omega), \forall i, i', j, i > 1, and\ i' < i \tag{5.10}$$

$$t_{ij}^{S1}(\omega) + M \cdot \left(1 - I_{ij}^{A} \cdot \left(1 - I_{ij}^{C1}(\omega)\right) \cdot \left(1 - I_{ij}^{N1}(\omega)\right)\right) \geq t_{i^{'}j}^{E2}(\omega), \forall i, i', j, i > 1, and\ i' < i \tag{5.11}$$

$$t_{ij}^{S2}(\omega) + M \cdot \left(1 - I_{ij}^{A} \cdot \left(1 - I_{ij}^{C2}(\omega)\right) \cdot \left(1 - I_{ij}^{N2}(\omega)\right)\right) \geq t_{i^{'}j}^{E1}(\omega), \forall i, i', j, i > 1, and\ i' < i \tag{5.12}$$

$$t_{ij}^{S2}(\omega) + M \cdot \left(1 - I_{ij}^{A} \cdot \left(1 - I_{ij}^{C2}(\omega)\right) \cdot \left(1 - I_{ij}^{N2}(\omega)\right)\right) \geq t_{ij}^{E1}(\omega), \forall i, j. \tag{5.13}$$

$$t_{ij}^{W1}(\omega) \geq t_{ij}^{S1}(\omega) - T_{ij}^{S}, \forall i, j. \tag{5.14}$$

$$t_{ij}^{W1}(\omega) \geq 0, \forall i, j. \tag{5.15}$$

$$t_{ij}^{W2}(\omega) \geq t_{ij}^{S2}(\omega) - T_{ij}^{S}, \forall i, j. \tag{5.16}$$

$$t_{ij}^{W2}(\omega) \geq 0, \forall i, j. \tag{5.17}$$

$$t_{j}^{O}(\omega) \geq t_{ij}^{E1}(\omega) - T_{j}^{E}, \forall i, j. \tag{5.18}$$

$$t_{j}^{O}(\omega) \geq t_{ij}^{E2}(\omega) - T_{j}^{E}, \forall i, j. \tag{5.19}$$

$$t_{j}^{O}(\omega) \geq 0, \forall j. \tag{5.20}$$

$$t_{j}^{I}(\omega) \geq T_{j}^{E} - T_{1j}^{S} + t_{j}^{O}(\omega) - \sum_{i}\left(t_{ij}^{L1}(\omega) + t_{ij}^{L2}(\omega)\right) - \sum_{k} t_{kj}^{Lw}(\omega), \forall j. \tag{5.21}$$

$$t_{j}^{I}(\omega) \geq 0, \forall j. \tag{5.22}$$

$$t_{ij}^{S1}(\omega) \geq t_{ij}^{A1}(\omega), \forall i, j. \tag{5.23}$$

$$t_{ij}^{S2}(\omega) \geq t_{ij}^{A2}(\omega), \forall i, j. \tag{5.24}$$

$$t_{ij}^{S1}(\omega) \geq T_{1j}^{S}, \forall i, j. \tag{5.25}$$

$$t_{ij}^{S2}(\omega) \geq T_{1j}^{S}, \forall i, j. \tag{5.26}$$

$$I_{kj}^{admit}(\omega) \leq I_{kj}^{Pw}(\omega), \forall k, j. \tag{5.27}$$

$$\sum_{j} I_{kj}^{admit}(\omega) \leq 1, \forall k. \tag{5.28}$$

$$t_{kj}^{Lw}(\omega) = I_{kj}^{admit}(\omega) \cdot dummy\_t_{k}^{Lw}(\omega), \forall k, j. \tag{5.29}$$

$$t_{kj}^{Aw}(\omega) = I_{kj}^{admit}(\omega) \cdot dummy\_t_{k}^{Aw}(\omega), \forall k, j. \tag{5.30}$$

$$t_{kj}^{Ew}(\omega) = t_{kj}^{Sw}(\omega) + t_{kj}^{Lw}(\omega), \forall k, j. \tag{5.31}$$

$$t_{kj}^{Sw}(\omega) \geq t_{kj}^{Aw}(\omega), \forall k, j. \tag{5.32}$$

$$t_{kj}^{Sw}(\omega) + M \cdot \left(1 - I_{kj}^{admit}(\omega)\right) \geq T_{1j}^{S}, \forall k, j. \tag{5.33}$$

$$t_{kj}^{Ww}(\omega) = t_{kj}^{Sw}(\omega) - t_{kj}^{Aw}(\omega), \forall i, j. \tag{5.34}$$

$$t_{kj}^{Sw}(\omega) - t_{ij}^{E1}(\omega) + M \cdot d_{kij}^{1} \geq 0, \forall k, i, j. \tag{5.35}$$

$$t_{ij}^{S1}(\omega) - t_{kj}^{Ew}(\omega) + M \cdot (1 - d_{kij}^{1}) \geq 0, \forall k, i, j. \tag{5.36}$$

$$t_{kj}^{Sw}(\omega) - t_{ij}^{E2}(\omega) + M \cdot d_{kij}^{2} \geq 0, \forall k, i, j. \tag{5.37}$$

$$t_{ij}^{S2}(\omega) - t_{kj}^{Ew}(\omega) + M \cdot (1 - d_{kij}^{2}) \geq 0, \forall k, i, j. \tag{5.38}$$

$$t_{kj}^{Sw}(\omega) + M \cdot \left(1 - I_{kj}^{admit}(\omega)\right) \geq t_{k'j}^{Ew}(\omega), \forall k > 1, k' < k. \tag{5.39}$$

$$t_{j}^{O}(\omega) \geq t_{kj}^{Ew}(\omega) - T_{j}^{E}, \forall k, j. \tag{5.40}$$

Beside the objective function, there are 39 constraints, as defined by Eqs. 5.2 – 5.40. To be more specific, Eqs. 5.2 – 5.3 are the service length constraint for patients with appointments, which make sure the service length equal to zero if the corresponding patient is no-show or cancels the appointment, or there is no such patient i.e. $I_{ij}^{A}=0$ or $I_{ij}^{D}=0$; otherwise, the service length is random drawn from a given distribution. Similarly, Eqs. 5.4 – 5.5 define the arrival time for patients with appointments. If the corresponding patient is no-show or cancels the appointment, or there is no such patient, then the patient arrival time will be equal to zero; otherwise the patient arrival time is drawn from a given distribution. Eq. 5.6 is the double booking constraint. It defines that if an appointment slot isn't single booked, then it cannot be double booked. In addition, Eqs. 5.7 – 5.8 define the relationship among the service start time, service ending time and service length for patients with appointments. Furthermore, Eqs. 5.9 – 5.13 are the service commitment constraints for patients with appointments. These constraints prevent a provider serving two or more patients at the same time based on the assumption that patients scheduled in early appointment slot should receive the service early, if they are not no-show or cancel the appointments. Eqs. 5.14 – 5.17 are the waiting time constraints for patients with appointments. The patient waiting time is defined as the real service start time minus the expected service start time or zero, whichever is greater. Furthermore, Eqs. 5.18 – 5.20 and Eq. 5.40 define the provider overtime, which should equal to the ending time of the last patient less the expected ending time of the clinic session, or zero, whichever is greater. In addition, Eqs. 5.21 – 5.22 define the provider idle time, which equals the length of clinic length plus the provider overtime and minus the sum of service time for each patient seen in the clinic session. Eqs. 5.23 – 5.26 are the service start time constraint for patients with appointments, which define the service start time of a patient should be earlier than the arrival time of the corresponding patient, as well as the expected service start time of the first appointment. Eqs. 5.27 – 5.28 are the walk-in patient admission constraints, which define that a walk-in patient can only be admitted by at most one of their preferred providers. Eq. 5.29 defines the service length of walk-in patients. If a walk-in patient hasn't been admitted by a provider, then the corresponding service length will be set to zero, otherwise the arrival time will be a random number drawn from a given distribution. Eq. 5.30 defines the arrival time of walk-in patients. If a walk-in patient hasn't been admitted by a provider, then the corresponding arrival time will be set to zeros, otherwise the arrival time will be a random number drawn from a given distribution. In addition, Eq. 5.31 defines the relationship among the service start time, service ending time and service length for walk-in patients. Eqs. 5.32 and 5.33 define the service starting time the walk-in patients. In general, Eq. 5.32 specifies that providers cannot see walk-in patient before their arrival; Eq. 5.33 specifies that, for admitted walk-in patients, they cannot be served before the expected starting time of the first appointment. Eqs. 5.34 – 5.39 are the service time commitment constraints for walk-in patients, which prevent a provider from serving two patients (one walk-in patient and one patient with appointment, or two walk-in patients) at the same time.

5.3. Solution approach

As we can see, the walk-in patient admission problem is formulated as a two-stage stochastic mix-integer programming model. For this type of model, it is relatively easy to evaluate the objective function for a given first-stage decision under certain scenario. However, it could be extremely difficult to evaluate the expectation of the recourse function for a given first-stage decision. A popular approach to solve the two-stage stochastic programming model is the Sample Average Approximation method, which uses the Monte Carlo sampling approach to approximate the expectation of the recourse function with sample mean of a fixed number of recourse function. In the following, we illustrate the SAA concept by using the generalized two-stage stochastic linear programming model, shown in Eqs. 5.41 and 5.42.

$$\begin{aligned}&\underset{x\in\mathbf{R}^n}{Min}\ c^{\mathrm{T}}x+\mathbf{E}[Q(x,\omega)]\\&s.t.\ Ax=b, x\geq 0,\end{aligned}\tag{5.41}$$

where $Q(x,\omega)$ is the optimal objective value of the second-stage problem:

$$\begin{aligned}&\underset{x\in\mathbf{R}^n}{Min}\ q(\omega)^{\mathrm{T}}y\\&s.t.\ T(\omega)x+W(\omega)y=h(\omega),\ y\geq 0.\end{aligned}\tag{5.42}$$

Note that ω represents random scenarios for the second stage. Here $(q(\omega), T(\omega), W(\omega), h(\omega))$ is random vector with respect to realization ω, which contains the data for the second stage.

In case there is only finite number of scenarios for the second stage, namely $\omega_1,\omega_2,\cdots\omega_n$, with respective probability $p_1, p_2,\cdots p_n$, the expectation term in Eq. 5.43 can be rewritten as:

$$\mathbf{E}[Q(x,\omega)]=\sum_{k=1}^{n}p_k\cdot Q(x,\omega_k)\ \ .\tag{5.43}$$

However, in most case, there is infinite number of scenarios for the second stage. It will be hard to calculate $\mathbf{E}[Q(x,\omega)]$ in this situation. Hence, as proposed by the SAA method, the expectation will be approximated by the sample mean of the optimal objective values obtained from sample scenarios in second stage, i.e.

$$\mathbf{E}[Q(x,\omega)]=\frac{1}{n}\cdot\sum_{k=1}^{n}Q(x,\omega_k)\,,\tag{5.44}$$

where $\omega_1,\omega_2,\cdots\omega_n$ is a sample of n scenarios drawn from population of infinite realizations.

For the walk-in admission optimization problem, there is infinite number of scenarios in the second stage, due to the possible continuous distribution of service time and patient arrival time. On other hand, even if there is finite number of second stage scenarios, the number will be extremely large and it will be hard to accurately estimate the probability of each scenario, given the random service time, patient no-show, appointment cancellation and patient arrival. Hence, we apply the SAA method to approximate the expected value of recourse function. To be more specific, the objective value function in Eq. 5.1 is approximated as $\frac{1}{n}\sum_{\omega\in\Omega_s}\left(c^{Wait}\cdot\sum_j\sum_i\left(t_{ij}^{W1}(\omega)+t_{ij}^{W2}(\omega)\right)+\sum_j\left(c^{Overtime}\cdot t_j^O(\omega)+c^{Idle}\cdot t_j^I(\omega)\right)\right)$, where n is the sample size, and Ω_s is set of sample scenarios randomly drawn from the entire scenario space.

As mentioned early, the admission decisions are made upon the arrival of walk-in patients. In practice, each walk-in patient has a unique arrival time. Hence, the admission decisions, which are the first-stage decision variables, are made for only one walk-in patient at a time. For example, upon the arrival of the k^{th} walk-in patient, only the variables $I_{kj}^{admit}, j=1,2,..J$ are the first stage decision variables, i.e. $I_{kj}^{admit}(\omega)=I_{kj}^{admit}$ for all ω. Note that variables $I_{k'j}^{admit}, k'=1,2,...k-1$, which are the admission decisions for previous walk-in patients, are used as parameters, which is already known at current time, to make admission decisions for the k^{th} walk-in patient. In addition, variables $I_{k'j}^{admit}(\omega), k'=k+1,k+2,.....K$, which are admission decisions for future walk-in patients will be the second-stage decision variables, which are determined under different scenarios. In this way, the model can be solved with the commercial mixed integer programming (MIP) optimization solvers, such as CPLEX and GAMS. For the state-of-the-art research on applying sample average approximation to SMIPs, we refer to Kleywegt et al. (2002), and Shapiro and Homem-de-Mello (2000). It is obvious that large scenario sample size can lead to better approximation of the expected value of recourse function and improve the solution quality. However, the computational time for find the optimal solution by using sample average approximation can be increase exponentially with the increase of scenario sample size drawn from the entire scenario space. Large sample size may lead to extremely large computation time (unacceptable) or even make it impossible to locate the optimal solution.

Unfortunately, even for a small sample size, this model cannot be efficiently solved to make the real time walk-in patient admission decisions. The reason is that SAA approach solves the first-stage and second-stage decision simultaneously, which leads to a large solution space and high computational time consumption. Based on this understanding, we propose an alternative solution approach, which is based on the above mentioned SAA method. In the following we summarize the alternative solution approach in the following:

Step 1: Find out the set of feasible solution candidates $S=\{s_1,s_2,s_3,\ldots\ldots s_m\}$for the first-stage decision variables.

Step 2: For each solution candidate "s_i" identified in Step 1, generate sample scenarios of size n, namely, $\omega_1, \omega_2, \cdots \omega_n$.

Step 3: For each solution candidate "s_i" identified in Step 1, solve the second-stage problem for each individual scenario ω_j, and calculate sample mean "f_i" of the optimal second-stage objective values, which is obtained under different the sample scenarios.

Step 4: The solution candidate s_i with the minimal "f_i" is the best first-stage solution.

By using this approach, the first-stage decision variables and second-stage decision variables are not determined simultaneously. In addition, the second-stage problems are solved individually for each scenario. In this way, the computational time is expected to increase linearly with increasing number of scenarios. This is because the number of scenarios only determines the number of second-stage problems, while the complexity of each second-stage problem remains unchanged. By comparing with the original SAA approach, in which the computational time increase exponentially with the increasing number of scenarios, the modified SAA approach is more efficient for problems need large number of scenarios. Note that this modified SAA approach is only suitable for problems with finite first-stage solution candidates, since we need to enumerate all first-stage solution candidates during the process. For our problem, the first stage decision variables are $I_{kj}^{admit}, j = 1,2,..J$, which only takes values of 0 or 1. Clearly, our problem has finite first-stage solution candidates. Hence, the modified SAA approached can be applied to solve this walk-in patient admission problem.

5.4. Numerical analysis

In this section, we conduct numerical analysis to examine the performance of the optimized the walk-in patient admission decisions, in term of clinic cost (objective value), which is the weighted sum of waiting time for patients with appointments, waiting time for walk-in patients, provider idle time and provider overtime. For this purpose, nine different cases are constructed, which considers the variability of patient non-attendance rate, patient service time, patient arrival pattern and cost coefficients related to patient waiting, provider idle and provider overtime. The nine cases are examined under two different walk-in patient demand rates.

5.4.1. Data collection and experiment design

In a most outpatient clinics, a typical 4-hour clinic session is evenly divided into multiple 15-minute or 30-minute slots (Giachetti et al. 2005, Green et al. 2007, Qu and Shi 2009). In the local clinic that motivates this study, the length a clinic session is 4 hours, which is divided into 8 30-minutes slots. Hence, in this study, we consider a 4-hour clinic session with 8 30-minutes slots and two providers (2 FTE) working at the same time in a clinic session. In addition, it is assumed that all appointment slots have been booked with one patient, i.e. double booking strategy is not considered at the planning stage. By doing this, we eliminate the impact of double booking on the patient waiting time, provider idle time and provider overtime.

For a patient with appointment, there is a possibility that he/she cancels the appointment or does not show up for the appointment, which is often referred as non-attendance in practice. The patient non-attendance rate could vary from clinic to clinic. According to the literature, the patient non-attendance rate in U.S. primary care clinics can be as low as 5% and as high as 55%. In the following, we summarize the no-show and appointment cancellation statistics found in the existing literature.

1) Johnson et al. (2007) indicate that the no-show rate vary from 3% to 42%, with an average of 17%.

2) George and Rubin (2003) report that the non-attendance rate (no-shows and cancellations) in U.S. primary care clinics range from 5% to 55%.

3) Al-Shammari (1992) and Hermoni et al. (1990) report non-attendance rates of 29.5% and 36%, respectively.

4) Moore et al. (2001) suggest that no-shows and cancelled appointments combined amount 31.1% of appointments.

For this study, we consider three different levels of non-attendance rate for patient with appointment, namely, low non-attendance rate (no-show: 3%, cancellation: 2%), medium non-attendance rate (no-show: 17%, cancellation: 13%), and high non-attendance rate (no-show: 42%, cancellation: 13%). Note that the low non-attendance rate is corresponding to the lower bound of non-attendance rate found in the literature, while the high non-attendance rate is corresponding to upper bound of non-attendance rate found in the literature. As for the medium non-attendance rate, it is calculated by averaging the lower and upper bound of the non-attendance rates found in the existing studies.

For the patients who show up for their appointment, they may arrive on time or early. In literature, many studies assume all patients arrive punctually for their appointment if not no-show or cancel their appointment (Zacharia and Pinedo, 2014; Muthuraman and Lawley, 2008; LaGanga and Lawrence, 2012). On the other hand, there are also tremendous studies assuming non-punctual arrivals of patients. In the following, we summarize the representative studies.

1) Alexopoulos et al. (2008) use the exponential distribution to model the interval time of patients. It is indicated that the parameters chosen for the exponential distribution may varies based on the clinics characteristic and patient population.

2) Fontantesi et al. (2002) indicate that patient arrivals tend to be “clumped” due to the common bus schedule, traffic light time, and availability of parking space. According to their study, most patients arrive 15 minutes earlier or 10 minutes late for their appointment. On average, the patients arrive 3 – 4 minutes before their scheduled appointment time.

3) Parmessar (2010) apply the appointment driven arrival in his simulation study for appointment optimization. The appointment driven arrivals assume that the patient should be arriving within a certain time interval, which is based on the schedule appointment time. For example, if a patient has an appointment at time t_0, then the patient will arrive at a random time drawn from the interval $[t_0 - a, t_0 + b]$, where a and b are positive constant that determine the width of the interval.

It can be seen that, the patient arrivals can be modeled in two different ways. One is to use the inter-arrival time between two patients. In this way, the inter-arrival time usually subjects to a given distribution, e.g., an exponential distribution. As a result, the arrival time of a patient depends on the arrival time of the previous patient. The other way to model the patient arrival is by using arrival lead time, which measures how long in advance a patient arrives for the appointment. In this way, the patient arrival lead time is assumed to follow a given distribution. In our study, we use the lead time to model to arrival patterns of patient with appointment. To be more specific, we consider three different patient arrival patterns. For the first patient arrival pattern, we consider punctual arrivals, i.e. arrival lead time equal to constant zero. For the second patient arrival pattern, we assume the arrival lead time to follow exponential distribution with the mean equal to 4 minutes, i.e. exponential (4) min. For the last patient arrival pattern, we assume patient arrives randomly within 2 hours before the scheduled appointment time, i.e. lead time follows uniform distribution with lower bound 0 min and upper bound 120 min. Note that for all arrival patterns, the actual patient arrival time equals to the scheduled appointment time less the arrival lead time, or zero, whichever is larger. Clearly, the first arrival pattern has the least variability (i.e., variance of the arrival lead time); the second arrival pattern has the medium variability, while the third arrival pattern has the highest variability.

As for arrival the patterns of walk-in patients, it cannot be modeled by using the arrival lead time, since the walk-in patients does not have an appointment. Instead, we use the arrival rate to capture the walk-in patient arrival pattern. According to the literature, the mean arrival rate of walk-in patients can be modeled as some fraction of the clinic capacity, e.g., 50%. In our study, the capacity of the clinic is 4 patients per hour. Based on this, we consider two different walk-in patient arrival rate, namely, current rate (1 per hour, 25% capacity), future rate (2 per hour, 50%). The current arrival rate corresponds to the current walk-in patient demand level of a local clinic. The future rate, which doubles the current arrival rate, is used to represent the possible future walk-in patient demand level. Note that for both walk-in patient arrival rates, it is assumed the number of arrived walk-in patient in one hour follows the Poisson distribution, i.e. the inter-arrival time of walk-in patients follows the exponential distribution. As for the walk-in patient's preference on providers, we assume that a walk-inpatient has preference only on "Provider 1" with probability 1/3, has preference only on "Provider 2" with probability 1/3, and has preferences on both providers with probability 1/3.

Although each appointment slot has a length of 30 minutes, the service time of patient does not necessarily have to be 30 minutes. According to the literature, the service time distribution varies among different clinics, patient populations, service types, providers, and etc. Many studies have used different distributions to describe the service time of patients. In the following, we summarize some of the representative studies.

1) LaGanga and Lawrence (2012) assume a constant patient service time, which equals to the length of an appointment slot in their clinic overbooking study.

2) Qu et al. (2013) assume the patient service time follows the lognormal distribution in their outpatient appointment optimization study based on a Women's clinic.

3) Shi et al. (2014) assume the patient service time follows the Gamma distribution in their patient flow simulation study based on a local VA medical center.

Clearly, there are various assumptions for the patient service time distribution, which are related to the studied clinics and patient groups. In the numerical study, we can consider three different service time distributions, which are Gamma (2.9898, 9.10383) minutes, Lognormal(3.0479, 0.71566) minutes, and constant 27.22 minutes. Parameters for the Gamma distribution are estimated based on the service time data collected from a local clinic, while the parameters for the lognormal distribution are calculated by using the same mean while doubling the variance of the Gamma distribution. The constant 27.22 minutes are chosen based on the mean service time drawn from the Gamma distribution.

The cost coefficients in the objective function should somehow represent the value of time for patients and providers. With this understanding, the cost coefficients are chosen based on the hourly wages of all occupation and primary providers in the United States. According to the Bureaus of Labor Statistics (BLS, 2013), the 10th, 50th and 90th percentiles of national hourly wage in 2012 are $8.7, $16.71, and $41.74, respectively, over all U.S. industry sectors. The average hourly wage of Family and General Practitioners is $86.95 in 2012. In addition, by considering the compensation for providers to work overtime, the hourly wage for providers working overtime is assumed to be 1.5 times of the regular hourly wage. In addition, since the walk-in patients don't have any appointment in advance, they should have lower priority as compare to the patient with appointments. In this consideration, we assume that the waiting cost per unit time for walk-in patient should be less as compared to that for patients with appointments. In particular, we set the cost coefficient related to walk-in patient waiting time to a quarter of the cost coefficient related to the waiting time of patients with appointments. As a result, in the numerical study, three sets of the cost coefficients for waiting time of walk-in patients, waiting time of patients with appointment, provider idle time, and provider overtime are considered. The three ratios are 0.25:1:10:15, 0.25:1:5.2:7.8 and 0.25:1:2.1:3.1 corresponding to the 10th, 50th and 90th percentiles of national hourly wage, respectively. It can be interpreted that the three ratios are corresponding to low income patients, medium income patients and high income patients.

Based on the above discuss, we have considered different types of patient non-attendance rate, patient arrival patterns, patient service time distribution and cost coefficients as summarized in Table 5.1 – 5.4. As mentioned above, we design nine different cases for the numerical analysis, namely, Case 0 – Case 8. Case 0 will be the base case, which uses the patient non-attendance rate, arrival pattern 2 for patient with appointment, medium arrival rate for walk-in patient, service time distribution 2 and cost coefficients that corresponding to medium income patients. The parameters used for Case 0 are shown in Table 5.5. As for Case 1 – Case 8, each of them represents a certain extreme condition by changing only one or a few parameters from the base case.

1) Cases 1 & 2 represent the situations of high attendance rate and low attendance rate, respectively, by altering the no-show rate and cancellation rate, simultaneously.

2) Cases 3 & 4 represent the situations high variance and low variance of patient service time by altering the patient service time distribution.

3) Cases 5 & 6 represent the situations of high variance and low variance of patient arrival time, respectively, by altering the patient arrival lead time distribution.

4) Cases 7 & 8 illustrate the situation of provider seeing low-income patients and high-income patients, respectively, by altering the cost coefficients, i.e. c^{Wait}, c^{Idle} and $c^{Overtime}$.

Table 5.1: Patient non-attendance rates for numerical analysis

Name	Parameters	Value
Low non-attendance rate	No-show rate	3%
	Cancellation rate	2%
Medium non-attendance rate	No-show rate	17%
	Cancellation rate	13%
High non-attendance rate	No-show rate	42%
	Cancellation rate	13%

Table 5.2: Patient arrival patterns for numerical analysis

Patient types	Arrival pattern	Description
Patient with appointment	Pattern 1	Punctual arrival: arrival lead time = 0
	Pattern 2	Medium variability arrival: arrival lead time follows exponential distribution with mean equal to 4 minutes.
	Pattern 3	High variability arrival: arrival lead time follows uniform distribution, with lower bound equal to 0 minute and upper bound equal to 120 minutes
Walk-in patient	Current rate	Inter-arrival time follows exponential distribution with mean equal to 60 minutes
	Future rate	Inter-arrival time follows exponential distribution with mean equal to 30 minutes

Table 5.3: Patient service time distributions for numerical analysis

Name	Description
Service time distribution 1	Service time distribution with low variance: constant 27.22 minutes
Service time distribution 2	Service time distribution with medium variance: Gamma (2.9898, 9.10383) minutes
Service time distribution 3	Service time distribution with high variance: Lognormal(3.0479, 0.71566) minutes

Table 5.4: Cost coefficients for numerical analysis

Name	Parameters	Value
Cost coefficient for low income patients	$c^{Wait.w}$	0.25
	$c^{Wait.a}$	1
	c^{Idle}	10
	$c^{Overtime}$	15
Cost coefficient for medium income patients	$c^{Wait.w}$	0.25
	$c^{Wait.a}$	1
	c^{Idle}	5.2
	$c^{Overtime}$	7.8
Cost coefficient for high income patients	$c^{Wait.w}$	0.25
	$c^{Wait.a}$	1
	c^{Idle}	2.1
	$c^{Overtime}$	3.1

The altered parameters for Cases 1-8 are shown in Table 5.6. Note that, in each case, except for the altered parameters, all the remaining parameters are the same as those in the base case. For example, in Case 8, the cost coefficients are changed to 0.25:1:2.1:3.1, which correspond to the 90th percentile of national hourly wage. However, all other parameters remain the same as the base case.

Table 5.5: Parameters used for the base case

Parameters	Rate/Distribution	Parameters	Rate/Distribution
Session length	4 hours	Patient arrival lead time distribution	Exponential(4) minutes
Number of appointment for each provider	8	$c^{Wait.w}$	0.25
Length of each appointment	30 minutes	$c^{Wait.a}$	1
No-show rate	17%	c^{Idle}	5.2
Cancellation rate	13%	$c^{Overtime}$	7.8
Service time distributions	Gamma(2.9898, 9.10383) minutes		

Table 5.6: Parameter adjustments for Cases 1-8 compared with base case

Case number	parameters	Rate /Distribution
Case 1	No-show rate	3%
	Cancellation rate	2%
Case 2	No-show rate	42%
	Cancellation rate	13%
Case 3	Service time distributions	Lognormal(3.0479, 0.71566) min
Case 4	Service time distributions	Constant 27.22 min
Case 5	Patient arrival lead time distribution	Uniform (0,120) min
Case 6	Patient arrival lead time distribution	0
Case 7	c^{Idle}	10
	$c^{Overtime}$	15
Case 8	c^{Idle}	2.1
	$c^{Overtime}$	3.1

5.4.2. Numerical analysis results

The proposed two-stage stochastic mixed-integer programming model is solved by using CPLEX solve for each case shown in section 4.1. The solver is run on a personal computer with an Intel 2.67GHz i5 dual-core processor and 2.9GB RAM. It takes 30 – 80 seconds to find the optimal solution for each of the 9 cases based on the scenario sample size of 100.

In Fig. 5.1, we compare the optimal objective values obtained through walk-in patient admission to the optimal objective values through overbooking. Clearly, for all cases, the walk-in patient admission achieves a lower objective value (cost) under both current and future walk-in demand levels. The reason is that walk-in patient admission decisions are made based on the real time situations at the appointment day, while the double booking decisions are made at the planning stage before the appointment day, based on the estimated no-show and appointment cancellation rate. In addition, by comparing the optimal objective values between current walk-in demand level with future walk-in demand level, it can be seen that the higher walk-in arrival rate (future level) leads to a lower objective value. Hence, a high walk-in patient arrival rate should be preferred by the clinic, since it creates more opportunities for providers to see more patients when they have time.

In Fig. 5.2, we compare the average waiting time of walk-in patients to the average waiting time of patient with appointments. In general, for all cases, the walk-in patient waiting time is much higher as compared to the waiting time of patient with appointment. The phenomenon is commonly seen in clinics, where walk-in patients have to wait until providers finish their work with patients who have appointments, since priorities are given to patients with appointments. In our model, we also assume priorities to patients with appointment by setting the cost coefficient of walk-in patient waiting time as a quarter of that of patient with appointment. In addition, when the patient waiting times between the current walk-in patient demand level and future demand level are compared, it can be seen that the patient waiting time for both walk-in patients and patients with appointments are higher for the future demand level (except for Case 4). Note that, for Case 4, the waiting time for patients with appointments is slightly higher under the current demand as compared to that under the future demand. This may be caused the by statistical error.

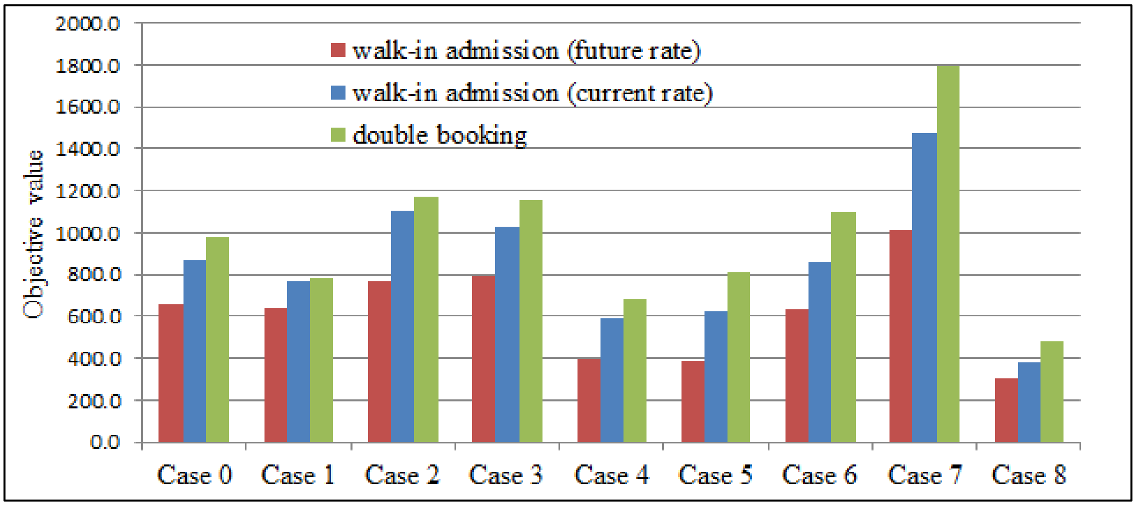

Fig. 5.1: Comparison of optimal objective values between walk-in admission strategy and double booking strategy

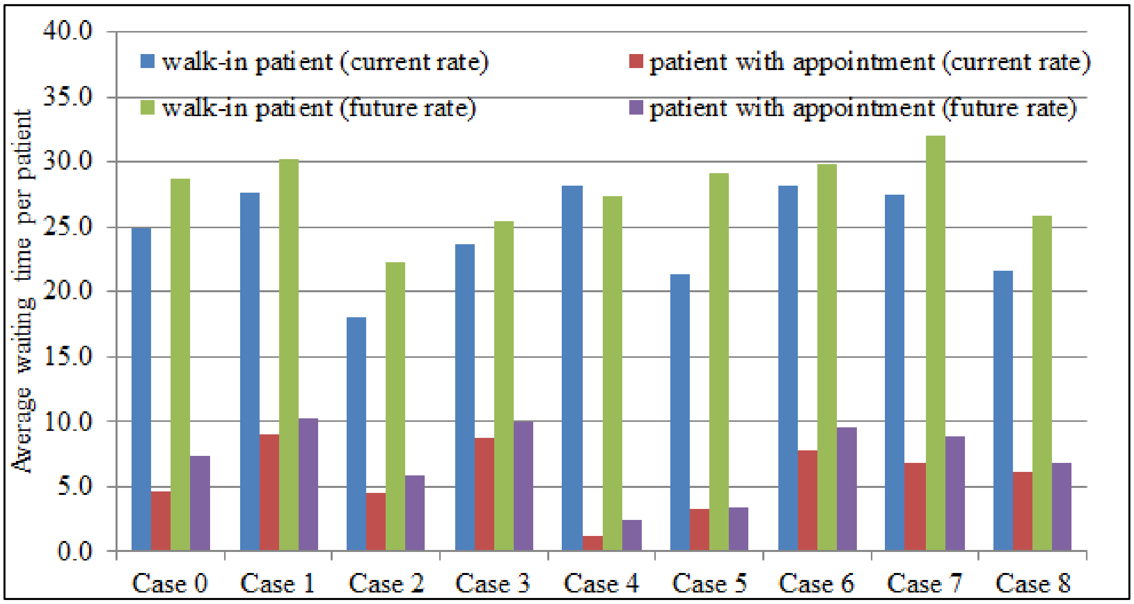

Fig. 5.2: Comparison of patient waiting times between walk-in patients and patients with appointments

In Fig. 5.3, we compare the provider idle time under the current walk-in patient demand level to that under the future walk-in demand level. Clearly, for all cases, the provider idle time is much less under the future walk-in demand (high) level than under the current walk-in demand (low) level. Hence, with more walk-in arrivals, the provider idle time, which is caused by patient no-show and appointment cancellation, can be significantly reduced. Similarly, in Fig. 5.4, we compare the provider overtime under current walk-in demand level to that under future walk-in demand level. As we can see, for all cases, the provider overtime is only slightly higher under the future walk-in demand (high) level. Hence, it can be concluded that high walk-in demand can help to reduce the provider idle time significantly without much increase in provider overtime. Note that "Case 4" has significantly less provider overtime as compare to other cases. This is because constant service time is assumed in "Case 4" and service time variability is one of the major issue that cause provider overtime.

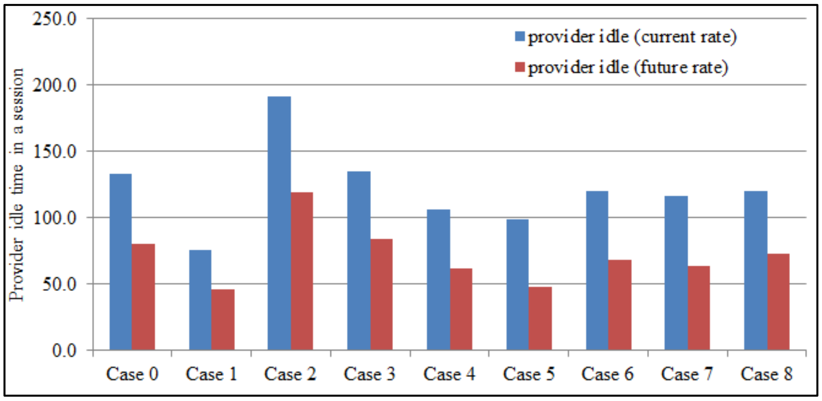

Fig. 5.3: Comparison of provider idle times under different walk-in patient arrival rates

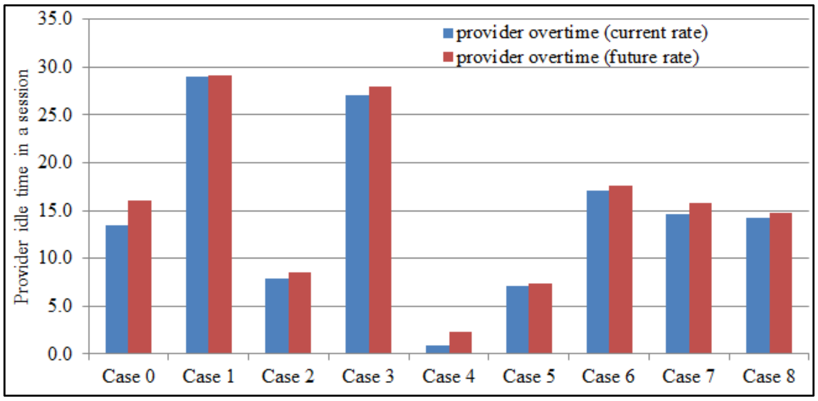

Fig. 5.4: Comparison of provider overtimes under different walk-in patient arrival rates

In Table 5.7, we summarize the descriptive statistics of performance metrics, including objective value, waiting time for patient with appointment, waiting time for walk-in patients, provider idle time, provider overtime, and number of walk-in patient admission, under the current walk-in demand level (low). Similarly, the descriptive statistics of performance metrics for future walk-in demand level (high) is summarized in Table 5.8. The results reveal a few interesting phenomena which are commonly seen in practice. For instance, Case 1 and Case 0 have the same clinic settings except for the attendance rate, where it is higher for Case 1. The statistics indicate that Case 1 has a lower objective value as compared with Case 0. This supports the general concept that high attendance rates are preferred in clinics, although walk-in admission is adopted to mitigate the adverse effect of patient no-show and appointment cancellation. This concept can also be observed by comparing Case 2 with Case 0, where the attendance rate is higher for Case 0. For another instance, Case 3 and Case 4 have the same clinic settings except for the patient service time distribution, where Case 4 has constant service time for all patients. As a result, Case 4 has a lower objective value than Case 3. This supports the general concept that service time variability is not preferred for clinics. In addition, Case 5 and Case 6 also have the same clinic settings except for the patient arrival lead time distribution, where Case 6 assumes punctual arrival of patients, i.e., arrival lead time equals to zero. As a result, Case 6 has a higher objective value. This supports the general concept that clinics prefer patients to arrive earlier for their appointments. To add more, Case 7 and Case 8 also have the same clinic settings except for the cost coefficient ratio, where Case 7 represents the scenario of low-income patients by using a high ratio. The statistics indicate that Case 7 have a higher objective value compared with Case 8. The implication is that high-income patients are preferred by the clinics.

Table 5.7: Summary of descriptive performance statistics for Cases 0- 8 (current rate)

		Case0	Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case8
Objective	mean	869.0	765.1	1106	1031	594.0	622.1	864.3	1478	379.5
	std of mean	8.42	7.44	8.42	10.61	6.18	7.78	7.99	15.50	3.25
Waiting (patient with appointment)	mean	48.2	130.0	33.3	96.3	13.6	35.5	84.5	74.8	66.7
	std of mean	1.08	2.48	1.02	3.62	0.67	1.47	1.85	1.70	1.65
Waiting (walk-in patients)	mean	87.4	61.7	67.6	78.1	78.9	72.4	84.5	83.2	61.2
	std of mean	2.32	1.46	1.95	1.80	1.59	2.05	2.27	2.23	1.55
Provider idle	mean	133.4	75.7	191.5	135.4	106.6	98.6	120.3	116.3	120.6
	std of mean	1.67	1.06	1.74	1.56	1.29	1.56	1.51	1.55	1.50
Provider overtime	mean	13.5	29.0	7.8	27.02	0.83	7.2	17.0	14.6	14.3
	std of mean	0.50	0.71	0.35	0.97	0.05	0.37	0.55	0.49	0.49
walk-in patient admissions	mean	3.5	2.2	3.8	3.5	3.4	3.5	3	3.4	3
	std of mean	0.05	0.04	0.04	0.17	0.14	0.18	0.17	0.17	0.16

Table 5.8: Summary of descriptive performance statistics for Cases 0- 8 (future rate)

		Case0	Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case8
Objective	mean	660.2	641.5	767.1	796.8	396.3	385.1	635.6	1009	305.3
	std of mean	7.58	7.51	7.91	10.35	4.65	6.32	7.00	13.43	3.01
Waiting (patient with appointment)	mean	78.5	147.4	43.6	109.5	26.3	37.5	105.2	96.5	75.2
	std of mean	1.79	2.52	1.11	3.54	0.73	1.50	1.96	1.82	1.66
Waiting (walk-in patients)	mean	154.7	108.5	144.0	137.2	123.4	154.1	151.9	165.1	122.0
	std of mean	2.46	2.06	2.97	2.48	2.30	3.48	3.19	3.45	2.36
Provider idle	mean	80.4	46.1	119.5	83.7	61.8	48.3	68.3	63.4	73.3
	std of mean	1.41	0.86	1.67	1.44	1.01	1.24	1.22	1.26	1.19
Provider overtime	mean	16.0	29.1	8.5	27.9	2.3	7.4	17.6	15.8	14.7
	std of mean	0.56	0.72	0.35	0.95	0.08	0.37	0.54	0.49	0.49
walk-in patient admissions	mean	5.4	3.6	6.5	5.4	5.2	5.4	5.1	5.4	5.3
	std of mean	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.05

In Fig. 5.5, the impact of patient attendance rate on the optimal number of walk-in patient admissions is shown. As we can see, for both low walk-in demand and high walk-in demand, the high number of walk-in admissions is associated with the low patient attendance rate, and the low number of walk-in admissions is associated with the high patient attendance rate. In addition, a linear relation between patient attendance rate and the number of walk-in admissions is observed under the high walk-in demand. As for the low walk-in demand, we are also expecting the linear relation between the patient attendance rate and the number of walk-in admissions. However, the low walk-in demand level could hinder the number of walk-in admission from further increasing for the low patient attendance rate. In the same manner, the impact of patient arrival lead time, patient service time distribution, and cost coefficient ratio on the optimal number of walk-in patient admissions are shown in Figs. 5.5 – 5.8. In general, the impacts of these factors are not significant, i.e., the number of optimal number does not vary significantly with the change of patient arrival lead time, patient service time distribution, and cost coefficient ratio. Note that although the low arrival lead time seems to reduce the walk-in patient admissions, as shown in Fig. 5.6, the effect is not as significant as that of patient attendance rate. Hence, it can be concluded that the patient attendance rate is the major factor that determines the optimal number of walk-in patient admissions, while other factors including arrival lead time, service time distribution and cost coefficient ratio, do not affect the number of walk-in patient admission significantly.

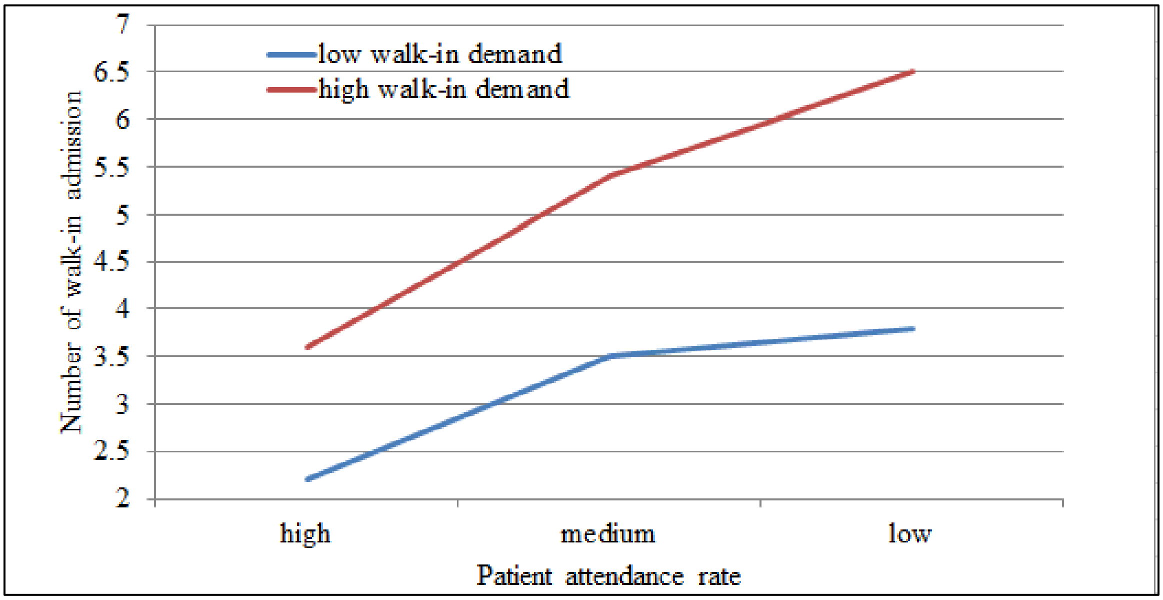

Fig. 5.5: Impact of patient attendance rate on the number of walk-in admissions

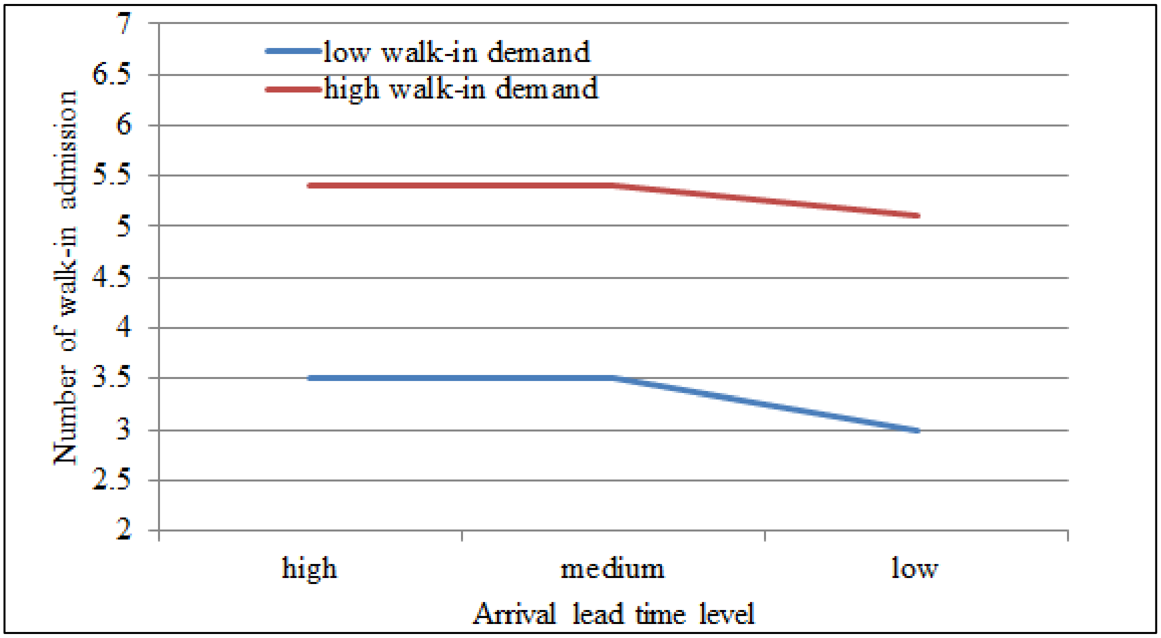

Fig. 5.6: Impact of patient arrival lead time on the number of walk-in patient admissions

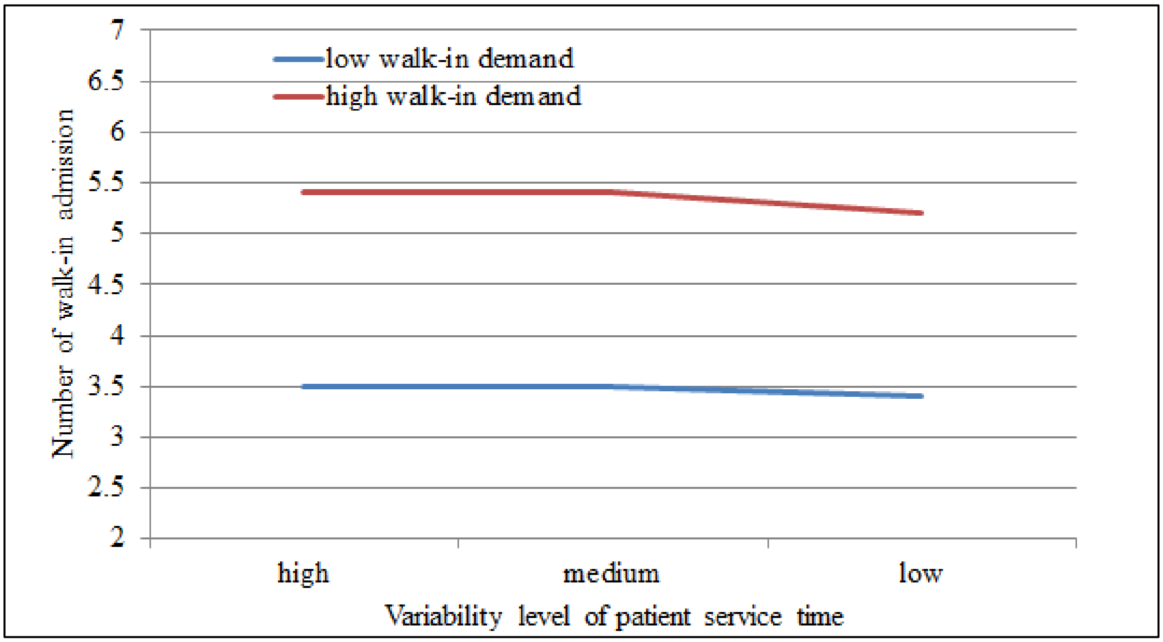

Fig. 5.7: Impact of patient service time on the number of walk-in admissions

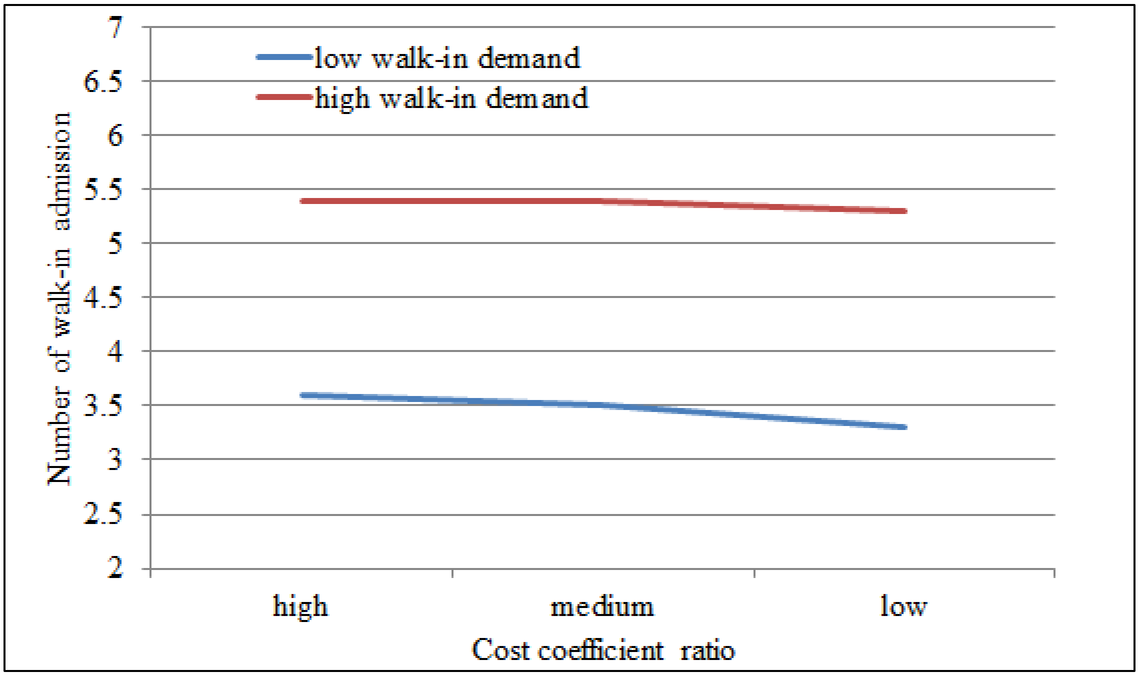

Fig. 5.8: Impact of cost coefficient ratio on the number of walk-in admissions

To further investigate whether there is a linear relation between the number of walk-in patient admissions and patient attendance rate, we consider a case with punctual patient arrival and constant service time of 30 minutes. Note that all other parameters, except for no-show and appointment cancellation rate, are set the same as the base case. In order to study the effect of no-show rate, the appointment cancellation rate is fixed as zero, vice versa. By doing so, the coupling effect of patient arrival time variability and service time variability is eliminated. Since patient no-show and appointment cancellation make no difference in our model, we only discuss the effect of patient no-show in the following. The effect of appointment cancellation is expected to be the same as the effect of patient no-show.

In Fig. 5.9, the impact of no-show rate on the optimal number of walk-in admissions is shown. As we can see, for both current and future walk-in demand level, the optimal number of walk-in admissions increases with the increase of patient no-show rate. In simple words, more walk-in admissions are expected for clinics with higher patient no-show rate. In addition, the number of walk-in admissions seems to increase linearly with the increasing no-show rate, when the no-show rate is low (less than or equal to 0.4). After that the increasing rate will become slower with the further increase of the no-show rate. One possible reason for this phenomenon is that the walk-in patient demand level is not high enough to supply the clinics' need when the no-show rate is high. Although clinics have the capacity to admit more walk-in patients, there are no more walk-in patient arrivals. Hence, when the no-show rate approaches to 1, the number of walk-in admissions approaches to the actually number of walk-in arrivals.

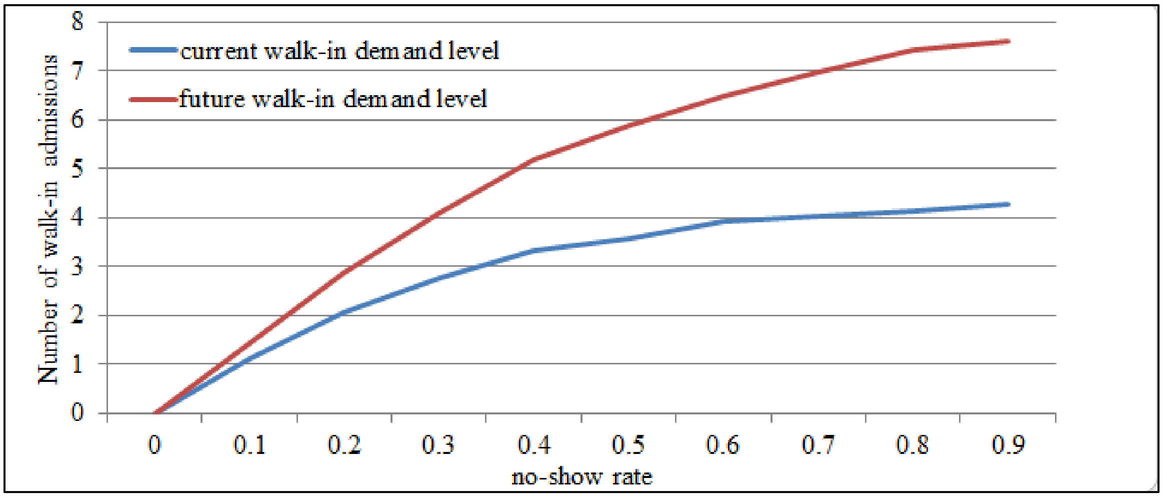

Fig. 5.9: Effect of no-show rate on the number of walk-in admissions

In Fig. 5.10, we compare the optimal objective values achieved under walk-in admission with that achieved under no walk-in admission. Clearly, for both current walk-in demand level and future walk-in demand level, a better (smaller) objective function value is obtained, i.e., walk-in admission helps to reduce the clinic cost in terms of patient waiting, provider idle and provider overtime. Note that the gap between blue line and green line is the reduced cost through walk-in admission under the current walk-in demand level, while the gap between red line and green line is the reduced cost through walk-in admission under the future walk-in demand level. Clearly, the further walk-in demand level, which is higher, leads to more significant cost reduction. In addition, the cost reduction through walk-in patient admission is also more significant for higher patient no-show rate.

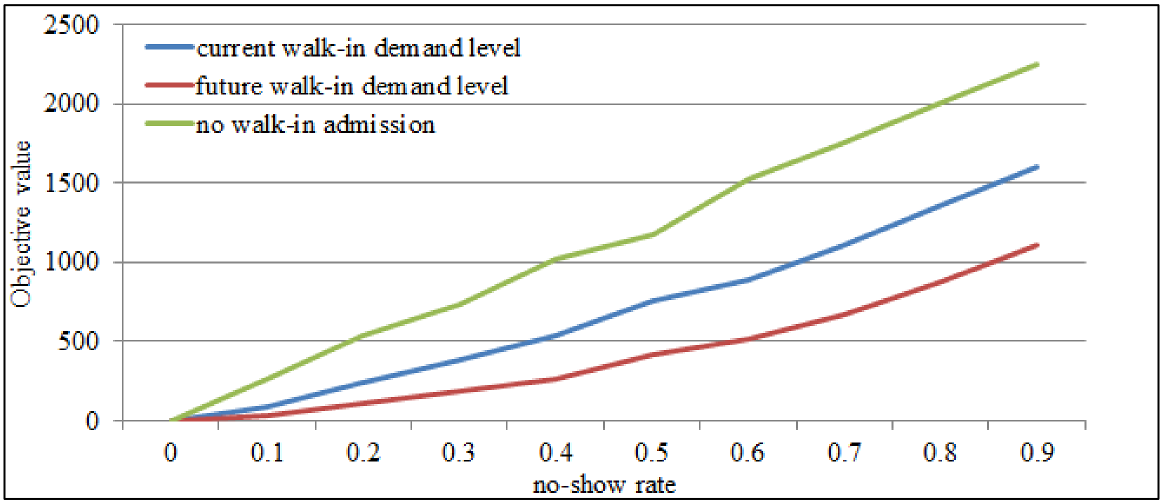

Fig. 5.10: Effect of no-show rate on the objective function value

5.4.3. Heuristic rule

Although the above discussed model can efficiently solve the small size walk-in patient admission problem (2-provider system), it cannot solve the large size problem (6-provider system) in the timely manner. With this consideration, two heuristic rules, which are shown in the following, are added to the model to improve the efficiency for solving the large size problem.

Heuristic rule 1: if a walk-in patient is admitted to the *i*th slot of provider *j*, then the provider should first see the first patient booked in the *i*th slot, after that the provider should see the second patient booked in the *i*th slot, at last the provider should see the walk-in patient.

Heuristic rule 2: all patients scheduled in or admitted to the *i*th slot of provider j should been seen earlier than any patients scheduled in or admitted to the *i*th slot of the same provider.

In order to enable these two heuristic rules in the model, a new second stage variable "I_{kij}^{assign}" need to be introduced. $I_{kij}^{assign} = 1$, if and only if the *k*th walk-in patient is admitted to the *i*th slot of provider *j*. With this variable, the relationship of service start time and end time among patient can be simplified and presented in the following:

$$t_{ij}^{S1}(\omega) + M \cdot \left(1 - I_{k(i-1)j}^{assign}(\omega)\right) \geq t_k^{Ew}(\omega), \forall i, j, k, i > 1. \tag{5.45}$$

$$t_{ij}^{S2}(\omega) \geq t_{ij}^{E1}(\omega), \forall i, j. \tag{5.46}$$

$$t_{ij}^{S1}(\omega) \geq t_{(i-1)j}^{E2}(\omega), \forall i, j, k, i > 1. \tag{5.47}$$

$$t_k^{Sw}(\omega) + M \cdot \left(1 - I_{kij}^{assign}(\omega)\right) \geq t_{ij}^{E2}(\omega), \forall i, j, k, i > 1. \tag{5.48}$$

With the incorporation with these two heuristic rules, we apply the model again to a case, in which all parameters are the same as the base case, except for the number of provider have been changed to six. It takes around 1minutes to solve model for 100 scenarios. The achieved minimum objective function value equal to 2823 minutes (the cost are measured in weighted sum of patient waiting time, provider idle time and provider over time), with the average patient time per session, average provider idle time and average provider over time equal to 225 minutes, 447 minutes, and 4 minutes, respectively. The average walk-in arrival is 8, while the average number of admitted walk-in patients is 4.

6. CONCLUSIONS AND FUTURE RESEARCH

6.1. Conclusions

In this research, we discuss two scheduling strategies (i.e., overbooking and admitting walk-in patients), which can reduce the adverse effect caused by the well-known patient non-attendance problem, if they are correctly implemented. Three novel mathematical models are developed for finding the best overbooking strategy and the optimized walk-in patient admission policy. To be more specific, a two-stage stochastic programming model is developed to solve the overbooking optimization problem with consideration of patients' choice; an MDP model is developed to find the (heuristic) optimal walk-in patient admission policy for single provider system; another two-stage stochastic programming model is developed to optimize the real time walk-in admission decisions. Besides the development of these models, the novel solution approaches are also proposed to fulfill the requirement (solution efficiency and accuracy) raised by the problems.

Some highlights of this study include: 1) Patients' choice on providers are properly captured by my models; 2) There are no constrains on patient service time distribution and patient arrival pattern for the overbooking optimization model and the two-stage walk-in patient admission optimization model (this allows the models to be applied to a wide range of clinics with different service time distribution and patient arrival pattern); 3) The cooperation schema, which allows providers to see each other's patients, is considered in overbooking optimization model. 4) The adjusted SAA algorithm is developed for solving two-stage walk-in patient admission optimization model, which is proved to be quite efficiency and effective. 5) Some interesting managerial insights are found through our carefully designed numerical cases. In the following, we summarize the findings from each of the three models.

The overbooking optimization model provides the optimized overbooking strategy given the characteristic of a clinic. By applying the overbooking optimization model to our carefully designed numerical cases, which consider different patient non-attendance rates, different patient arrival patterns, different service time distributions and different patient income levels, it is found that:

1) The optimal number of overbooked slot has a positive linear relationship with the patient non-attendance rate.

2) The patients who are booked in the overbooked slots have significantly higher waiting time than patients who are booked in the single-booked slots.

3) The overbooking strategy are only effective for clinics with high patient non-attendance rates (>20% in our case). Applying overbooking strategy to clinics with low patient non-attendance rates only increases patient waiting time.

4) The cooperation schema among providers can further reduce the clinic cost (weighted sum of patient waiting time, provider idle time and provider overtime) as compare to the non-cooperation schema, i.e. there is benefit for clinics to allow providers seeing each other's patients.

The MDP model solves the walk-in patient admission optimization problem in the single-provider clinic. The properties of the MDP model are derived to discover the optimal walk-in patient admission policies under most states. Meanwhile, heuristic admission policies are proposed for other possible states, and then are compared over 36 scenarios representative of the possibilities, which considered four arrival patterns of patients with appointments, three arrival rates of patients without appointments, and three overbooking policies. The experimental results demonstrate that:

1) Admitting all walk-in patients is a simple and good rule in clinics with walk-in patient arrival rates not greater than 20% of service rate.

2) In clinics with walk-in patient arrival rates greater than 20% of service rate, walk-in patients should be admitted when the total remaining slots are more than the expected number of patients needed to be seen.

The two-stage stochastic mixed-integer programming model provides the optimal walk-in admission decision upon the arrival of walk-in patients. In the study, we integrate heuristic rules and the adjusted SAA method to overcome the major challenge of this model, which is identified as the solution efficiency (the admission decision needs to be made in real time). After applying the model to our carefully designed numerical cases, it is found that:

1) In general, our model can be solved within 1 minute, although the computational time could vary from case to case.

2) The computational time of the proposed solution approach only increase linearly with the increase of problem size (number of providers in the clinic and number of appointment slots in a clinic session.

3) The walk-in patients in general spend more time waiting in the clinics as compare to the patients with appointments.

4) As compare to the overbooking strategy, the walk-in patient admission strategy is effective even for clinic with low patient non-attendance rate.

5) Given the same clinic characteristic (i.e. patient non-attendance rate, service time distribution, patient arrive pattern, and cost coefficient), the optimal walk-in admission strategy always lead to a lower cost as compare to the optimal overbooking strategy. As a result, we

consider the walk-in patient admission as a more superior strategy for mitigating the negative effect of patient non-attendance as compare to the overbooking strategy.

6.2. Future research

Although we modeled the patients’ choice in the overbooking and walk-in patient admission optimization problems, the impact of patients’ choice have not been revealed from this study. Given various patients’ preference, the optimal overbooking and walk-in patient admission strategy could change substantially. There is no double that the clinic cost and patient satisfaction can be improved by incorporating patients’ preference into the scheduling practice. Hence, one future direction is to investigate how the patients’ choice could affect the overbooking and walk-in admission decisions by running large number of numerical cases with varying patient’s preference and applying data mining tools.

In the MDP model, it is assumed that a provider only sees his/her own patients and new patients. Thus the walk-in patient admission to one provider is independent of the walk-in patient admission to other providers. As a result, the walk-in patient admission policy could be determined for each provider. However, some clinics group several providers as a provider team to increase scheduling flexibility. The next step of research could be extending the proposed MDP model to optimize the admission policy of pooled walk-in patients for a provider team. Also, the proposed MDP model is based on the assumption of constant consultation time per patient. In our future study, the impact of the variation in patient consultation time on the performance of walk-in patient admission policies will be investigated. In addition, heuristic walk-in patient admission rules were compared under the assumption that the waiting cost per time unit for walk-in patients is much lower than that for patients with appointments. In the future, we will investigate the heuristic admission rules in clinics in which walk-in patients are considered as part of major customers and their satisfaction is as important as that of patients with appointments.

Besides this, another interesting future direction is to further improve the solution approach for the proposed models in this study, as current solutions are only satisfactory but not perfect. For example, the current solution for the MDP model can only provide heuristic optimal rules for some the states. For another example, the current solution approach for the two-stage stochastic optimization models needs the incorporation of heuristic rules, which hinder the research of global optimal solution, to solve large size problem efficiently. Hence, with the increase of problem size there is an urgent demand for a more power solution approach. One potential way is to apply the decomposition methodologies such as the benders' decomposition and Wols Dantzig–Wolfe decomposition.

At last, it will be interesting to extend the current problem, which is based on the single clinic setting, to the clinic network setting. The complexity of the new problem grows in terms of not only the problem size, but also the new features, such as patient transportation between clinics and information exchange between clinics. For example, the clinics in the network could cooperate by referring walk-in patients among them to achieve the maximal welfare of the entire network. On the other hand, the clinics could also compete with each other to achieve the maximal profit of each single clinic. To make the problem more complex, a clinic could cooperate with part of the clinics in the network while compete with the rest of clinics in the same network. For this new walk-in patient admission problem, the model developed in this study is not sufficient to capture the new considerations. Given the development of the clinic networks, it is in urgent needs that other methodologies such as patient transportation model and game theory to be incorporated for solving the walk-in admission problem under the networks environments.

REFERENCES

Ashton, R., Hague, L., Brandreth, M., Worthington, D., & Cropper, S. (2005). A simulation-based study of a NHS walk-in centre. Journal of the Operational Research Society, 56 (2), 153-161.

Al-Shammari, S.A. (1992). Failures to keep primary care appointments in Saudi Arabia. Family Practice Research, 12(2), 171–176.

Alexopoulos, C., Goldsman, D., Fontanesi, J., Kopald, D., & Wilson, J. R. (2008). Modeling patient arrivals in community clinics. Omega, 36(1), 33-43.

Agency for Healthcare Research and Quality (AHRQ). (2007). Primary care doctors account for nearly half of physician visits but less than one-third of expenses. AHRQ News and Numbers. http://archive.ahrq.gov/news/nn/nn042507.htm

Parmessar, A. (2010) "Optimizing appointment driven systems via IPA with applications to health care systems". Available at: http://www.math.vu.nl/~sbhulai/theses/werkstuk-parmessar.pdf

White, M.B., & Pike, M.C.. (1964). Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. Medical Care, 2, 133-145.

Bureau of Labor Statistics (BLS). (2011). Occupational Outlook Handbook, 2010-11 Edition. http://www.bls.gov/oco/ocos074.htm#earnings

Bureau of Labor Statistics. (2013). May 2012 National Occupational Employment and Wage Estimates United States. Retrieved at http://www.bls.gov/oes/current/oes_nat.htm

BLS. (2013). Employment, Hours, and Earnings from the Current Employment Statistics survey (National). Retrieved at http://data.bls.gov/timeseries/CES6562000001?data_tool=XGtable

Birge, J.R., & Louveaux, F. (2011) Introduction to Stochastic Programming, 2nd Edition. Springer, New York, NY

Bundy, D.G., Randolph, G.D., Murray, M., Anderson, J., & Margolis, P.A. (2005). Open access in primary care: results of a North Carolina pilot project. Pediatrics, 116(1), 82–87.

Balasubramanian, H., Muriel, L., & Wang, L. (2012). The impact of provider flexibility and capacity allocation on the performance of primary care practices. Flexible Services and Manufacturing Journal, 24(4), 422-447.

Bean, A., & Talaga, J. (1992). Appointment breaking: causes and solutions. Journal of Health Care Marketing, 12, 14-25.

Bean, A.G., & Talaga, J. (1995). Predicting appointment breaking. Journal of Health Care Marketing, 15 (1), 29-34.

Buckley, O., Ward, E., Ryan, A., Colin, W., Snow, A., & Torreggiani, W.C. (2009). European obesity and the radiology department. What can we do to help?. European radiology, 19(2), 298-309.

Bailey, N.T. (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. Journal of the Royal Statistical Society. Series B (Methodological), 185-199.

Bosch, P.M.V., & Dietz, D.C. (2000). Minimizing expected waiting in a medical appointment system. IIE Transactions, 32(9), 841-848.

Bosch, P.M.V., & Dietz, D.C. (2001). Scheduling and sequencing arrivals to an appointment system. Journal of Service Research, 4(1), 15-25.

Crismani, C., & Galletly, C. (2011). 'Walk-ins': Developing a nursing role to manage unscheduled presentations to a community mental health clinic. Contemporary nurse, 39(1), 12-19.

Chmiel, C., Huber, C.A., Rosemann, T., Zoller, M., Eichler, K., Sidler, P., & Senn, O. (2011). Walk-ins seeking treatment at an emergency department or general practitioner out-of-hours service: a cross-sectional comparison. BMC health services research, 11(1), 94.

Campbell, K. (2011). Waste not, Want not: Curtailing healthcare costs in the US today. Available from http://www.whvheart.com/waste-not-want-not-curtailing-healthcare-costs-in-the-us-today/

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. Productionand Operations Management, 12, 519-549.

Cohen, A.D., Dreiher, J., Vardy, D.A., & Weitzman, D. (2008). Nonattendance in a dermatology clinic–a large sample analysis. Journal of the European Academy of Dermatology and Venereology, 22(10), 1178-1183.

Cameron, S., Sadler, L., & Lawson, B. (2010). Adoption of open-access scheduling in an academic family practice. Canadian Family Physician, 56(9), 906-911.

Cayirli, T., Yang, K.K., & Quek, S.A. (2012). A universal appointment rule in the presence of no-shows and walk-ins. Production and Operations Management, 21(4), 682–697.

Cayirli, T., Veral, E., & Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. Health Care Management Science, 9(1), 47-58.

Ciechanowski, P., Russo, J., Katon, W., Simon, G., Ludman, E., Von Korff, M., & Lin, E. (2006). Where is the patient? The association of psychosocial factors and missed primary care appointments in patients with diabetes. General hospital psychiatry, 28(1), 9-17.

Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., DeLaurentis, P.C., & Sands, L. (2010). Using no-show modeling to improve clinic performance.Health informatics journal, 16(4), 246-259.

Dreiher, J., Goldbart, A., Hershkovich, J., Vardy, D.A., & Cohen, A.D. (2008). Factors associated with non-attendance at pediatric allergy clinics. Pediatric Allergy and Immunology, 19(6), 559-563.

Dobson, G., Hasija, S., & Pinker, E.J. (2011). Reserving capacity for urgent patients in primary care. Production and Operations Management, 20(3), 456–473.

Fetter, R.B., & Thompson, J.D. (1966). Patients' waiting time and doctors' idle time in the outpatient setting. Health Services Research, 1(1), 66.

Fontanesi, J., Alexopoulos, C., Goldsman, D., DeGuire, M., Kopald, D., Holcomb, K., & Sawyer, M.H. (2001). Non-punctual patients: planning for variability in appointment arrival times. The Journal of medical practice management: MPM, 18(1), 14-18.

González-Arévalo, A., Gómez-Arnau, J.I., DelaCruz, F.J., Marzal, J.M., Ramírez, S., Corral, E.M., & García-del-Valle, S. (2009). Causes for cancellation of elective surgical procedures in a Spanish general hospital.Anaesthesia, 64(5), 487-493.

Garuda, S., Javalgi, R., & Talluri, V. (1998). Tackling no-show behavior: a market-driven approach. Health Marketing Quarterly, 15, 25-44.

Geraghty, M., Glynn, F., Amin, M., & Kinsella, J. (2007). Patient mobile telephone text reminder: a novel way to reduce non-attendance at the ENT out-patient clinic. The Journal of Laryngology and Otology, 122(3), 296–298.

Goldman, L., Freidin, R., Cook, E.F., Eigner, J., & Grich, P. (1982). A multivariate approach to the prediction of no-show behavior in a primary care center. Archives of Internal Medicine, 142(3), 563-567.

Gupta, D., & Denton, B.T. (2008). Health care appointment systems: Challenges and opportunities.IIE Transactions, 40, 800-819.

Guy, R., Hocking, J., Wand, H., Stott, S., Ali, H., & Kaldor, J. (2012). How effective are short message service reminders at increasing clinic attendance? a meta-analysis and systematic review. Health Services Research, 47(2), 614–632.

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

Giachetti, R.E. (2008). A simulation study of interventions to reduce appointment lead-time and patient no-show rate. In Simulation Conference, 2008. WSC 2008. Winter (pp. 1463-1468). IEEE.

Gupta, D., Potthoff, S., Blowers, D., & Corlett, J. (2006). Performance metrics for advanced access. Journal of Healthcare Management, 51(4), 246-259.

Gail, G.W. (2007). You can make walk-ins work. Medical Economics, 84(16), 44-6, 49-50. Retrieved from http://search.proquest.com/docview/227819816?accountid=6766

George, A., & Rubin, G. (2003). Non-attendance in general practice: A systematic review and its implications for access to primary health care. Family Practice, 20, 178-184.

Hashim, M. J., Franks, P., & Fiscella, K. (2001). Effectiveness of telephone reminders in improving rate of appointments kept at an outpatient clinic: a randomized controlled trial. The Journal of the American Board of Family Practice, 14(3), 193-196.

Henry, S.R., Goetz, M.B., & Asch, S.M. (2012). The effect of automated telephone appointment reminders on HIV primary care no-shows by veterans. Journal of the Association of Nurses in AIDS Care, 23(5), 409-418.

Hermoni, D., Mankuta, D., & Reis, S. (1990). Failure to keep appointments at a community health centre. Analysis of causes. Scandinavian Journal Primary Health Care, 8(2), 107–111.

Herriott, S. (1999). Reducing delays and waiting times with open-office scheduling. Family Practice Management, 6, 38-43.

Ho, C. J., & Lau, H.S. (1992). Minimizing total cost in scheduling outpatient appointments. Management Science, 38(12), 1750-1764.

Ho, C.J., & Lau, H.S. (1999). Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. European Journal of Operational Research, 112(3), 542-553.

Howard, M., Goertzen, J., Kaczorowski, J., Hutchison, B., Morris, K., Thabane, L., & Papaioannou, A. (2008). Emergency department and walk-in clinic use in models of primary care practice with different after-hours accessibility in Ontario. Health Policy, 4(1), 73-88.

Huang, Y., & Zuniga, P. (2012). Dynamic overbooking scheduling system to improve patient access. Journal of the Operational Research Society, 63(6), 810-820.

Huarng, F. (2003). The Impact of Walk-in Ratio on Outpatient Department. In2003 International Conference on Technology and Management.

Krueger, A.B. 2009. A hidden cost of health care: Patient time. New York Times Economix Blog. http://economix.blogs.nytimes.com/2009/02/09/a-hidden-cost-of-health-care-patient-time/

Kros, J., Dellana, S., & West, D. (2009). Overbooking increases patient access at East Carolina University's student health services clinic. Interfaces, 39(3), 271-287.

Kerssens, J.J., Groenewegen, P.P., Sixma, H.J., Boerma, W.G.W., & Van Der Eijk, I. (2004). Comparison of patient evaluations of health care quality in relation to WHO measures of achievement in 12 European countries. Bulletin of the World Health Organization, 82 (2), 106-114.

Klassen, K.J., & Rohleder, T.R. (1996). Scheduling outpatient appointments in a dynamic environment. Journal of Operations Management, 14(2), 83-101.

Kaandrop, G.C., & Koole, G. (2007). Optimal outpatient appointment scheduling. Health Care Management Science, 10(3), 217-229.

Kennedy, J.G., & Hsu, J.T. (2003). Implementation of an open access scheduling system in a residency training program. Family Medicine, 35(9), 666-670.

Kopach, R., DeLaurentis, P. C., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Willis, D., et al. (2007). Effects of clinical characteristics on successful open access scheduling. Health Care Management Science, 10(2), 111-124.

Kim, S., & Giachetti, R.E. (2006). A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 36(6), 1211-1219.

Kleywegt, A.J., Shapiro, A., & Homem-de-Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. SIAM Journal on Optimization, 12(2), 479-502.

Lacy, N., Paulman, A., Reuter, M., & Lovejoy, B. (2004). Why we don't come: Patient perceptions on no-shows. Annals of Family Medicine, 2, 541-545.

Lehmann, T.N.O., Aebi, A., Lehmann, D., Balandraux Olivet, M., & Stalder, H. (2007). Missed appointments at a Swiss university outpatient clinic. Public health, 121(10), 790-799.

Liu, N. (2013). Optimal Choice for Appointment Scheduling Window under Patient No-show Behavior, Working Paper, Columbia University

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

LaGanga, L.R., & Lawrence, S.R. (2008). Yield management in health care clinics with walk-in traffic. Proceedings of Tradition and Innovation in Operations Management, 15th International Annual EurOMA Conference, University of Groningen, the Netherlands – June 15-17, 2008.

LaGanga, L.R., & Lawrence, S.R. (2009). Comparing walk-in, open access, and traditional appointment scheduling in outpatient health care clinics. Productions and Operations Management Society Conference. Orlando, FL.

LaGanga, L.R., & Lawrence, S.R. (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. Production and Operations Management, 21(5), 874-888.

Liu, N., Ziya, S., & Kulkarni, V. (2010). Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manufacturing and Services Operations Management, 12(2), 347-365.

Lee, S., & Yih, Y. (2010). Analysis of an open access scheduling system in outpatient clinics: A simulation study. Simulation, 86(8-9), 503-518.

Lee, V.J., Earnest, A., Chen, M.I., & Krishnan, B. (2005). Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. BMC Health Services Research, 5, 51-58.

Moore, C.G., Wilson-Witherspoon, P., & Probst, J.C. (2001). Time and money: Effects of no-shows at a family practice residency clinic. Family Medicine, 33(7), 522-527.

Murray, M., & Tantau., C. (1999). Redefining open access to primary care. Manage Care Quartely, 7(3), 45-55.

Murray, M., & Berwick, D.M. (2003). Advanced access: reducing waiting and delays in primary care. Journal of the American Medical Association, 289 (8), 1035-1040.

Murray, M., & Tantau, C. (2000). Same-day appointments: exploding the access paradigm. Family Practice Management, 7, 45–50.

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

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

Mallard, S.D., Leakeas, T., Duncan, W.J., Fleenor, M.E., & Sinsky, R.J. (2004). Same-day scheduling in a public health clinic: a pilot study. Journal of Public Health Management and Practice, 10(2), 148-155.

Macharia, W.M., Leon, G., Rowe, B.H., Stephenson, B.J., & Haynes, R.B. (1992). An overview of interventions to improve compliance with appointment keeping for medical services. The Journal of the American Medical Association 267(13) 1813–1817.

Norris, J.B., Kumar, C., Chand, S., Moskowitz, H., Shade, S.A., & Willis, D.R. (2014). An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. Decision Support Systems, 57, 428-443.

O'Connor, M.E., Matthews, B.S., & Gao, D. (2006). Effect of open access scheduling on missed appointments, immunizations, and continuity of care for infant well-child care visits. Archives Pediatrics & Adolescent Medicine, 160(9), 889–893.

O'Hare, C.D., & Corlett, J. (2004). The outcomes of open-access scheduling. Family Practice Management, 11(2), 35-38.

Olson, C.L., Schumaker, H.D., & Yawn, B.P. (1994). Overweight women delay medical care. Archives of family medicine, 3(10), 888-892.

Pandhi, N., & Saultz, J.W. (2006). Patients' perceptions of interpersonal continuity of care. Journal of the American Board of Family Medicine, 19 (4), 390-397.

Parikh, A., Gupta, K., Wilson, A.C., Fields, K., Cosgrove, N.M., & Kostis, J.B. (2010). The effectiveness of outpatient appointment reminder systems in reducing no-show rates. The American journal of medicine, 123(6), 542-548.

Parente, D.H., Pinto, M,B., & Barber, J.C. (2005). A pre-post comparison of service operational efficiency and patient satisfaction under open access scheduling. Health Care Manage Review, 30(3), 220–228.

Pinto, M.B., Parente, D.H., & Barber, J.C. (2002). Selling open access health care delivery to patients and administrators: what's the hook? Health Marketing Quarterly, 19 (3), 57-69.

Patrick, J. (2012). A Markov decision model for determining optimal outpatient scheduling. Health Care Management Science, 15, 91–102.

Phatak, P., Skoretz, L., Linda, L. (2011). Improving walk-in patient flow thrugh primary care. available at: https://srd.vssc.med.va.gov/Training/FlowAcademy/Documents/Loma%20Linda%20-%20Improving%20Walk-in%20Patient%20Flow%20thru%20Primary%20Care%20-%20(Dr.Skoretz+Phatak)%20-%20Final%20Version%205-18-2011.ppt

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

Qu, X., J. Shi. (2009). Effects of two-level provider capacities on the performance of open access clinics. Health Care Management Science, 12, 99-114.

Qu, X., Rardin, R.L., & Williams, J.A.S. (2011). Single versus hybrid time horizons for open access scheduling. Computers & Industrial Engineering, 60(1), 56.

Qu, X., Rardin, R.L., & Williams, J.A.S. (2012). A mean–variance model to optimize the fixed versus open appointment percentages in open access scheduling systems. Decision Support Systems, 53, 554–564.

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

Rodriguez, H.P., Rogers, W.H., Marshall, R.E., & Safran, D.G. (2007). The effects of primary care physician visit continuity on patients' experiences with care. Journal of General Internal Medicine, 22, 787-793.

Rohleder, T.R., & Klassen, K.J. (2000). Using client-variance information to improve dynamic appointment scheduling performance. Omega, 28(3), 293-302.

Rose, K.D., Ross, J.S., & Horwitz, L.I. (2011). Advanced access scheduling outcomes: A systematic review. Archives of Internal Medicine, 171(13), 1150-1159.

Robinson, L., & Chen, R. (2010). A comparison of traditional and open access policies for appointment scheduling. Manufacturing and Services Operations Management, 12(2), 330-347.

Saultz, J.W., & Lochner, J. (2005). Interpersonal continuity of care and care outcomes: A critical review. Annals of Family Medicine, 3 (2), 159-166.

Samorani, M., & LaGanga, L.R. (2013). Outpatient appointment scheduling given individual day-dependent no-show predictions. Available from http://www.business.ualberta.ca/MicheleSamorani/~/media/business/FacultyAndStaff/AOIS/MicheleSamorani/Documents/papers/Outpatient_Appointment_Scheduling.pdf

Su, S., & Shih, C.L. (2003). Managing a mixed-registration-type appointment system in outpatient clinics. International journal of medical informatics, 70(1), 31-40.

Sussman. (2013). How walk-in clinics are changing the healthcare business. Available at: http://www.forbes.com/sites/vannale/2013/11/20/minuteclinic-coo-andrew-sussman-on-how-walk-in-clinics-are-changing-the-healthcare-business/

Schoen, C., Osborn, R., Huynh, P.T., Doty, M., Davis, K., Zapert, K., & Peugh, J. (2004). Primary care and health system performance: Adults' experiences in five countries. Health Affairs, Web Exclusives, W4, 487-503.

Shapiro, A., & Homem-de-Mello, T. (2000). On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. SIAM journal on optimization, 11(1), 70-86.

Shi, J., Peng, Y., & Erdem, E. (2014). Simulation analysis on patient visit efficiency of a typical VA primary care clinic with complex characteristics. Simulation Modelling Practice and Theory, 47, 165-181.

Vanden Bosch, P.M., Dietz, D.C., & Simeoni, J.R. (1999). Scheduling customer arrivals to a stochastic service system. Naval Research Logistics (NRL), 46(5), 549-559.

Vissers, J. (1979). Selecting a suitable appointment system in an outpatient setting. Medical Care, 1207-1220.

Woodcock, E. (2003). Mastering Patient Flow: More Ideas to Increase Efficiency and Earnings. MGMA.

Woods, R. (2010). The effectiveness of reminder phone calls on reducing no-show rates in ambulatory care. Nursing economic$, 29(5), 278-282.

WHO. (2011). World health statistics 2011. Retrieved from http://www.who.int/whosis/whostat/2011/en/index.html

Welch, J.D., & Bailey, N.T.J., 1952. Appointment systems in hospital outpatient departments. The Lancet, 259 (6718), 1105-1108.

Zeng, B., Zhao, H., & Lawley, M. (2013). The impact of overbooking on primary care patient no-show. IIE Transactions on Healthcare Systems Engineering, 3(3), 147-170.

Zeng, B., Turkcan, A., Lin, J., & Lawley, M. (2010). Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Annals of Operations Research, 178(1), 121-144.

Zacharias, C., & Pinedo, M. (2014). Appointment Scheduling with No-Shows and Overbooking. Production and Operations Management, 23(5), 788-801.

APPENDIX. PROOFS OF PROPOSITIONS

Proof of proposition 2. $\sum_{i=1}^{M} x_i^m > 0$ means that at least one elective patient is waiting in the clinic for services, while n^m>0 means at least one walk-in patient is waiting in the clinic for services. According to condition 4.15 in Proposition 1, any valid action must satisfy $d_s^m + d_w^m = 0$ or $d_s^m + d_w^m = 1$. If some patients are waiting in the clinic, the actions satisfying $d_s^m + d_w^m = 0$ lead to longer patient waiting time and may result in less walk-in patients seen and longer provider idle time and overtime compared with the actions satisfying $d_s^m + d_w^m = 1$. Therefore, for any state $\mathbf{s}^m$ satisfying $\sum_{i=1}^{M} x_i^m > 0$ or n^m>0, the optimal action $\pi^*(\mathbf{s}^t)$ must satisfy $\bar{d}_s^m + \bar{d}_w^m = 1$. ■

Proof of corollary 4. For a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m = 0$, $n^m = 0$ and $w^m = 1$, the valid action set $A(\mathbf{s}^m)$ is {(0,0,0), (1,0,0), (1,0,1)}. If the walk-in patient arriving between decision stages (m – 1) and m is admitted (i.e. $d_a^m = 1$), seeing this patient between decision stages m and (m – 1) could reduce patient waiting time and provider idle time without any loss because $\sum_{i=1}^{M} x_i^m = 0$ and $n^m = 0$. As a result, $V(\mathbf{s}^m, (1,0,1)) > V(\mathbf{s}^m, (1,0,0))$, which means that action (1,0,1) is better than action (1,0,0) for a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m = 0$, $n^m = 0$ and $w^m = 1$.

Next, we compare actions (0,0,0) and (1,0,1). From a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m = 0$, $n^m = 0$ and $w^m = 1$, the process transitions to the same state $\mathbf{s}^{m+1}$ when either (0,0,0) or (1,0,1) is taken. That means $\mathbf{s}^{m+1}(\mathbf{s}^m, (0,0,0)) = \mathbf{s}^{m+1}(\mathbf{s}^m, (1,0,1))$, where $\mathbf{s}^{m+1}(\mathbf{s}^m, \mathbf{a}^m)$ denotes the state at decision stage m+1 when an action $\mathbf{a}^m$ is taken in state $\mathbf{s}^m$. Thus $E[V^*(\mathbf{s}^{m+1}|\mathbf{s}^m,(0,0,0)] = E[V^*(\mathbf{s}^{m+1}|\mathbf{s}^m,(1,0,1)]$. According to Equation (7), we obtain $R^m(\mathbf{s}^m,(0,0,0)) = -c_i$ and $R^m(\mathbf{s}^m,(1,0,1)) = r_w$. Then according to Equation (11), $V(\mathbf{s}^m, (1,0,1)) > V(\mathbf{s}^m,(0,0,0))$ because $E[V^*(\mathbf{s}^{m+1}|\mathbf{s}^m,(0,0,0)] = E[V^*(\mathbf{s}^{m+1}|\mathbf{s}^m,(1,0,1)]$ and $R^m(\mathbf{s}^m,(1,0,1)) - R^m(\mathbf{s}^m,(0,0,0)) = r_w + c_i > 0$. Therefore, (1,0,1) is optimal for a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{M} x_i^m = 0$, $n^m = 0$ and $w^m = 1$. ■

Proof of proposition 3. For convenience, in this proof, denote $\mathbf{a}_{JKL}^{m}$ the action (*J*,*K*,*L*) at decision stage *m*. According to Proposition 2, for a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{m} x_i^m > 0$, the optimal action $\pi^*(\mathbf{s}^t)$ satisfies

$d_s^m + d_w^m = 1$. Therefore, one of actions $\mathbf{a}_{010}^{m}$, $\mathbf{a}_{001}^{m}$, $\mathbf{a}_{110}^{m}$ and $\mathbf{a}_{101}^{m}$ is optimal for such a state. Since not all four actions $\mathbf{a}_{010}^{m}$, $\mathbf{a}_{001}^{m}$, $\mathbf{a}_{110}^{m}$ and $\mathbf{a}_{101}^{m}$ are valid in any state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{m} x_i^m > 0$, we divide all states $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{m} x_i^m > 0$ into three subsets depending on whether actions $\mathbf{a}_{010}^{m}$, $\mathbf{a}_{001}^{m}$, $\mathbf{a}_{110}^{m}$ and $\mathbf{a}_{101}^{m}$ are valid in state $\mathbf{s}^m$. The three subsets are defined as

$S_{\text{I}} = \{\mathbf{s}^m \mid \mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{m} x_i^m > 0$ and $n^m + w^m = 0\}$,

$S_{\text{II}} = \{\mathbf{s}^m \mid \mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{m} x_i^m > 0$, $n^m > 0$ and $w^m \geq 0\}$,

and $S_{\text{III}} = \{\mathbf{s}^m \mid \mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{m} x_i^m > 0$, $n^m = 0$ and $w^m > 0\}$.

In a state $\mathbf{s}^m \in S_{\text{I}}$, since $n^m + w^m = 0$, actions $\mathbf{a}_{001}^{m}$ and $\mathbf{a}_{101}^{m}$ are invalid, and taking either $\mathbf{a}_{010}^{m}$ or $\mathbf{a}_{110}^{m}$ results in the identical expected total net reward in the (*M*–*m*+1) remaining stages. Thus, both $\mathbf{a}_{010}^{m}$ and $\mathbf{a}_{110}^{m}$ are optimal in a state $\mathbf{s}^m \in S_{\text{I}}$.

Actions $\mathbf{a}_{010}^{m}$, $\mathbf{a}_{110}^{m}$ and $\mathbf{a}_{101}^{m}$ are valid in state $\mathbf{s}^m \in S_{\text{III}}$, while all four actions $\mathbf{a}_{010}^{m}$, $\mathbf{a}_{001}^{m}$, $\mathbf{a}_{110}^{m}$ and $\mathbf{a}_{101}^{m}$ are valid in state $\mathbf{s}^m \in S_{\text{II}}$. First, we prove that actions $\mathbf{a}_{010}^{M}$ and $\mathbf{a}_{110}^{M}$ are optimal in any state $\mathbf{s}^M \in S_{\text{II}} \cup S_{\text{III}}$ at decision stage *M*. According to Equations 4.8, 4.9 and 4.12, the expected total net rewards of actions $\mathbf{a}_{010}^{M}$, $\mathbf{a}_{001}^{M}$, $\mathbf{a}_{110}^{M}$ and $\mathbf{a}_{101}^{M}$ in a state $\mathbf{s}^M \in S_{\text{II}}$ are

$$V(\mathbf{s}^M, \mathbf{a}_{010}^M) = V(\mathbf{s}^M, \mathbf{a}_{110}^M) = r_s \sum_{i=1}^{M} x_i^M - c_o \left(\sum_{i=1}^{M} x_i^M - 1 \right) - \frac{c_s}{2} \sum_{i=1}^{M} x_i^M \left(\sum_{i=1}^{M} x_i^M - 1 \right),$$

and $V(\mathbf{s}^M, \mathbf{a}_{001}^M) = V(\mathbf{s}^M, \mathbf{a}_{101}^M) = r_w + (r_s - c_o) \sum_{i=1}^{M} x_i^M - \frac{c_s}{2} \sum_{i=1}^{M} x_i^M \left(\sum_{i=1}^{M} x_i^M + 1 \right).$

Since $c_o > r_w$ and $c_s \geq 0$, we obtain $V(\mathbf{s}^M, \mathbf{a}_{010}^M) = V(\mathbf{s}^M, \mathbf{a}_{110}^M) > V(\mathbf{s}^M, \mathbf{a}_{001}^M) = V(\mathbf{s}^M, \mathbf{a}_{101}^M)$. Thus, actions $\mathbf{a}_{010}^{M}$ and $\mathbf{a}_{110}^{M}$ are optimal in a state $\mathbf{s}^M \in S_{\text{II}}$. In a state $\mathbf{s}^M \in S_{\text{III}}$, action $\mathbf{a}_{001}^{M}$ is invalid. Similarly, according to

Equations 4.8, 4.9 and 4.12, we could obtain $V(\mathbf{s}^M,\mathbf{a}_{010}^M)=V(\mathbf{s}^M,\mathbf{a}_{110}^M)>V(\mathbf{s}^M,\mathbf{a}_{101}^M)$. Thus, actions $\mathbf{a}_{010}^M$ and $\mathbf{a}_{110}^M$ are optimal in a state $\mathbf{s}^M \in S_{\text{III}}$, too. Therefore, we conclude that actions $\mathbf{a}_{010}^M$ and $\mathbf{a}_{110}^M$ are optimal in any state $\mathbf{s}^M \in S$ satisfying $\sum_{i=1}^m s_{ai}^m > 0$. Since all walk-in patients who could not be served in the current clinic session are dismissed, action $\mathbf{a}_{010}^M$ is more practical.

Next, we prove that one of actions $\mathbf{a}_{010}^m$ and $\mathbf{a}_{110}^m$ is an optimal action in any state $\mathbf{s}^m \in S_{\text{II}}$ at decision stage $m < M$. We compare $\mathbf{a}_{010}^m$ to $\mathbf{a}_{001}^m$, and $\mathbf{a}_{110}^m$ to $\mathbf{a}_{101}^m$ in a state $\mathbf{s}^m \in S_{\text{II}}$ for $m < M$. Let $\mathbf{a}_{C10}^m$ and $\mathbf{a}_{C01}^m$ denote a couple of actions in which $d_a{}^m$ takes the same value C. According to Equation 4.7, the difference of the immediate net reward between actions $\mathbf{a}_{C10}^m$ and $\mathbf{a}_{C01}^m$ in the state $\mathbf{s}^m \in S_{\text{II}}$ is

$$R^m(\mathbf{s}^m,\mathbf{a}_{C10}^m)-R^m(\mathbf{s}^m,\mathbf{a}_{C01}^m)=\left[r_s-c_s\left(\sum_{i=1}^m x_i^m-1\right)-c_w\left(n^m+w^mC\right)\right]-\left[r_w-c_s\sum_{i=1}^m x_i^m-c_w\left(n^m+w^mC-1\right)\right]$$

$$=(r_s-r_w)+(c_s-c_w)\,. \tag{A1}$$

When action $\mathbf{a}_{C01}^m$ is taken in the state $\mathbf{s}^m \in S_{\text{II}}$, the process transitions to the next state $\mathbf{s}^{m+1}(\mathbf{s}^m,\mathbf{a}_{C01}^m)$ satisfying $\sum_{i=1}^{m+1} x_i^{m+1}=\sum_{i=1}^m x_i^m+x_{m+1}^m+\widetilde{X}_{m+1}^m>0$, $n^{m+1}=n^m+w^mC-1\geq 0$ and $w^{m+1}=\widetilde{W}^m+\sum_{i=1}^m \widetilde{X}_i^m \geq 0$. On the other hand, when action $\mathbf{a}_{C10}^m$ is taken in the state $\mathbf{s}^m \in S_{\text{II}}$, the process transitions to the next state $\mathbf{s}^{m+1}(\mathbf{s}^m,\mathbf{a}_{C10}^m)$ satisfying $\sum_{i=1}^{m+1} x_i^{m+1}=\sum_{i=1}^m x_i^m+x_{m+1}^m+\widetilde{X}_{m+1}^m-1\geq 0$, $n^{m+1}=n^m+w^mC>0$ and $w^{m+1}=\widetilde{W}^m+\sum_{i=1}^m \widetilde{X}_i^m \geq 0$.

Assume $\pi^*(\mathbf{s}^{m'} \mid \mathbf{s}^m,\ \mathbf{a}_{C01}^m)$ be the first optimal action satisfying $\bar{d}_s^{m'}=1$ and $\bar{d}_w^{m'}=0$ after taking a series of optimal actions $\pi^*(\mathbf{s}^{m''} \mid \mathbf{s}^m,\ \mathbf{a}_{C01}^m)$ for $m < m'' < m'$ from the state $\mathbf{s}^{m+1}(\mathbf{s}^m,\mathbf{a}_{C01}^m)$. In other word, all optimal actions $\pi^*(\mathbf{s}^{m''} \mid \mathbf{s}^m,\ \mathbf{a}_{C01}^m)$ for $m < m'' < m'$ satisfy $\bar{d}_s^{m''}=0$ and $\bar{d}_w^{m''}=1$. Let $\bar{C}^{m''}$ be the value of $\bar{d}_a^{m''}$ in

$\pi^*(\mathbf{s}^{m''} \mid \mathbf{s}^m,\ \mathbf{a}_{C01}^m)$ for $m < m'' \leq m'$. Since $\mathbf{s}^{m+1}(\mathbf{s}^m,\mathbf{a}_{C01}^m)$ satisfies $\sum_{i=1}^{m+1} x_i^{m+1}>0$ and $\bar{d}_s^{m''}=0$ in $\pi^*(\mathbf{s}^{m''} \mid \mathbf{s}^m,\ \mathbf{a}_{C01}^m)$ for all $m < m'' < m'$, state $\mathbf{s}^{m'}$ reached after taking a series of optimal actions $\pi^*(\mathbf{s}^{m''} \mid \mathbf{s}^m,\ \mathbf{a}_{C01}^m)$ for

$m < m'' < m'$ satisfies $\sum_{i=1}^{m'} x_i^{m'} > 0$, which implies $m' \leq M$ because actions $\mathbf{a}_{010}^{M}$ and $\mathbf{a}_{110}^{M}$ are optimal in any state $\mathbf{s}^M \in S$ satisfying $\sum_{i=1}^{M} x_i^{M} > 0$. If $m' < M$, the process transitions to the same state $\mathbf{s}^{m'+1}$ when either the series of actions $\{\mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01}, \mathbf{a}^{m'}_{\bar{C}^{m'}01}\}$ or the series $\{\mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01}, \mathbf{a}^{m'}_{\bar{C}^{m'}10}\}$ is taken starting from the state $\mathbf{s}^m \in S_{\mathrm{II}}$. Thus,

$$E[V^*(\mathbf{s}^{m'+1} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01}, \mathbf{a}^{m'}_{\bar{C}^{m'}01})] = E[V^*(\mathbf{s}^{m'+1} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01}, \mathbf{a}^{m'}_{\bar{C}^{m'}10})]. \tag{A2}$$

If $m' = M$, the number of remaining elective patients at the end of the session will be same when either series of actions is taken starting from the state $\mathbf{s}^m \in S_{\mathrm{II}}$. That means

$$R^{M+1}(\mathbf{s}^M, \mathbf{a}^{M}_{\bar{C}^{M}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{M-1}_{\bar{C}^{(M-1)}01}) = R^{M+1}(\mathbf{s}^M, \mathbf{a}^{M}_{\bar{C}^{M}10} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{M-1}_{\bar{C}^{(M-1)}01}). \tag{A3}$$

Since the action series $\{\mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01}, \mathbf{a}^{m'}_{\bar{C}^{m'}10}\}$ is the optimal action at decision stages (m+1) to m' for the process starting the state $\mathbf{s}^{m+1}(\mathbf{s}^m, \mathbf{a}^{m}_{C01})$, the maximum expected total net reward in the (M–m) remaining stages is

$$E[V^*(\mathbf{s}^{m+1} | \mathbf{s}^m, \mathbf{a}^{m}_{C01})]$$

$$= \begin{cases} \displaystyle\sum_{m''=m+1}^{m'-1} E[R^{m''}(\mathbf{s}^{m''}, \mathbf{a}^{m''}_{\bar{C}^{m''}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m''-1}_{\bar{C}^{(m''-1)}01})] + E[R^{m'}(\mathbf{s}^{m'}, \mathbf{a}^{m'}_{\bar{C}^{m'}10} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01})] \\ \qquad + E[V^*(\mathbf{s}^{m'+1} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01}, \mathbf{a}^{m'}_{\bar{C}^{m'}10})], \qquad \text{for } m' < M \\ \displaystyle\sum_{m''=m+1}^{m'-1} E[R^{m''}(\mathbf{s}^{m''}, \mathbf{a}^{m''}_{\bar{C}^{m''}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m''-1}_{\bar{C}^{(m''-1)}01})] + E[R^{m'}(\mathbf{s}^{m'}, \mathbf{a}^{m'}_{\bar{C}^{m'}10} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01})] \\ \qquad + R^{M+1}(\mathbf{s}^M, \mathbf{a}^{M}_{\bar{C}^{M}10} | \mathbf{s}^m, \mathbf{a}^{m}_{C01}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{M-1}_{\bar{C}^{(M-1)}01}), \qquad \text{for } m' = M \end{cases} \tag{A4}$$

according to Equation (11). Meanwhile, the expected total net reward in the (M–m) remaining stages when taking the action series $\{\mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01}, \mathbf{a}^{m'}_{\bar{C}^{m'}01}\}$ from the state $\mathbf{s}^{m+1}(\mathbf{s}^m, \mathbf{a}^{m}_{C10})$, is

$$E[V(\mathbf{s}^{m+1}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'}_{\bar{C}^{m'}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C10},)]$$

$$= \begin{cases} \displaystyle\sum_{m''=m+1}^{m'-1} E[R^{m''}(\mathbf{s}^{m''}, \mathbf{a}^{m''}_{\bar{C}^{m''}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m''-1}_{\bar{C}^{(m''-1)}01})] + E[R^{m'}(\mathbf{s}^{m'}, \mathbf{a}^{m'}_{\bar{C}^{m'}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01})] \\ \qquad + E[V^*(\mathbf{s}^{m'+1} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'}_{\bar{C}^{m'}01})], \qquad \text{for } m' < M \\ \displaystyle\sum_{m''=m+1}^{m'-1} E[R^{m''}(\mathbf{s}^{m''}, \mathbf{a}^{m''}_{\bar{C}^{m''}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m''-1}_{\bar{C}^{(m''-1)}01})] + E[R^{m'}(\mathbf{s}^{m'}, \mathbf{a}^{m'}_{\bar{C}^{m'}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{m'-1}_{\bar{C}^{(m'-1)}01})] \\ \qquad + R^{M+1}(\mathbf{s}^M, \mathbf{a}^{M}_{\bar{C}^{M}01} | \mathbf{s}^m, \mathbf{a}^{m}_{C10}, \mathbf{a}^{m+1}_{\bar{C}^{(m+1)}01}, \ldots \mathbf{a}^{M-1}_{\bar{C}^{(M-1)}01}), \qquad \text{for } m' = M \end{cases}. \tag{A5}$$

According to Equations (A2)–(A5), we derive

$$E[V(\mathbf{s}^{m+1},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m'}_{\overline{C}^{m'}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10},)]-E[V^{*}(\mathbf{s}^{m+1}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01})]$$

$$=\sum_{m''=m+1}^{m'-1}\left\{E[R^{m''}(\mathbf{s}^{m''},\mathbf{a}^{m''}_{\overline{C}^{m''}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m''-1}_{\overline{C}^{(m''-1)}01})]-E[R^{m''}(\mathbf{s}^{m''},\mathbf{a}^{m''}_{\overline{C}^{m''}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m''-1}_{\overline{C}^{(m''-1)}01})]\right\}$$

$$+\left\{E[R^{m'}(\mathbf{s}^{m'},\mathbf{a}^{m'}_{\overline{C}^{m'}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m'-1}_{\overline{C}^{(m'-1)}01})]-E[R^{m'}(\mathbf{s}^{m'},\mathbf{a}^{m'}_{\overline{C}^{m'}10}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m'-1}_{\overline{C}^{(m'-1)}01})]\right\}. \quad \text{(A6)}$$

According to Equations (1)–(7), the difference of the expected net reward obtained at decision stage *m″* for *m* < *m″* < *m′* is

$$E[R^{m''}(\mathbf{s}^{m''},\mathbf{a}^{m''}_{\overline{C}^{m''}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m''-1}_{\overline{C}^{(m''-1)}01})]-E[R^{m''}(\mathbf{s}^{m''},\mathbf{a}^{m''}_{\overline{C}^{m''}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m''-1}_{\overline{C}^{(m''-1)}01})]$$

$$=E\left[r_w-c_s\left(\sum_{i=1}^{m''}x_i^m+\sum_{i=m+1}^{m''}\sum_{k=m}^{i-1}\widetilde{X}_i^k-1\right)-c_w\left(n^m+w^mC+\sum_{k=m+1}^{m''}w^k\overline{C}^k-(m''-m)\right)\right]$$

$$-E\left[r_w-c_s\left(\sum_{i=1}^{m''}x_i^m+\sum_{i=m+1}^{m''}\sum_{k=m}^{i-1}\widetilde{X}_i^k\right)-c_w\left(n^m+w^mC+\sum_{k=m+1}^{m''}w^k\overline{C}^k-(m''-m)-1\right)\right]=E\left[c_s-c_w\right]=c_s-c_w\,, \quad \text{(A7)}$$

and the difference of the expected net reward obtained at decision stage *m′* is

$$E[R^{m'}(\mathbf{s}^{m'},\mathbf{a}^{m'}_{\overline{C}^{m'}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m'-1}_{\overline{C}^{(m'-1)}01})]-E[R^{m'}(\mathbf{s}^{m'},\mathbf{a}^{m'}_{\overline{C}^{m'}10}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m'-1}_{\overline{C}^{(m'-1)}01})]$$

$$=E\left[r_w-c_s\left(\sum_{i=1}^{m'}x_i^m+\sum_{i=m+1}^{m'}\sum_{k=m}^{i-1}\widetilde{X}_i^k-1\right)-c_w\left(n^m+w^mC+\sum_{k=m+1}^{m'}w^k\overline{C}^k-(m'-m)\right)\right]$$

$$-E\left[r_s-c_s\left(\sum_{i=1}^{m'}x_i^m+\sum_{i=m+1}^{m'}\sum_{k=m}^{i-1}\widetilde{X}_i^k-1\right)-c_w\left(n^m+w^mC+\sum_{k=m+1}^{m'}w^k\overline{C}^k-(m'-m)\right)\right]=E\left[r_w-r_s\right]=r_w-r_s\,. \quad \text{(A8)}$$

Substituting Equations (A7) and (A8) into Equation (A6), we obtain

$$E[V(\mathbf{s}^{m+1},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m'}_{\overline{C}^{m'}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10},)]-E[V^{*}(\mathbf{s}^{m+1}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01})]=(m'-m-1)(c_s-c_w)+(r_w-r_s)\,.$$

According to Equation (10), we know $E[V^{*}(\mathbf{s}^{m+1}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10})]\geq E[V(\mathbf{s}^{m+1},\mathbf{a}^{m+1}_{\overline{C}^{(m+1)}01},\ldots,\mathbf{a}^{m'}_{\overline{C}^{m'}01}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10},)]$. Thus

$$E[V^{*}(\mathbf{s}^{t+1}|\mathbf{s}^{t},\mathbf{a}^{t}_{C10})]-E[V^{*}(\mathbf{s}^{t+1}|\mathbf{s}^{t},\mathbf{a}^{t}_{C01})]\geq(m'-m-1)(c_s-c_w)+(r_w-r_s)\,. \quad \text{(A9)}$$

According to Equations (11) and (A1) and Inequality (A9), we obtain

$$V(\mathbf{s}^{m},\mathbf{a}^{m}_{C10})-V(\mathbf{s}^{m},\mathbf{a}^{m}_{C01})=\left\{R^{m}(\mathbf{s}^{m},\mathbf{a}^{m}_{C10})+E[V^{*}(\mathbf{s}^{m+1}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10})]\right\}-\left\{R^{m}(\mathbf{s}^{m},\mathbf{a}^{m}_{C01})+E[V^{*}(\mathbf{s}^{m+1}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01})]\right\}$$

$$=\left\{R^{m}(\mathbf{s}^{m},\mathbf{a}^{m}_{C10})-R^{m}(\mathbf{s}^{m},\mathbf{a}^{m}_{C01})]\right\}+\left\{E[V^{*}(\mathbf{s}^{m+1}|\mathbf{s}^{m},\mathbf{a}^{m}_{C10})-E[V^{*}(\mathbf{s}^{m+1}|\mathbf{s}^{m},\mathbf{a}^{m}_{C01})]\right\}\geq(m'-m)(c_s-c_w)\,.$$

Since $c_s > c_w$ and $m' > m$, we know $V(\mathbf{s}^m, \mathbf{a}^m_{C10}) > V(\mathbf{s}^m, \mathbf{a}^m_{C01})$. Therefore, one of actions $\mathbf{a}^m_{010}$ and $\mathbf{a}^m_{110}$ is an optimal action in any state $\mathbf{s}^m \in S_{\mathrm{II}}$ for $m < M$.

Similar to the proof of $V(\mathbf{s}^m, \mathbf{a}^m_{C10}) > V(\mathbf{s}^m, \mathbf{a}^m_{C01})$ in a state $\mathbf{s}^m \in S_{\mathrm{II}}$ for $m < M$, we could obtain $V(\mathbf{s}^m, \mathbf{a}^m_{110}) > V(\mathbf{s}^m, \mathbf{a}^m_{101})$ in a state $\mathbf{s}^m \in S_{\mathrm{III}}$ for $m < M$. Since action $\mathbf{a}^m_{001}$ is invalid in a state $\mathbf{s}^m \in S_{\mathrm{III}}$, either $\mathbf{a}^m_{010}$ or $\mathbf{a}^m_{110}$ is optimal in any state $\mathbf{s}^m \in S_{\mathrm{III}}$ at decision stage $m < M$.

In summary, for a state $\mathbf{s}^m \in S$ satisfying $\sum_{i=1}^{m} x_i^m > 0$, the optimal action $\pi^*(\mathbf{s}^m)$ satisfies $\bar{d}_s^m = 1$ and $\bar{d}_w^m = 0$.