APPLYING SIMULATION AND GENETIC ALGORITHM FOR PATIENT APPOINTMENT SCHEDULING OPTIMIZATION

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

By

Yidong Peng

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Program:
Industrial Engineering and Management

October 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

APPLYING SIMULATION AND GENETIC ALGORITHM FOR PATIENT APPOINTMENT SCHEDULING OPTIMIZATION

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

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Jing Shi
Chair

Kambiz Farahmand

Jun Zhang

Joseph Szmerekovsky

Approved:

11/06/2013
Date

Canan Bilen-Green
Department Chair

ABSTRACT

In this study, we discuss the implementation of integrated simulation and genetic algorithm for patient scheduling optimization under two different settings, namely the "traditional" scheduling system and the "open access" scheduling system. Under the "traditional" setting, we propose a two-phase approach for designing a weekly scheduling template for outpatient clinics providing multiple types of services. Our results demonstrate that the two-phase approach can efficiently find the promising weekly appointment scheduling templates for outpatient clinics. Under the "open access" setting, we propose a discrete event simulation and genetic algorithm (DES-GA) approach to find the heuristic optimal scheduling template for the clinic allowing both open access and walk-in patients. The solution provides scheduling templates consisting of not only the optimal number of reservations for open access appointments and walk-ins, but also the optimized allocation of these reserved slots, by minimizing the average cost per admission of open access or walk-in patient.

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. Jun Zhang, and Dr. Joseph Szmerekovsky, for offering time and individual expertise to support and guide me though this exciting and painstaking journey.

In addition, I would like to thank Dr. Xiuli Qu and Dr. Nan Kong, for their diligent guide and support for my research.

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

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Overview of Health Care Delivery System

The US health care industry, which began as a volunteer and charitable system, has now developed into the largest business in the country. It is reported that the United States spent 17.9% ($ 2.7 trillion) of GDP on health care in 2011, which is more than any other countries in the world (WHO, 2011). In the meantime, the health care spending in the U.S. is growing at 3.7% - 4.1% during 2009-2011 (Hartman et al., 2013), which is higher than the growth rate of the nation income. According to the statistics (BLS, 2013), the health care industry is employing more than 17 million workers with a projection of 5.7 million new jobs by the year 2012.

It is argued that the U.S. has the most formidable medical force and most modern medical technology in the world (Francois Sainfort, 2004). Despite the large amount of money, advanced technology and labor force invested in the health care industry, the health care delivery system in the U.S. has long been criticized for inadequate quality and lack of efficiency. Back to 2000, the World health Organization (WHO, 2000) ranked the U.S. health care delivery system 37th among 191 countries worldwide. In addition, the extremely high costs related to health care services put tremendous burden on the patients. For many years, the entire health care industry has been under pressure to reduce care cost while improving care quality (Institution of Medicine, 2001).

According to the report "Healthcare Delivery System in United States" conducted by Francois (2004), the healthcare industry is a very large, complex and inefficient industry. In the meanwhile, the Institute of Medicine (2011) has directed the inadequate quality of care in the U.S. to four underlying reasons: 1) the growing complexity of science and technology, 2) the increase in chronic condition, 3) the poorly organized health care delivery system, and 4) the constraints on exploiting the revolution in information technology. Due to the characteristic of

the current U.S. health care delivery system, Francois (2004) indicates the unlimited potential of industrial and system engineering to contribute and make significant change to the system. The capability of applying quantitative tools and model based analysis from engineering prospective has been highlighted for design of new system with great complexity and re-design of current health care delivery system. It is further indicated that the healthcare has recently been viewed as an "industry", because of the rapid change to the sector, which is brought in by the increasing managed care. As such, it is concluded (Francois, 2004) that reengineering the delivery of healthcare services through innovative development, application, and use of proven and novel operations research and management sciences methods, theories, and tools coupled with modern and novel information and communication technology solutions can lead to tremendous cost savings and improved access to healthcare services, as well as improved quality of life for all citizens.

1.2. The Outpatient Scheduling System

As a key component of the entire healthcare delivery system, the outpatient clinics are experiencing long patient waiting time, which has long been identified as the source of inefficiency and rising cost in clinics. Many outpatient clinics have reported that operational excellence in appointment scheduling helps smoothen the patient flow, reduce patient waiting in clinics, and eventually leads to improved quality and reduced cost of care (Cayirli and Veral, 2003; Armstrong et al., 2005; LaGanga and Lawrence, 2007; Gupta and Denton, 2008). In general, the objectives of appointment scheduling system have been defined as (Liu et al., 2010): (i) provide better service to customers by assigning them a short time window, during which they are guaranteed with a service; (ii) protect the system against demands fluctuation from time to time, which can lead to possible lacking of facility utilization at some time, or overloading at

other time. To be more specific, the objective of outpatient scheduling is to find an appointment system for which a particular measure of performance is optimized in a clinical environment (i.e., an application of resource scheduling under uncertainty), as defined by Cayirli and Veral (2003). As for the means of appointment booking process, typically, patients make phone calls or visit clinic in person to make appointments. Clinics schedule patients in available slots upon request. It is not until recently that, with the development of information technology, the online appointment scheduling systems have become available through a secured access to the websites of some clinics.

It is well known that the outpatient scheduling systems have gone through many changes. In this study, we categorize the outpatient scheduling systems into two classes, namely, "traditional" and "open access" appointment scheduling 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, cancellation rate, and late arrival rate are likely to increase due to the long waiting list for appointments. Numerous studies indicate that patient no-shows, cancellations and late arrivals increase volatility to the standard clinic process, which would in turn increases the healthcare expenditure and decrease clinic efficiency and patient accessibility. For this reason, in 1990s open access scheduling (or advanced access scheduling) was proposed

to mitigate the negative effects of patient no-shows, cancellations, and late arrivals, promote 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. As we can see, the key difference between traditional scheduling and open access scheduling system is that no slots are reserved for the same-day appointment of non-acute illness under the setting of traditional scheduling.

1.3. Problem Statement

In this study, the problem we are facing is to find the best scheduling template for a given clinic setting. A weekly/daily scheduling template refers to a set of rules that specifies the number of appointments and the appointment times that are reserved for each type of services in each clinic session during a week/day. In practice, when a patient requests an appointment, a scheduler in the clinic first compares the current schedule of each week/day to the weekly/daily scheduling template to find available appointment times for him/her, and then schedules an appointment for him/her at one of the available times that best matches his/her preference. The best scheduling template is defined as the one that minimizes the weighted sum of patient waiting time, provider idle time and provider over time during a clinic session under the setting of traditional scheduling system. However, this definition is modified into minimizing the cost per same day appointment during a day under the "open access" setting, where the cost is measured as weighted sum of patient waiting time, provider idle time and provider overtime during the day. The patient waiting time is measured as the time difference between the actual appointment start time and scheduled appointment start time. In case that the actual appointment starts before the scheduled appointment start time, the waiting time will be defined as zero. The

provider overtime is the time that a 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.

As introduced in Section 1.2, we classify the outpatient scheduling systems into "traditional" and "open access" scheduling systems. Hence, we have two separate tasks. In Figure 1.1, the scheme of the problem studied in this thesis is clearly illustrated. Task 1 is to develop the best scheduling template for a "traditional" scheduling system with various service categories, while Task 2 is to find the best scheduling template under the setting of "open access" scheduling with unique features. For the "traditional" scheduling system, we aim to develop a weekly scheduling template, which can balance the workload among sessions during the week and minimize weighted total of patient waiting time, provider idle time and provider overtime of each session. Note that, different types of appointment might be provided among sessions. However, only one category of appointment is offered in each session. The existing studies in literature have not been able to address the scheduling template issue for clinics with different appointment services. As such, the success of Task 1 fills an important research gap and contributes to the clinic scheduling research. For the "open access" scheduling system, we need to find a daily scheduling template that minimizes the cost per same day appointment, since the nature of open access scheduling requires a daily template rather than weekly template. It is worthwhile to mention that the walk-in patients are also considered under the "open access" setting. In addition, the double booking strategy, patient no-shows and patient cancellations are considered under both settings. The contribution of Task 2 is also clear – it for the first time addresses the optimization problem of scheduling template for open access clinics that admits walk-in patients. For a detailed description of these two tasks, we refer reader to Sections 3.1 and

4.1, where the problem descriptions and formulations are given to Task 1 and Task 2, respectively.

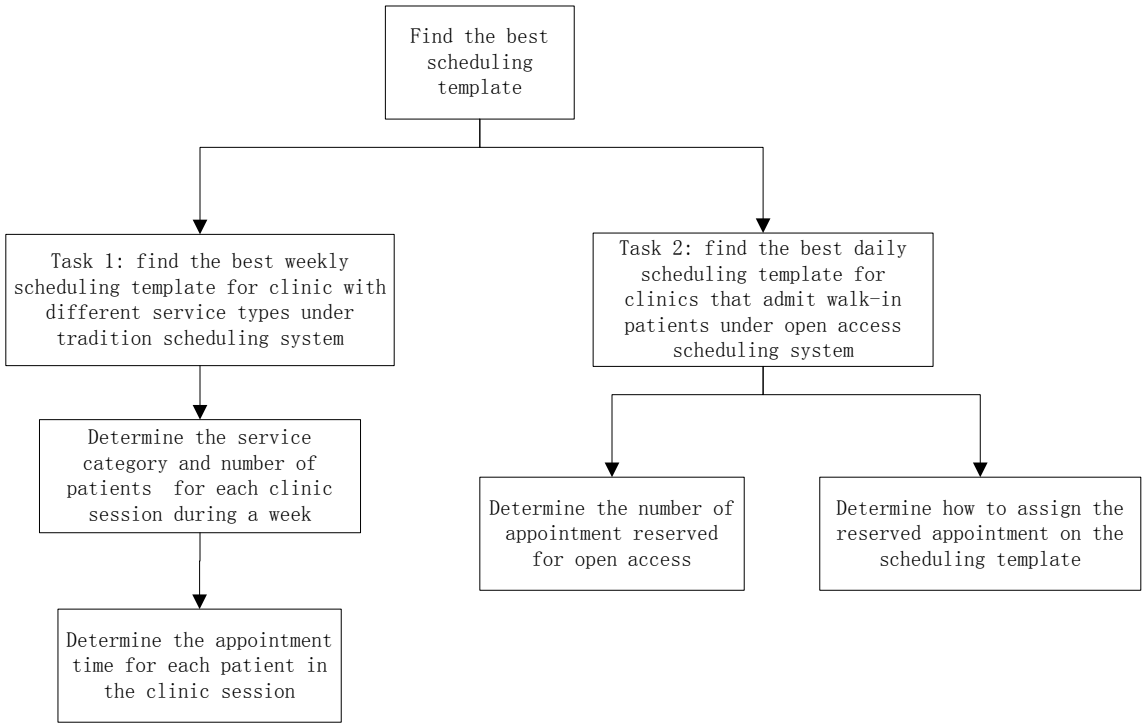

Fig. 1.1: The scheme of the studied problem

1.4. Organization of the Thesis

The remainder of the thesis is organized as follows. In Chapter 2, we provide a literature review of the existing studies on the patient scheduling problem and the relevant methodologies of this study. In Chapter 3, we propose a two-phase approach, which includes an integer linear programming model to balance the workload among sessions, a two-stage SMIP model to locate the best scheduling template under the "traditional" setting, and a solution approach for the SMIP model. A case study is also provided in Chapter 3, in order to demonstrate the performance of the proposed two-phase approach. In Chapter 4, the DES-GA solution method is developed for optimizing the scheduling template under "open access" setting. In addition, a case study is also presented in Chapter 4, where numerical examples are summarized for sensitivity

analysis. Insightful discussion is provided based on the results. At last, we draw the conclusion in Chapter 5.

2. LITERATURE REVIEW

The appointment scheduling problem in healthcare settings has attracted the interest of many researchers and practitioners over the past 60 years (Cayirli and Veral, 2003). There has been a rich body of operations research (OR) literature on outpatient appointment scheduling. For a comprehensive review of outpatient appointment scheduling, we refer to Cayirli and Veral (2003), where taxonomy of methodologies developed in previous literature is provided, and general problem formulations and modeling considerations are presented. For more literature on patient scheduling, we refer the readers to the review paper written by Gupta and Denton (2008), where the practical issues related to appointment scheduling and the art of modeling and optimization are discussed.

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 H-S, 1992, Ho and H-S, 1996, 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 identical Erlang distribution and punctual patient arrivals. Vanden Bosch and Dietz (2000 and 2001) extend the model by considering general phase-type distributed service times as well as patient no-shows. Kaandrop 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.

In first task of our study, we formulate a scheduling optimization problem under "traditional" appointment scheduling setting with a Stochastic Mixed Integer Programming (SMIP) model. Stochastic programming has been applied to the appointment scheduling problems in healthcare settings, including operating rooms (Denton and Gupta, 2003; Denton et al., 2007; Batun et al., 2011) and outpatient clinics (Robinson and Chen, 2003; Begen and Queyranne, 2011; Begen et al., 2012). However, to the best of our knowledge, it has not been applied in multiple-provider outpatient appointment scheduling. The first task of our study differs in three ways from the previous appointment scheduling studies applying stochastic

programming in the literature. First, unlike the studies (Denton and Gupta, 2003; Denton et al., 2007; Batun et al., 2011; Robinson and Chen, 2003; Begen and Queyranne, 2011; Begen et al., 2012) assuming that the sequence(s) of all or subsets of surgeries/jobs is given or pre-determined by heuristic rules, in our study, the service sequence in each clinic session is determined by solving the proposed SMIP. Secondly, unlike the multiple-server scheduling problems studied by Batun et al. (2011), and Robinson and Chen (2003) which determine the continuous starting times of surgeries in each OR, in the first half of our scheduling problem, the services of each type are assigned with discrete appointment times, but not assigned to individual physicians (i.e. servers). In the problem, the patients with appointments in a clinic session will wait in a priority queue formed with their appointment times until being seen by one of the physicians working in the session. This is similar to a queuing system with multiple servers and a single priority waiting line. Thirdly, we consider discrete appointment times and continuous random service times in the problem. In the previous studies, continuous appointment times and continuous random service times are considered by Denton and Gupta (2003), Denton et al. (2007), Batun et al. (2011), and Robinson and Chen (2003), and discrete appointment times and discrete random service times are considered by Begen and Queyranne (2011), and Begen et al. (2012).

In the second task of our study, we develop the appointment scheduling optimization problem under the setting of "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.

In the meantime, there are also scheduling optimization studies considering walk-in patients. Kim et al. (2006) develop a stochastic mathematical overbooking model, which considers the probability distribution of walk-in patients during the process of determining the optimal number of appointments to be scheduled in order to maximize the expected total profits in diverse healthcare environment. Oh and Chow (2011) conduct a discrete event simulation to evaluate the impact of different patient appointment arrangement and patient-doctor allocation strategies on the patient cycle time of clinic visit. An exclusive allocation strategy is developed for walk-in patient seeking consultation for non-chronic conditions. Cayirili et al. (2012) propose a "Dome" appointment rule, which is formulated as a function of the clinic specified parameter, a "Dome" pattern parameter and the mean and variance of consulting time, for clinics with no-shows and walk-in patients. In addition, they propose a model to adjust the mean and variance of consulting time considering the effect of no-shows and walk-in patients, while the simulation and nonlinear regression models are applied to estimate the clinic specified parameter.

It can be seen that a number of qualitative and quantitative studies have been performed in the area of open access scheduling, but there is still a lack of investigations on how to allocate the reserved appointment slots in a scheduling template. In addition, no study has addressed the allocation of appointment slots complicated by admitting walk-in patients to open access clinics to the best of our knowledge. As such, to bridge this gap, we develop the mathematical programming model and a Discrete Event Simulation and Genetic Algorithm (DES-GA)

approach in this study to find the heuristic optimal scheduling template for open access clinics that admit walk-in patients.

Throughout our study, the hybrid simulation and genetic algorithm are implemented as the general solution methodology. In literature, the combination of simulation and genetic algorithm (GA) has been widely used for solving optimization problems in different fields. A large number of studies apply simulation/GA approach to solve job shop scheduling problems (Nicoara et al., 2011; Gholami and Zandieh, 2009; Jeong et al. 2006). Many other studies can be found that apply simulation and GA to solve maintenance scheduling problems (Manbachi et al., 2011; Cheu et al., 2004; Ma et al., 2004). More applications of simulation/GA approach can be found in other areas. Amiri et al. (2012) use this approach to optimize buffer allocation in unreliable production lines; Huang et al. (2012) adopt this approach to develop an optimum design for the arrangement of moisture-buffering materials in order to achieve a reliable indoor humidity environment; Lin et al. (2012) also employ this approach to optimize the scheduling of dispatching earthmoving trucks. In terms of simulation/GA applications in health care industry, only a few studies can be found. Yeh and Lin (2007) use the approach to improve the quality of service at a hospital emergency department by appropriately adjusting nurses' schedules without hiring additional staff. Gul et al. (2011) propose a simulation and bi-criteria GA model to find the best scheduling heuristic for an outpatient procedure center, where the simulation is used to evaluate the performance of 12 different sequencing and patient appointment time-setting heuristics, while the genetic algorithm is used to determine if better solutions can be obtained for this single day scheduling problem. Nevertheless, each of these simulation/GA studies is problem-specific, and no general model, which could be easily adopted and customized for solving all problems, is available.

3. APPOINTMENT SCHEDULING OPTIMIZATOIN FOR TRADITIONAL SCHEDULING SYSTEM

3.1. Problem Description and Formulation

In many outpatient specialty clinics, the overall appointment scheduling process involves assignment decisions at two phases. In the first phase, a master scheduling template is developed to allocate healthcare service capacity to clinic sessions in response to projected patient appointment requests. Then in the second phase, appointments are scheduled to various time slots in each clinic session given the service type specified by the master scheduling template. When making first-phase decisions, various service types are clustered based on medical specialists and equipment they require. It is often the case that not any arbitrary pair of services can be scheduled in the same clinic session, as equipment changeover time needed between the two services is prohibitive or different medical specialties required cannot be covered by the same set of health professionals.

In our study, we develop a two-phase approach for multi-category appointment scheduling in an outpatient specialty clinic to incorporate this feature. Our approach closely matches the real-world appointment scheduling process. We a priori cluster those services that do not require substantial changeover time into the same category. Hence, we require that services scheduled in the same clinic session belong to the same category. Although this restriction is a common feature in almost all outpatient clinics that provide multiple categories of services, to the best of our knowledge, multi-category appointment scheduling with incorporation of this feature has not been fully investigated by the healthcare management science research community. It is worth noting that we are aware of the fact that it is likely to achieve more reliable solutions when taking an approach that integrates category assignment and

slot assignment. However, such an integrated approach would lead to a stochastic optimization problem of much larger scale and the solution may not seem interactive in real-world practice. Furthermore, we incorporate uncertainty arising in patient no-show and service time. High patient no-show and large variation in service time are two common features in many outpatient clinics (Cayirli and Veral, 2003, Gupta and Denton, 2008).

As mentioned above, this women's clinic provides multiple types of services such as pre-pregnancy checkup, routine prenatal visitation, obstetric examination for high-risk pregnancy, and routine gynecology examination for new and follow-up patients. In this women's clinic, these different services are clustered into categories according to the requirements on service equipments. For example, for the obstetric examination of a high-risk pregnant patient, more advanced testing procedures and devices (e.g., transvaginal ultrasound) are needed to identify risks for the patient and her fetus at early stage of the pregnancy, whereas standard procedures and devices are sufficient for low-risk patients. Significant amount of time incurred by changing medical equipments makes it undesirable to schedule these two types of services in the same clinic session. In order to reduce the changeover time, this women's clinic, like many other outpatient clinics around the nation, simply does not schedule a session with different types of services between which significant changeover time is required. This restriction, however, presents challenges in appointment scheduling. In addition, the studied women's clinic, like many small-scale clinics that primarily serve the nation's low-income populations, has serious issues with patient no-shows and large variation of service times. For certain service type, nearly half of the appointments are missed and the service time varies from a few minutes to more than an hour. The aforementioned restrictions and characteristics motivate us to study the multi-

category appointment scheduling problem with considerations of patient no-shows and large variation of service times.

On the other hand, appointments in the women's clinic are scheduled for all available physicians instead of individual physicians. As a result, the patients scheduled in a session will be seen in a sequence based on their appointment times, and may be seen by any physician working in the session. Since multiple physicians are available in any clinic session, more than one appointment could be scheduled at the same appointment time.

The aforementioned issues and characteristics in the women's clinic motivate us to study this multi-category outpatient appointment scheduling problem, for which we propose a two-phase mathematical model. In Phase I, each clinic session is assigned to one of the given service categories and specified with the number of appointments for each service type belonging to the assigned category. The goal of the Phase I assignment is to balance provider workload among sessions. In Phase II, an appointment time is determined for each appointment that is reserved for each service type in each session, with the objective of minimizing patient waiting time, provider idle time, and provider overtime. Since appointment times in outpatient clinics are only multiples of 5 minutes, only a finite number of discrete time points could be possible appointment times. To deal with discrete appointment times, each clinic session is divided into multiple time slots of predetermined equal length, which correspond to possible appointment times. In Phase II, the appointments reserved for each service type in each session are allocated into time slots in the session. Allocating an appointment into a time slot only indicates that its appointment time is the beginning of the slot, and the service time for the appointment could last over multiple time slots. To summarize, our two-phase approach is intended to identify promising weekly scheduling templates with incorporation of the uncertainty arising in patient no-shows and service times,

subject to the restriction that all appointments in each session must belong to the same service category.

3.1.1. MILP model for decision making in Phase I

In Phase I, we assign a service category to each clinic session and determine how many appointments should be scheduled in the clinic session for each service type that belongs to the assigned service category. Given the demand forecast, the average service time, and the anticipated no-show rate of each service type within a prespecified decision period (e.g., typically one week), we consider a deterministic static assignment problem between a given set of appointments and a given set of clinic sessions in the period with the restriction that each clinic session can only contain appointments for services in the same category. We formulate this assignment problem with an integer program.

Let M be the set of service types, C be the set of service categories, and S be the set of clinic sessions during the decision period. For each service category $c \in C$, we denote $M(c) \subseteq M$ to be the subset of service types that belong to category c. We use decision variable x_{ms} to represent the number of appointments to be scheduled for service of type $m \in M$ in clinic session $s \in S$. We also use binary decision variable w_{cs} to indicate whether service category $c \in C$ is assigned to clinic session $s \in S$. In addition, parameters $\bar{r}_m$, γ_m and D_m denote the average service time, the anticipated no-show rate, and forecasted demand, respectively, for service type $m \in M$. The notation used in the integer program is summarized in Table 3.1.

In Phase I, a healthcare manager intends to balance the workload among clinic sessions during each decision period because the balanced provider capacity utilization can reduce the differences of total patient waiting time and provider idle time among clinic sessions. We use the

expected total service time in a clinic session as a proxy for the expected workload in a session and present a mixed-integer program (P1) as follows.

Table 3.1: Notation used in Phase I

Index	
c	Service category index
m	Service type index
s	Clinic session index
Set	
C	Index set of service categories
M	Index set of service types
$M(c)$	Index subset of service types belonging to service category c
S	Index set of clinic sessions in a decision period
Parameter	
D_m	Demand forecast on type m service in a decision period
γ_m	Patient no-show rate of service type m
$\bar{r}_m$	Average service time of type m service
Decision Variable	
x_{ms}	Number of appointments to be scheduled for type m service in clinic session s
w_{cs}	Indicator whether service category c is assigned to clinic session s

$$\text{(P1)} \quad \min \quad \frac{1}{2}\sum_{\substack{s_1,s_2\in S \\ s_1\neq s_2}}\left|\sum_{m\in M}(1-\gamma_m)\bar{r}_m(x_{ms_1}-x_{ms_2})\right| \tag{3.1}$$

$$\text{s.t.} \quad \sum_{\forall s\in S} x_{ms} = D_m, \text{ for } m\in M; \tag{3.2}$$

$$\sum_{\forall c\in C} w_{cs} = 1, \text{ for } s\in S; \tag{3.3}$$

$$x_{ms} \le D_m w_{cs}, \text{ for } c\in C, \forall m\in M(c), \forall s\in S; \tag{3.4}$$

$$x_{ms} \in Z_+, \text{ for } m\in M, \forall s\in S; \tag{3.5}$$

$$w_{cs} \in \{0,1\}, \text{ for } c\in C, \forall s\in S. \tag{3.6}$$

In (P1), the objective function 3.1 minimizes the aggregate absolute difference on the expected total service time over all pairs of clinic sessions. Constraints 3.2 enforce that for each service type, all requests for such service must be scheduled within the decision period. Constraints 3.3 and 3.4 ensure that in each clinic session, only appointments for services in the

same category can be scheduled. Note that in (P1) nonlinearity only appears in the objective function due to the consideration of the absolute difference. To linearize the objective function, we introduce two groups of nonnegative auxiliary decision variables of each pair of sessions (s1, s2)$\in$S, s1 $\neq$ s2 ,

$$u_{(s_1,s_2)} = \begin{cases} \sum_{m\in M}(1-\gamma_m)\overline{R}_m(x_{ms_1} - x_{ms_2}), & \text{if } \sum_{m\in M}(1-\gamma_m)\overline{R}_m(x_{ms_1} - x_{ms_2}) > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3.7}$$

and $$v_{(s_1,s_2)} = \begin{cases} -\sum_{m\in M}(1-\gamma_m)\overline{R}_m(x_{ms_1} - x_{ms_2}), & \text{if } \sum_{m\in M}(1-\gamma_m)\overline{R}_m(x_{ms_1} - x_{ms_2}) < 0 \\ 0, & \text{otherwise} \end{cases}. \tag{3.8}$$

Then the objective function is rewritten in the linear form as

$$\min \quad \frac{1}{2}\sum_{\substack{s_1,s_2\in S \\ s_1\neq s_2}} \left(u_{(s_1,s_2)} + v_{(s_1,s_2)}\right)$$

and additional sets of constraints are added to (P1) as:

$$u_{(s_1,s_2)} \geq \sum_{m\in M}(1-\gamma_m)\overline{R}_m(x_{ms_1} - x_{ms_2}), \quad \text{for } \forall s_1, s_2 \in S, s_1 \neq s_2; \tag{3.9}$$

$$v_{(s_1,s_2)} \geq -\sum_{m\in M}(1-\gamma_m)\overline{R}_m(x_{ms_1} - x_{ms_2}), \quad \text{for } \forall s_1, s_2 \in S, s_1 \neq s_2; \tag{3.10}$$

$$u_{(s_1,s_2)},\ v_{(s_1,s_2)} \geq 0, \quad \text{for } \forall s_1, s_2 \in S, s_1 \neq s_2. \tag{3.11}$$

3.1.2. SMIP model for decision making in Phase II

In Phase I, a mixed-integer program (P1) is solved to obtain x^*_{ms}, the optimal number of appointments to be scheduled for each type of service in each session. The solution x^*_{ms} is used to specify the number of appointments of service type *m* assigned to time slots in clinic session *s* in Phase II. When scheduling appointments into slots, we take into account the potential no-shows and service time uncertainty pertaining to each appointment. Hence, we develop a two-stage stochastic mixed-integer programming (SMIP) model. The objective of the problem is to

improve the expected operational performance in terms of three commonly used measures, namely patient waiting time, provider idle time and provider overtime (Cayirli and Veral, 2003). For an introduction to stochastic integer programming, we refer to Birge and Louveaux (2011). To develop the SMIP model, we make the following assumptions.

1) Each session is evenly divided into time slots; multiple appointments can be scheduled in one time slot.
2) Once an appointment is made, it cannot be modified unless it is canceled by the patient.
3) All patients must be seen once they arrive for their appointments. Each patient must be seen before any other patients with later appointments than hers.
4) Patients arrive punctually for their appointments; otherwise, they are no-shows. Providers arrive punctually at the beginning of the session.
5) For each service type, random service times are independent and identically distributed.
6) Patient no-shows are independent of each other with known probability for each type of service.

Many of the above assumption are commonly made in the outpatient appointment scheduling literature (Cayirli and Veral, 2003, Gupta and Denton, 2008). The others can be justified by the real-world setting investigated in this paper. For example, it is well observed that in many outpatient specialty clinics, patients tend to arrive earlier with the willingness of waiting. When a patient does not arrive in time, it typically implies that she would not come for the appointment. Literature indicates that more than 75% of the patients arrive early while fewer than 10% arrive late for longer than 5 minutes (O'Keefe ,1985).

We introduce the additional notation used in Phase II, as summarized in Table 3.2. In Phase I, we specify the set of service types to be scheduled in a clinic session and the number of appointments to be scheduled for each type of service. These specifications become the input to the second-phase problem. Let M'_s and D'_{ms} denote the set of service types assigned to session s and the number of appointments to be scheduled for type m service in session s, respectively. Since we consider appointment assignment to time slots in individual sessions in Phase II, we suppress the clinic session index s in the remainder of the section for notational simplicity. In addition, we denote N and T_Δ to be the number of appointment slots in a clinic session and the length of each slot. Then we define $T := NT_\Delta$ to be the total length of a clinic session.

In the developed two-stage stochastic program, we define y_{mn} to be the first-stage decision variables, which represent the number of appointments scheduled for type $m \in M'$ service in time slot $n = 1,\ldots,N$. We define $b_{jk}(\omega)$ and $z^{jk}_{imn}(\omega)$ to be the second-stage decision variables, which indicate whether j is greater than or equal to the total number of patients seen by provider $k = 1,\ldots, K$ in the clinic session, and whether the j^{th} patient seen by provider k has the i^{th} appointment for type $m \in M'$ service in the n^{th} slot, respectively. We model each scenario, denoted by ω, as a "snapshot" of the patient arrivals of each service type and the service time for each appointment scheduled in the clinic session. We let Ω denote the set of all scenarios. To generate the scenarios, we take advantage of the assumptions that patient no-shows are independent of each other. Let Q_{mn} be the random variable representing the number of patients who arrive for their appointments for type $m \in M'$ service scheduled in the n^{th} slot. Then the assumption of independent no-shows implies that Q_{mn} follows a binomial distribution with parameters y_{mn} and $1 - \gamma_m$, i.e., $Q_{mn} \sim \text{Binomial}(y_{mn}, 1 - \gamma_m)$. For each scenario $\omega \in \Omega$, we define

$q_{mn}(\omega)$ to be the number of patients who actually arrive for their appointments under the scenario and thus $J(\omega) := \sum_{m \in M} \sum_{n=1}^{N} q_{mn}(\omega)$ denotes the total number of patients seen in the clinic session.

Table 3.2: Notation used in Phase II

Index	
i	Appointment index
j	Patient index
k	Provider index
m	Service type index
n	Appointment slot index
ω	Scenario index
Set	
M'	Index set of service types assigned to the considered clinic session
Ω	Index set of scenarios
Parameter	
D'_m	Number of appointments scheduled for type m service in the considered clinic session
K	Total number of providers available in the considered clinic session
N	Number of appointment slots in the considered clinic session
T	Total length of the considered clinic session
T_Δ	Length of an appointment slot
γ_m	No-show rate of a patient with an appointment for type m service
η_I	Weighting coefficient of provider idle time
η_O	Weighting coefficient of provider overtime
η_W	Weighting coefficient of patient waiting time
Random Variable	
$q_{mn}(\omega)$	Number of patients who actually show up for their appointments scheduled for type m service in the n^{th} slot under scenario ω
$J(\omega)$	Total number of patients who actually show up for their appointments scheduled under scenario ω
$r_{imn}(\omega)$	Service time for the i^{th} appointment scheduled for type m service in the n^{th} slot under scenario ω
Decision Variable	
y_{mn}	Number of appointments scheduled for type m service in the n^{th} slot (first-stage decision variable)
$b_{jk}(\omega)$	$= \begin{cases} 0, & \text{if } j \text{ is greater than the total number of patients seen by provider } k \text{ under scewnario } \omega \\ 1, & \text{otherwise} \end{cases}$
$z_{imn}^{jk}(\omega)$	$= \begin{cases} 1, & \text{if the } j^{th} \text{ patient seen by provider } k \text{ has the } i^{th} \text{ appointment for type } m \text{ service in the } n^{th} \text{ slot under scenario } \omega \\ 0, & \text{otherwise} \end{cases}$
$t_{jk}^{A}(\omega)$	Appointment time of the j^{th} patient seen by provider k under scenario ω
$t_{jk}^{S}(\omega)$	Actual starting time of the service for the j^{th} patient provided by provider k under scenario ω
$t_{jk}^{E}(\omega)$	Actual completion time of the service for the j^{th} patient provided by provider k under scenario ω
$\tau_{jk}(\omega)$	Actual service time of the j^{th} patient seen by provider k under scenario ω
$t_{jk}^{W}(\omega)$	Actual waiting time of the j^{th} patient seen by provider k under scenario ω
$t_k^I(\omega)$	Actual idle time of provider k in the considered session under scenario ω
$t_k^O(\omega)$	Actual overtime of provider k in the considered session under scenario ω

Furthermore, for each scenario $\omega \in \Omega$, we define $r_{imn}(\omega)$ to be the service time of the i^{th} appointment for type $m \in M'$ service scheduled in the n^{th} slot, $n = 1, \ldots, N$. Finally, for each scenario $\omega \in \Omega$, we introduce auxiliary decision variables $\tau_{jk}(\omega)$, $t^A_{jk}(\omega)$, $t^S_{jk}(\omega)$, $t^E_{jk}(\omega)$, $t^W_{jk}(\omega)$, $t^I_k(\omega)$, and $t^O_k(\omega)$ in the formulation to determine the patient waiting time, provider idle time, and provider overtime. Thus, a scenario-based formulation for the problem in Phase II is presented as:

(P2) min $$\eta_W E_\Omega\left[\sum_{k=1}^{K}\sum_{j=1}^{J(\omega)} t^W_{jk}(\omega)\right] + \eta_I E_\Omega\left[\sum_{k=1}^{K} t^I_k(\omega)\right] + \eta_O E_\Omega\left[\sum_{k=1}^{K} t^O_k(\omega)\right] \quad (3.12)$$

s.t. $$\sum_{n=1}^{N} y_{mn} = D'_m, \qquad \text{for } m \in M'; \quad (3.13)$$

$$\sum_{i=1}^{q_{mn}(\omega)}\sum_{j=1}^{J(\omega)}\sum_{k=1}^{K} z^{jk}_{imn}(\omega) = q_{mn}(\omega), \text{ for } m \in M', n = 1,...,N, \text{ and } \omega \in \Omega; \quad (3.14)$$

$$\sum_{n=1}^{N}\sum_{m\in M'}\sum_{i=1}^{q_{mn}(\omega)} z^{jk}_{imn}(\omega) = b_{jk}(\omega), \text{ for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.15)$$

$$b_{j'k}(\omega) \geq b_{jk}(\omega), \text{ for } j, j' = 1,\ldots,J(\omega), j' \leq j, k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.16)$$

$$\sum_{n=1}^{N}\sum_{m\in M'}\sum_{i=1}^{q_{mn}(\omega)} (n-1)T_\Delta z^{jk}_{imn}(\omega) = t^A_{jk}(\omega), \text{ for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.17)$$

$$\sum_{n=1}^{N}\sum_{m\in M}\sum_{i=1}^{q_{mn}(\omega)} r_{imn}(\omega) z^{jk}_{imn}(\omega) = \tau_{jk}(\omega), \text{ for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.18)$$

$$t^E_{jk}(\omega) = t^S_{jk}(\omega) + \tau_{jk}(\omega), \text{ for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.19)$$

$$t^S_{jk}(\omega) \geq t^A_{jk}(\omega), \qquad \text{for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.20)$$

$$t^S_{jk}(\omega) \geq t^E_{j-1,k}(\omega), \qquad \text{for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.21)$$

$$t^W_{jk}(\omega) \geq (t^S_{jk}(\omega) - t^A_{jk}(\omega)) - 2T(1-b_{jk}(\omega)), \text{ for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.22)$$

$$t^W_{jk}(\omega) \leq (t^S_{jk}(\omega) - t^A_{jk}(\omega)) + 2T(1-b_{jk}(\omega)), \text{ for } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega; \quad (3.23)$$

$$t^I_k(\omega) \geq t^E_{J(\omega)k}(\omega) - \sum_{j=1}^{J(\omega)} \tau_{jk}(\omega), \qquad \text{for } k = 1,\ldots,K \text{ and } \omega \in \Omega; \quad (3.24)$$

$$t_k^I(\omega) \geq T - \sum_{j=1}^{J(\omega)} \tau_{jk}(\omega), \quad \text{for } k = 1,\ldots,K \text{ and } \omega \in \Omega; \tag{3.25}$$

$$t_k^O(\omega) \geq t_{J(\omega)k}^E(\omega) - T, \quad \text{for } k = 1,\ldots,K \text{ and } \omega \in \Omega; \tag{3.26}$$

$$y_{mn} \in Z_+, \quad \text{for } m \in M' \text{ and } n = 1,..., N; \tag{3.27}$$

$$b_{jk}(\omega) \in \{0,1\}, \quad \text{for } j = 1,\ldots,J(\omega), \text{ and } \omega \in \Omega; \tag{3.28}$$

$$z_{imn}^{jk}(\omega) \in \{0,1\}, \quad \text{for } i = 1,\ldots, q_{mn}(\omega), m \in M', n = 1,..., N,$$
$$j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega \tag{3.29}$$

$$\tau_{jk}(\omega), t_{jk}^S(\omega), t_{jk}^E(\omega), t_{jk}^W(\omega) \geq 0, \text{ for, } j = 1,\ldots,J(\omega), k = 1,\ldots,K, \text{ and } \omega \in \Omega \tag{3.30}$$

$$t_k^I(\omega), t_k^O(\omega) \geq 0, \quad \text{for } k = 1,\ldots,K \text{ and } \omega \in \Omega \tag{3.31}$$

In the objective function 3.12 of (P2), the three terms capture the expected costs of patient waiting time, provider idle time, and provider overtime, respectively. The weights for patient waiting time, provider idle time, and provider overtime are specified by the clinic manager. Constraints 3.13 are first-stage constraints that enforce the number of appointments to be scheduled for each type of service in the considered clinic session. Constraints 3.14 – 3.26 are second-stage constraints. Different from the standard two-stage stochastic program with recourse, (P2) may lead to recourse problems of different sizes for various first-stage decisions and scenarios. Without complicating the model presentation, we conveniently use $q_{mn}(\omega)$ in Constraint 3.14 to specify the number of patients who arrive for their appointments for type $m \in M'$ service scheduled in time slot *n*, which depends on the first-stage decision y_{mn} and given scenario $\omega \in \Omega$. In other words, the coupling between the two stages in (P2) is reflected in the scenario-wise specification of $q_{mn}(\omega)$.

For each scenario, Constraints 3.14 – 3.26 guarantee a valid assignment of appointments to slots in a clinic session. Constraints 3.14 guarantee that all patient showing up for their

appointments will be seen by some provider. Constraints 3.15 and 3.16 determine the sequence of patients seen by each provider. Constraints 3.17 and 3.18 specify the appointment times and service times of a sequence of patients seen by a provider. Constraints 3.19 validate the relationship of the starting time and the completion of each service and its service duration. Constraints 3.20 and 3.21 ensure that the starting time of each service must not be earlier than either its appointment time or the competition time of the previous service, i.e., for $\omega \in \Omega$, $t_{jk}^{S}(\omega) = \min(t_{jk}^{A}(\omega), t_{(j-1),k}^{E}(\omega))$ for j = 1,…,J(ω), and k = 1,…,K. Constraints 3.22 and 3.23 are used to specify the waiting time of each patient, i.e., $t_{jk}^{W}(\omega) = b_{jk}(\omega)(t_{jk}^{S}(\omega) - t_{jk}^{A}(\omega))$ for j = 1,…,J(ω), k = 1,…,K, and $\omega \in \Omega$. Constraints 3.24 and 3.25 specify the idle time of each provider, and constraints 3.26 specify the overtime of each provider.

In clinics with significant patient no-shows and highly variable service times, there may be appointment scheduling templates that are insignificantly inferior to the optimal solution in terms of the objective function 3.12. Hence, it is more desirable in clinical practice to present the scheduling practitioners with these templates. It is also more desirable to select a subset of them that are expected to increase the satisfaction of both patients and providers, as well as increase the robustness of the schedule under uncertainty. Therefore, a secondary objective can be used to select a promising template from a small set of templates that are near-optimal in terms of the primary objective in 3.12. To increase patient satisfaction, a secondary objective considered in this study is to minimize the maximal expected waiting time among patients scheduled in different slots, i.e.,

$$\min \max_{n=1,\ldots,N} \left\{ E_{\Omega}\left[\sum_{m \in M} \sum_{i=1}^{q_{mn}(\omega)} \sum_{k}^{K} \sum_{j=1}^{J(\omega)} t_{jk}^{W}(\omega) z_{imn}^{jk}(\omega) \right] \Big/ \sum_{m \in M} y_{mn} \right\}. \tag{3.32}$$

The above secondary objective function is nonlinear, thus it is not incorporated into the objective function 3.12. Instead, from a set of near-optimal templates in terms of 3.12, we select the template according to this secondary objective, which achieves the highest satisfactory equity among the waiting times of patients scheduled into different time slots of the clinic session. Note that the genetic algorithm introduced in Section 5 readily provides a set of templates rather than only one template.

3.2. Solution Approach

The MILP model in Phase I can be quickly solved by most commercial optimization software packages such as CPLEX, LINGO, and GAMS. However, the SMIP model in Phase II presents severe computational intractability (e.g., see Klein and Van, 1999). Therefore, we focus on the solution approach in Phase II.

In the decision-making problem in Phase II, it is relatively easy to evaluate the objective function for a given first-stage decision under certain scenario. However, it is difficult to evaluate the expected 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 described with continuous random variables and may follow a variety of distributions (e.g., gamma, lognormal, and Weibull distributions), depending on the service type and patient population (Cayirli and Veral, 2003, Bailey, 1952 Klassen and Rohleder, 1996, Cayirli et al., 2006). For two-stage SMIPs with continuous random variables, the expected recourse functions, in general, cannot be analytically derived for first-stage decisions. Hence, we apply Monte Carlo sampling to estimate the expected recourse. Furthermore, it is computationally challenging to identify promising first-stage decisions as the search space of feasible first-stage solutions is non-convex for the derived SMIP problems due to the integrality in the first stage (Klein and

Van, 1999, Schultz R, 2003). For the state-of-the-art research on applying Monte Carlo sampling to SMIPs, we refer to Kleywegt et al. (2002). Finally, in real-world practice, it is desirable to present to clinic managers several promising suboptimal schedule templates rather than a provably global optimal template. Therefore, we develop a genetic-algorithm-based approach to identify promising first-stage decisions.

Genetic algorithms are inspired by the biological evolution theory that the fittest individuals have more opportunities to reproduce. They define the population as a collection of solutions (termed *chromosomes* in the literature) to an optimization problem and generate new sets (termed *generations* in the literature) of solutions by combining pairs of good solutions in the existing population (Mitchell M, 1996). Along the evolution process, promising solutions are preserved and poor solutions are eliminated. By incorporating Monte Carlo (MC) sampling into genetic algorithm (GA), we develop the following procedure, named GA-MC procedure, to identify promising suboptimal solutions. The detailed procedures are summarized in the following:

3.2.1. Step 1: initialization

During initialization, the program will decide the value of population size p, the subpopulation size p_c, the mutation rate β, the Monte Carlo sample sizes for iteration (n_1) and best solution selection (n_2), and the iteration limit $g_{\max}$.

After the parameters are initialized, the problem will set the iteration index g to 0, and randomly generate the initial generation of p candidate solutions, which are represented by chromosomes. Note that, the i^{th} candidate in the initial population will be denoted by $\mathbf{x}^{(0)}(i)$.

After the initial population is created, a Monte Carlo simulation with sample size n_1 is run for each chromosome in the population to estimate the corresponding expected total cost. Note that the estimate expected cost for chromosome $\mathbf{x}^{(0)}(i)$ will be denoted by $w(i)$.

3.2.2. Step 2: population evolution

During population evolution, p_c chromosomes will be selected based on the roulette-wheel rule. The selected chromosome will be dived into $p_c/2$ pairs, and a two-point crossover operation will be conducted on each pair in order to generate new chromosomes. A mutation operation is then executed on each newly generated chromosomes to maintain the diversity of the population.

After the new chromosomes are generated, the Monte Carlo simulation with sample size n_1 is run to estimate the corresponding expected total cost of new chromosomes. The new chromosomes will replace the worst p_c chromosomes (with the largest $w(i)$) in the old generation "g", and form the new generation "g+1".

Update $w(i)$ of chromosomes in the new generation and set the current generation number to "g+1", i.e. $g = g+1$. If $g < g_{max}$, then the evolution process is ended and the program will proceed to best solution selection; otherwise the evolution process will be repeated.

3.2.3. Step 3: best solution selection in the last generation

The Monte Carlo simulation with sample size n_2 is run for each chromosome in the last generation to estimate the corresponding expected total cost and standard deviation. Note that, the i^{th} candidate in the last generation will be denoted by $\mathbf{x}^{(g)}(i)$ and the corresponding expected total cost is denoted by $w(i)$.

At last, the program will report the chromosomes with the smallest $w(i)$, as well as the chromosomes with $w(i)$ that is not greater with statistical significance than the smallest $w(i)$.

The proposed GA-MC procedure is coded using MATLAB 7.12.0 (MathWorks, Inc., 2012), and experiments are conducted to choose proper values for the parameters used in the GA-MC procedure, including the sample sizes n_1 and n_2. Based on our experimental results, a sample size of 200 is selected for iterative sampling, while a sample size of 2000 for the last

generation. Using the GA-MC procedure with these two sample sizes, it takes about 50 minutes to find a suboptimal solution for one instance of (P2). It is worth noting that when a new weekly scheduling template is implemented in the studied clinic, it generally takes at least a few weeks to test its performance before it is further implemented. Hence, the computational time required in the proposed GA-MC procedure is acceptable.

3.3. Case Study

In this section, we report a case study that demonstrates how well the proposed two-phase approach performs in the scheduling template design for a real clinic. The clinic characteristics and patient demand data used in the case study are collected from the women's clinic that motivates our study. The values of the parameters p, p_c, β, n_1, n_2, and $g_{\max}$ in the GA-MC procedure are selected through preliminary numerical experiments. The results of the case study are analyzed to identify the patterns of heuristic optimal scheduling templates.

3.3.1. Data collection and study design

The studied women's clinic offers eight types of services in eight 4-hour clinic sessions per week. Two physicians are available in each clinic session. Among the eight types of services, the service of Developmental Pediatrics for newly-born babies is not considered in this case study due to its long service time, which is approximately 2 hours on average. Currently, two clinic sessions are allocated to schedule Developmental Pediatrics visits. Thus, we consider the case to allocate the remaining seven types of services in six clinic sessions. These seven types of services are clustered into three categories according to the requirement on service equipment. For each service type, the no-show rate of the appointments is estimated based on a two-month patient no-show data set, and the probability distribution of service times is obtained by analyzing a one-week clinical operations data. Similar to existing studies (Klassen and Rohleder, 1996, Klassen and Yoogalingam, 2009, Cayirli et al., 2006), we assume that service times for

each service type are independent and lognormally distributed. Table 3.3 presents the no-show rate, the distribution of service times and the average number of weekly requests for each service type. The current average number of weekly requests for each service type was estimated by the clinic manager based on the historical clinical operations data. Due to the observed trend of increasing demand, the manager also predicted two levels of future demands, which are labeled as "Future I" and "Future II" in Table 3.3.

Table 3.3: Weekly demand, no-show rate, and service time distribution for each service type

Service category	Service type	No-show rate	Service time (in minutes)			Average number of requests for service		
			Avg.	Std.	Distribution LN(μ,σ^2)	Current	Future I	Future II
Low Risk OB	New Low Risk OB	0.162	25	8	LN(3.17,0.10)	4	8	11
	Follow Up Low Risk OB	0.053	6	3	LN(1.68,0.22)	22	43	64
High Risk OB	Follow Up High Risk OB	0.080	10	6	LN(2.15,0.31)	35	70	105
Gynecology	New GYN	0.488	18	12	LN(2.71,0.37)	16	32	47
	MAU GYN	0.487	13	3	LN(2.54,0.05)	4	8	12
	Established GYN	0.384	10	5	LN(2.19,0.22)	17	33	49
	Results GYN	0.321	15	4	LN(2.67,0.07)	5	9	14

In the case study, each clinic session is divided into 16 slots with equal length of 15 minutes. The weighting coefficients for patient waiting time, provider idle time and overtime are determined based on the average hourly wage and the average annual obstetrics-gynecology salary. Krueger (2009) reports that the average U.S. hourly wage is $17.4 per hour in 2009. Based on surveys of physicians in different medical specialties, the average annual obstetrics-gynecology salary in the U.S. is $261,000 in 2011 (Medical Resource Group, 2011). Since on average an obstetrics-gynecology physician works 50 weeks per year and 40 hours per week, the average hourly cost of hiring an obstetrics-gynecology physician is $130.5. Meanwhile, considering the compensation for physicians' unwillingness to work overtime, the provider overtime cost per hour is assumed to be 1.5 times of the regular payment. Thus, it is assumed in the case study that the weighting coefficients for patient waiting time, provider idle time and

overtime are 1, 7.5 and 11.25, respectively. These weighting coefficients as well as the parameters of clinic setting are summarized in Table 3.4.

Table 3.4: Clinic setting parameters and weighing coefficients in the case study

Notation	Description	Value
K	Total number of physicians available in each clinic session	2
N	Number of appointment slots in each clinic session	16
T_s	Total length of each clinic session	4 hours
T_Δ	Length of an appointment slot	15 minutes
η_I	Weighting coefficient of provider idle time	7.5
η_O	Weighting coefficient of provider overtime	11.25
η_W	Weighting coefficient of patient waiting time	1

3.3.2. Algorithmic parameter selection for the GA-MC procedure

Several numerical experiments are conducted to determine good combinations of parameters used in the GA-MC procedure. Table 3.5 presents the optional values tested for the parameters (p, p_c, β, n_1 and n_2). In the experiments, 16 appointments for New GYN service, 4 appointments for MAU GYN service, 17 appointments for Established GYN service and 5 appointments for Results GYN service are scheduled to time slots in a clinic session. Note that these different types of services belong to the same category.

Table 3.5: Experiment design for selecting the parameters used in the GA-MC procedure

Notation	Description	Levels
β	Mutation probability	0.5%, 1%, 5%, 10%
n_1	Sample size for interior sampling	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000
n_2	Sample size for the comparison of the last-generation chromosomes	100, 200, 500, 1000, 2000, 5000
p	Population size	50, 100, 150
p_c	Population size of subdivision for crossovers	$p_c = r_c \times p$
r_c	Percentage of chromosomes selected for crossover	0.2, 0.5, 0.8

First, the sample size for the comparison of the last-generation chromosomes (n_2) is determined according to the accuracy target that the 95% confidence interval of a weighted total cost estimated by Monte Carlo simulation is less than 3% of its average. Table 3.6 presents the

average and the standard error of the weighted total cost for a chromosome. These results demonstrate that it suffices to use 2000 for n_2.

Table 3.6: Weighted total cost of a chromosome estimated using each candidate sample size n_2

	Weighted total cost estimated by Monte Carlo simulation		
Sample Size (n_2)	Average	Standard Error	95% Confidence Interval
100	1254	37	1254±73
200	1260	25	1260±49
500	1275	15	1275±29
1000	1268	11	1268±22
2000	1259	8	1259±16
5000	1268	5	1268±10

Next, we investigate the best combination for parameters p, p_c and β by comparing the performance of the GA-MC procedure under the 36 combinations for p, p_c and β as listed in Table 3.5, in terms of the convergence rate and the best objective function value found. For each optional value of β, we identify the best pair of values for p and r_c, and Fig. 3.1 presents the performance of the GA-MC procedure under these best pairs of values. The results in this figure show that the best combination for p, p_c, and β are 100, 50, and 1%, respectively. Meanwhile, Fig. 3.1 also illustrates that the GA-MC procedure using the best combination for p, p_c, and β converges to a constant objective function value after about 100 iterations. Thus, the iteration limit (g_{max}) determined in the case study is 100.

Finally, the sample size for iterative sampling (n_1) is determined based on the computational time and the best objective function value found. Three replications are run using each candidate sample size n_1. Fig. 3.2 illustrates the average, the minimum, and the maximum of the computational times and the best objective function values found using different sample sizes. It is observed in Fig. 3.2 that as the sample size n_1 increases, the computational time roughly linearly increases, and the best objective function value found converges. To obtain a reasonably good solution within acceptable time, a sample size of 200 is chosen for iterative

sampling. Table 3.7 summarizes the values for parameters p, p_c, β, n_1, n_2, and g_{max} of the GA-MC Procedure used in the case study.

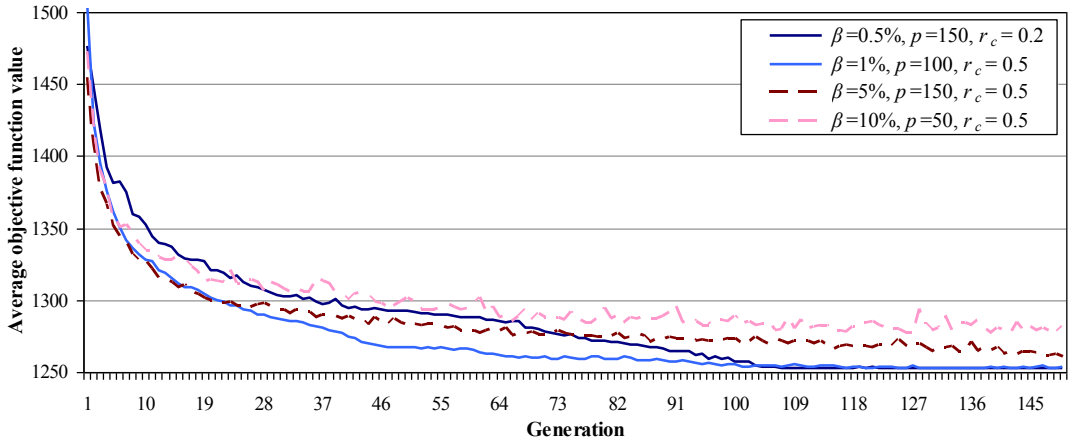

Fig. 3.1: Convergence of the average objective value under 4 mutation rates

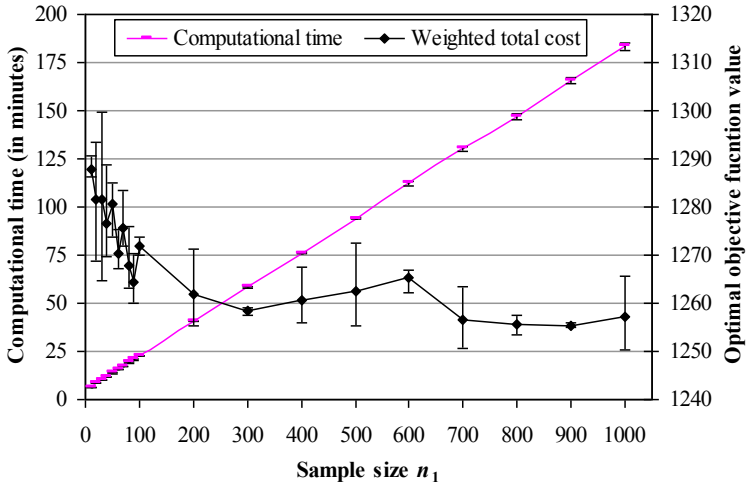

Fig. 3.2: Computational times and best objective function values for 19 optional values of n1

3.3.3. Case study results

For the case study, we investigate the optimal scheduling template design in three cases, which correspond to the three demand levels in Table 3.3. Tables 3.3 and 3.4 summarize the

clinic characteristics and the weighting coefficients for patient waiting time, provider idle time, and provider overtime used in the three cases. For each case, a master scheduling template is developed first by solving (P1). Then, according to the assignment of service types specified in the master scheduling template, the appointments for each service type are allocated to 15-minute slots by solving (P2).

Table 3.7: Parameters of the GA-MC Procedure used in the numerical study

Notation	Description	Value
β	Mutation probability	0.01
g_{max}	Iteration limit	100
n_1	Sample size for interior sampling	200
n_2	Sample size for the comparison of the last-generation chromosomes	2000
P	Population size	100
p_c	Population size of subdivision for crossovers	50

Tables 3.8 – 3.10 present the optimal master scheduling templates for the three cases, respectively. These master scheduling templates show that in the women's clinic, two clinic sessions should be allocated to the services in each category. Meanwhile, the results in Table 3.8 demonstrate that the utilization of provider capacity in any clinic session is lower than 35% in Case 1 comparing to physician capacity (two physicians available in each 4-hour clinic session). This observation implies that the current provider capacity in the studied women's clinic can handle an increased demand. Tables 3.9 and 3.10 show that due to the increased demand, the utilization of provider capacity ranges from 42% to 67% in Case 2, and from 61% to 102% in Case 3. Furthermore, Table 3.11 presents the master scheduling template currently used in the studied women's clinic and the provider capacity utilization under this template. The comparison of the master scheduling templates in Table 3.8 – 3.11 reveals that the optimal master scheduling

templates, presented in Tables 3.8 – 3.10, balance provider capacity utilization over clinic sessions much better than the master scheduling template used in the studied clinic.

Table 3.8: Optimal master scheduling template for Case 1 *

	Number of appointments for each service type in each clinic session					
Service Type	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
New Low Risk OB	2	2				
Follow Up Low Risk OB	11	11				
Follow Up High Risk OB			17	18		
New GYN					4	12
MAU GYN					2	2
Established GYN					12	5
Results GYN					4	1
Expected total service time (in minutes)	104.4	104.4	156.5	165.7	164.8	164.8

* Case 1 corresponds to the "Current" level of the weekly demand in Table 3.3.

Table 3.9: Optimal master scheduling template for Case 2 *

	Number of appointments for each service type in each clinic session					
Service Type	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
New Low Risk OB	6	2				
Follow Up Low Risk OB	14	29				
Follow Up High Risk OB			35	35		
New GYN					12	20
MAU GYN					4	4
Established GYN					20	13
Results GYN					6	3
Expected total service time (in minutes)	205.3	206.7	321.5	321.5	205.3	206.7

* Case 2 corresponds to the "Future I" level of the weekly demand in Table 3.3.

Table 3.10: Optimal master scheduling template for Case 3 *

	Number of appointments for each service type in each clinic session					
Service Type	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
New Low Risk OB	2	9				
Follow Up Low Risk OB	45	19				
Follow Up High Risk OB			53	52		
New GYN					32	15
MAU GYN					3	9
Established GYN					20	29
Results GYN					4	10
Expected total service time (in minutes)	297.6	296.5	487.8	478.6	478.6	478.6

* Case 3 corresponds to the "Future II" level of the weekly demand in Table 3.3.

Table 3.11: Master scheduling template used in the studied women's clinic

Service Type	Number of appointments for each service type in each clinic session			
	Wednesday morning	Monday and Thursday morning	Monday and Thursday afternoon	Friday morning
New Low Risk OB	3			
Follow Up Low Risk OB	15			
Follow Up High Risk OB		40		
New GYN			9	3
MAU GYN			3	1
Established GYN			9	4
Results GYN			2	1
Expected total service time (in minutes)	148.1	368.0	178.8	69.2

Tables 3.12 – 3.14 present the optimal assignments of seven types of services to time slots in the three cases, respectively. For Case 1, the results in Table 3.12 demonstrate that the total patient waiting time, the waiting time of individual patients and provider overtime in the heuristic optimal scheduling template are relatively insignificant compared to the provider idle time. The reason is the low utilization of provider capacity in Case 1. The heuristic optimal scheduling template in Table 3.12 implies the following rules for cases with low capacity utilization.

- No appointment should be scheduled in the last time slot, which could eliminate provider overtime.
- For the low-risk or high-risk OB sessions, appointments should be evenly scheduled in time slots.
- For the GYN sessions, more appointments for New GYN service should be scheduled in early time slots. Due to the higher variation in the service time among New GYN visits, scheduling them in later time slots increases the chance that provider overtime occurs.

Compared to Case 1, the total patient waiting time increases and the total provider idle time decreases in Cases 2 and 3. In Case 2, the waiting time of individual patients is acceptable, and the provider overtime is still relatively insignificant. The heuristic optimal scheduling

template, as presented in Table 3.13, shows similar observations and provides similar managerial insights for cases with middle capacity utilization. In Case 3, the waiting time of individual patients and provider overtime significantly increase in the four sessions in which provider capacity utilization is about 100%. The heuristic optimal assignments of appointments in these four sessions imply the following rules for cases with high capacity utilization.

- For the high-risk OB sessions, appointments should be evenly scheduled in time slots except the first slot. More appointments should be scheduled in the first slot.
- For the GYN sessions, more appointments for Established GYN service should be scheduled in early time slots, and appointments for New GYN should be evenly scheduled. The reason is that scheduling visits with lower-variation service times could reduce patient waiting time.

Table 3.12: Heuristic optimal scheduling templates for Case 1

	Sessions 1 and 2		**Session 3**	**Session 4**	**Session 5**				**Session 6**			
Optimal Scheduling Templates												
Slot Index	New Low Risk OB	Follow Up Low Risk OB	Follow Up High Risk OB	Follow Up High Risk OB	New GYN	MAU GYN	Established GYN	Results GYN	New GYN	MAU GYN	Established GYN	Results GYN
1	-	-	2	2	1	-	-	1	-	-	2	-
2	-	1	1	1	-	-	-	-	-	-	1	1
3	1	-	1	1	-	-	2	-	2	-	-	-
4	-	1	1	2	-	2	1	-	1	-	-	-
5	-	1	1	1	-	-	1	-	1	-	-	-
6	1	-	2	1	-	-	1	-	1	-	-	-
7	-	-	-	1	1	-	2	1	2	1	-	-
8	-	1	2	1	1	-	1	-	1	-	-	-
9	-	1	1	1	1	-	1	-	1	-	-	-
10	-	1	1	1	-	-	1	-	1	-	-	-
11	-	1	1	1	-	-	2	-	2	-	-	-
12	-	1	1	2	-	-	-	-	-	-	-	-
13	-	1	2	1	-	-	-	2	-	1	1	-
14	-	1	1	2	-	-	-	-	-	-	1	-
15	-	1	-	-	-	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-	-	-	-	-	-
Total appointments	**2**	**11**	**17**	**18**	**4**	**2**	**12**	**4**	**12**	**2**	**5**	**1**
Performance of Scheduling Template *												
Patient waiting time	0.0		1.4727	2.1	4.8				7.9			
Provider idle time	376.0		323.2	314.0	314.9				313.9			
Provider overtime	0.0		0.0	0.0	0.0				0.1			
Weighted total cost	2820		2425.7	2357.1	2366.6				2362.5			
Max. waiting time *	0.0033		0.2806	0.3111	0.6214				1.3211			

The performance of the scheduling template is evaluated in terms of total patient waiting time, total idle time of two physicians, total overtime of two physicians, the weighted total waiting cost, and the maximum of the average waiting time per slot. All these times are in minutes.

* Max. waiting time represents the maximum of the average patient waiting time per slot.

Table 3.13: Heuristic optimal scheduling templates for Case 2

	Session 1		**Session 2**		**Sessions 3 and 4**	**Session 5**				**Session 6**			
Optimal Scheduling Templates													
Slot Index	New Low Risk OB	Follow Up Low Risk OB	New Low Risk OB	Follow Up Low Risk OB	Follow Up High Risk OB	New GYN	MAU GYN	Established GYN	Results GYN	New GYN	MAU GYN	Established GYN	Results GYN
1	-	2	-	2	3	1	-	2	1	3	-	2	-
2	-	2	1	1	2	1	-	2	-	2	-	1	-
3	1	1	-	1	2	2	1	1	-	1	-	1	-
4	1	-	-	2	2	-	1	1	-	-	-	1	1
5	-	1	-	2	3	1	-	2	-	1	-	1	1
6	1	-	-	2	2	2	-	1	-	3	-	1	-
7	-	1	-	2	2	-	-	2	-	2	-	-	-
8	-	1	-	3	3	1	-	1	1	3	-	-	-
9	1	1	-	2	2	-	-	1	1	1	-	-	-
10	1	-	-	2	3	2	-	1	-	2	-	1	-
11	-	1	-	2	2	2	-	-	1	1	-	-	-
12	-	1	-	2	2	-	2	-	-	-	1	2	-
13	1	1	1	1	2	-	-	1	1	1	1	-	-
14	-	1	-	1	2	-	-	1	1	-	1	1	1
15	-	1	-	2	3	-	-	3	-	-	1	2	-
16	-	-	-	2	-	-	-	1	-	-	-	-	-
Total appointments	**6**	**14**	**2**	**29**	**35**	**12**	**4**	**20**	**6**	**20**	**4**	**13**	**3**
Performance of Scheduling Template *													
Patient waiting time	1.9		6.3		64.1	94.0				114.8			
Provider idle time	274.6		272.9		157.8	160.4				158.2			
Provider overtime	0.0		0.1		2.1	1.2				1.1			
Weighted total cost	2061.8		2053.9		1271.8	1310.1				1314.5			
Max. waiting time *	0.7211		1.2647		3.5762	4.2118				6.2966			

The performance of scheduling template is evaluated in terms of total patient waiting time, total idle time of two physicians, total overtime of two physicians, weighted total waiting cost, and maximum of the average waiting time per slot. All these times are in minutes.

* Max. waiting time represents the maximum of the average patient waiting time per slot.

Table 3.14: Heuristic optimal scheduling templates for Case 3

	Session 1		**Session 2**		**Session 3**	**Session 4**	**Session 5**				**Session 6**			
Optimal Scheduling Templates														
Slot Index	New Low Risk OB	Follow Up Low Risk OB	New Low Risk OB	Follow Up Low Risk OB	Follow Up High Risk OB	Follow Up High Risk OB	New GYN	MAU GYN	Established GYN	Results GYN	New GYN	MAU GYN	Established GYN	Results GYN
1	-	3	1	1	6	5	4	-	2	1	1	2	2	2
2	-	3	-	1	2	4	2	1	2	-	-	-	4	-
3	-	3	2	-	4	3	2	-	1	-	2	-	1	1
4	-	3	-	-	3	3	3	-	2	-	-	-	2	2
5	1	3	-	2	3	3	1	-	2	-	1	-	2	-
6	-	2	1	2	4	3	-	1	1	1	2	-	3	-
7	-	3	-	2	3	4	2	-	1	-	-	2	1	-
8	-	3	1	1	3	3	3	-	1	-	2	1	1	-
9	-	3	-	2	3	3	-	-	1	2	-	-	2	1
10	-	3	2	-	3	3	2	-	2	-	2	-	2	-
11	1	2	1	-	4	3	2	-	2	-	1	1	3	-
12	-	2	-	1	2	3	4	-	1	-	-	1	1	1
13	-	3	1	1	4	3	2	-	-	-	2	-	1	1
14	-	4	-	2	3	3	1	1	1	-	-	1	1	-
15	-	3	-	2	4	3	3	-	-	-	2	-	3	1
16	-	2	-	2	2	3	1	-	1	-	-	1	-	1
Total appointments	**2**	**45**	**9**	**19**	**53**	**52**	**32**	**3**	**20**	**4**	**15**	**9**	**29**	**10**
Performance of Scheduling Template *														
Patient waiting time	83.8		33.2		720.0	644.4	637.3				589.9			
Provider idle time	182.7		182.1		21.6	25.5	40.9				39.7			
Provider overtime	0.1		0.1		31.5	26.3	41.0				38.0			
Weighted total cost	1455.5		1400.4		1235.8	1130.8	1405.8				1315.0			
Max. waiting time **	2.9102		4.2010		19.8888	15.9322	24.2974				18.3732			

* The performance of scheduling template is evaluated in terms of the total patient waiting time, the total idle time of two physicians, the total overtime of two physicians, the weighted total waiting cost, and the maximum of the average waiting time per slot. All these times are in minutes.

** "Max. waiting time" represents the maximum of the average patient waiting time per slot.

Table 3.15: Scheduling template used in the studied women's clinic

	Wednesday morning		**Monday and Thursday morning**	**Monday and Thursday afternoon**			
Optimal Scheduling Templates							
Slot Index	New Low Risk OB	Follow Up Low Risk OB	Follow Up High Risk OB	New GYN	MAU GYN	Established GYN	Results GYN
1	-	-	2	2	-	-	-
2	1	1	2	-	-	2	-
3	-	-	2	1	1	-	-
4	1	1	-	-	-	-	-
5	-	-	4	1	1	-	-
6	1	1	3	-	-	2	-
7	-	-	3	2	-	-	-
8	-	2	4	-	-	-	-
9	-	2	3	1	1	-	-
10	-	2	3	-	-	2	-
11	-	2	4	-	-	1	1
12	-	2	3	2	-	-	-
13	-	2	3	-	-	-	-
14	-	-	4	-	-	2	1
15	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-
Total appointments	**3**	**15**	**40**	**9**	**3**	**9**	**2**
Performance of Scheduling Template *							
Patient waiting time	2.5		340.6	9.4			
Provider idle time	332.0		113.5	301.4			
Provider overtime	0.0		3.6	0.1			
Weighted total cost	2492.2		1232.6	2270.6			
Max. waiting time *	0.39		16.2	0.95			

The performance of the scheduling template is evaluated in terms of total patient waiting time, total idle time of two physicians, total overtime of two physicians, weighted total waiting cost, and maximum of the average waiting time per slot. All these times are in minutes.

* Max. waiting time represents the maximum of the average patient waiting time per slot.

To justify the promising performance of heuristic optimal scheduling templates in Tables 3.12 – 3.14, we present in Table 3.15 the scheduling template currently used in the studied women's clinic and its performance estimate via Monte Carlo sampling. The weekly patient demand with the scheduling template used in the clinic is comparable to the weekly demand in Case 2. The comparison between results in Tables 3.13 and 3.15 shows that the worst

performance of the scheduling template in Case 2 is better than that of the currently used scheduling template in terms of each considered performance metric. This performance improvement attributes to a better balance of provider capacity utilization in the master scheduling template obtained by solving (P1) and a better assignment of appointments to time slots for each individual clinic session obtained by solving (P2). The comparison result supports the conclusion that a better balance of provider capacity utilization can help reduce the longest waiting time of individual patients, provider idle time and overtime over various clinic sessions.

To further assess the quality of solutions found by the proposed GA-MC procedure, we construct two small-size instances of (P2) whose optimal solutions can be found in affordable time using exhaustive search. In the two instances (Instance I and Instance II), five and six follow-up high-risk OB appointments, respectively, need to be allocated into 16 equal-length time slots in a clinic session. In both instances, the weighting coefficients of 1, 12 and 18 are used for patient waiting time, provider idle time and overtime, respectively. Table 3.16 summarizes the sizes of the solution spaces of the two instances, the smallest objective function values found, and the computational times when using exhaustive search and the proposed GA-MC procedure. A sample size of 2000 is chosen to estimate the expected objective function values during exhaustive search, while the parameter values in Table 3.7 are used in the GA-MC procedure. The results in Table 3.16 demonstrate that the GA-MC procedure finds an identical solution as exhaustive search for each instance but in significantly shorter time. Furthermore, the computational time using the GA-MC procedure is hardly affected by the size of the solution space, whereas the computational time using exhaustive search increases linearly with the increase in the solution space size. This implies that the GA-MC procedure is very efficient, especially when solving large-size instances of (P2).

Table 3.16: Performance assessment of the proposed GA-MC procedure with two small-size instances

	Number of feasible solutions	95% confidence interval of the objective function value		Computational time*	
		GA-MC procedure	Exhaustive search	GA-MC procedure	Exhaustive search
Instance I	15504	5208±8	5208±8	31.7 minutes	10.8 hours
Instance II	54264	5098±9	5098±9	32.1 minutes	38.4 hours

* Each instance was solved separately by the GA-MC procedure and exhaustive search on an average personal computer.

4. APPOINTMENT SCHEDULING OPTIMIZATOIN FOR OPEN ACCESS SCHEDULING SYSTEM

4.1. Problem Description and Formulation

In this problem, we consider a primary care clinic admitting three types of patients, namely, patients with pre-booked appointments (Type 1), open access patients (Type 2), and walk-in patients (Type 3). The clinic has a fixed number of slots during each clinical day and reserves a portion of these slots for Type 2 and Type 3 patients. This is the only way to allow Type 2 and Type 3 patients to be admitted by this clinic, because the primary care clinics usually have sufficient demand of Type 1 patient to fill all the slots of a clinical day in the United States. Meanwhile, we assume that the clinic adopts double booking policy to mitigate the adverse effect of patient no-shows and short-notice cancellations. Hence, each slot during a clinic day can be single booked, or double booked for Type 1 patients, or reserved for a Type 2 or Type 3 patient. This way, an appointment schedule will be formed. The objective of this problem is to find the optimal scheduling template, which minimizes the average cost per unit of patient accessibility during a clinical day. In this paper, the cost is measured as the weighted sum of total patient waiting time, provider idle time, and provider overtime, while the patient accessibility is evaluated based on the total number of Type 2 and Type 3 patients admitted. In the following, the three types of patients are described in detail. Also, the patient accessibility is measured by the total number of same-day appointment patients and work-in patient admissions.

4.1.1. Type 1: patients with pre-booked appointments

Type 1 patients are scheduled days/weeks/months before their appointment dates. The patients can arrive for appointments on time, cancel their appointments, or do not show up. If they arrive on time, the provider is expected to see them on the scheduled appointment times. If an appointment is cancelled, the slot will be re-opened for booking patients of other types

immediately after the cancellation. If Type 1 patients do not show up for their appointments, the slots will be provided to walk-in patients who are still in the clinic. It is assumed that both no-shows and cancellations are independent of patients, e.g., whether a patient is a no-show or not does not influence the probability of no-show for another patient. Late arrivals are not considered as a special case since we assume that the patients are treated as walk-ins if they arrive late.

4.1.2. Type 2: open access patients

For any open access clinic, a certain number of slots are reserved for the patients calling for same-day appointments. If there are slots available at the time of requesting, the request will be accepted and the patient will be scheduled into one of the available slots based on his/her choice; otherwise, the request will be rejected. Similarly to Type 1 patients, Type 2 patients with same-day appointments can arrive on time, cancel their appointments, or do not show up for their appointments. However, it is assumed that Type 2 patients have lower no-show rate and cancellation rate, compared with Type 1 patients. It is worthwhile to note that the re-opened appointments are also available for Type 2 patients.

4.1.3. Type 3: walk-in patients

The walk-in patients generally visit clinic without appointments. For clinics that admit walk-in patients, the walk-in patients are usually put into the first available slot if there are any slots left at the time of their arrivals. Otherwise, the walk-in patients need to wait for the re-opened slots in the clinic. It is reasonable to assume that walk-in patients will stay in the clinic for a while in order to get a re-opened slot. If any slot is re-opened, it will be given to the walk-in patient who has the longest waiting time in the clinic. In case there are no walk-in patients at the time of re-opening, this slot will be kept available for both Type 2 and Type 3 patients until the time past the starting time of this appointment slot. However, if a walk-in patient cannot obtain an appointment within a certain waiting period, he/she will leave the clinic.

4.1.4. Appointment scheduling template

We represent the appointment scheduling template as a set of N numbers, where N equals the number of slots in a clinical day. For each slot in the scheduling template, we need to decide how many Type 1 patients (0, 1, or 2) need to be scheduled in it. If the number is 0 for a slot, the slot is reserved for Type 2 or Type 3 patients; if the number is 1, only one patient with regular appointment is scheduled; if the number is 2, this slot is double booked. Given an appointment schedule, the provider expects to see the patients at the beginning of their appointment times. No patients will be seen before their appointment times. However, if a provider sees a patient after his/her appointment time, the delay between the starting time to see the patient and the appointment time is considered as the patient waiting time, which is a portion of the clinical day cost. Another portion of the clinical day cost is the provider idle time, which is measured as the total time when the provider does not see patients during working hours. The last portion of the clinical day cost is the provider overtime, which is measured by the total time that the provider spends to see patients after working hours. Note that the walk-in patient waiting period for the re-opened slot is not taken into consideration of the clinical day cost. This is because walk-in patients usually expect a period of waiting and the clinic has no obligation to serve these patients.

To better illustrate how the appointment schedule templates influence the average cost per unit of patient accessibility, a simple example is presented in Fig. 4.1. In the example, a clinic provides eight 30-minute appointment slots. The clinic opens at time 0, and the first appointment starts at time 30. The working hours of the provider are from time 30 to time 270. It can be seen from the figure that there are two different appointment templates, both of which have 4 patients pre-booked but in different slots. If all the patients arrive on time, and open access appointment requests and walk-in patients are coming at the same time under both scheduling templates, i.e., two open access requests occur at time 90 and 150, one walk-in

patient arrives at time 210. It is clear that using template 1, the clinic can admit all these three patients in the reserved slots (5, 6, 7 and 8). However, using template 2, only the open access appointment request at 90 minutes could be accepted. As a result, 3 reserved slots in template 2 will be wasted. Apparently, in this example, template 2 generates more cost due to longer provider being idle, while it also admits less Type 2 and Type 3 patients.

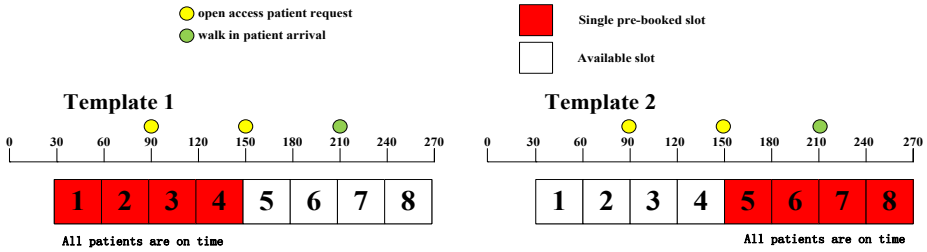

Fig. 4.1: A simple example showing the effect of scheduling template

4.1.5. Formulation

To formulate the problem, we introduce the notations in the following:

Table 4.1: Indices and parameters

Indices	
i	Appointment index
j	Index of patient seen by a provider
p	Type 2 patient index
q	Type 3 patient index
t, t'	Time index
Parameters	
T_Δ	Length of an appointment
T	Length of a clinical day
Δt	Time unit in the system
N	Number of slots in a clinical day
α_R	Cancellation rate of Type 1 patients
α_O	Cancellation rate of Type 2 patients
β_R	No-show rate of Type 1 patients
β_O	No-show rate of Type 2 patients
θ	Arrival rate of Type 2 patient requests
λ	Arrival rate of Type 3 patients
η_W, η_I, η_O	Cost coefficients of patient waiting time, provider idle time, and provider overtime, respectively

Table 4.2: Random variables and decision variables

Random variables	
P	Total number of appointment requests by Type 2 patients
Q	Total number of Type 3 patient arrivals
$C_i^R(t)$	Number of Type 1 patient cancellations of slot i occurred between time t and $t+\Delta t$.
$C_i^O(t)$	Number of Type 2 patient cancellations of slot i occurred between time t and $t+\Delta t$
$N_i^R(t)$	Number of Type 1 patient no-shows of slot i occurred between time t and $t+\Delta t$
$N_i^O(t)$	Number of Type 2 patient no-shows of slot i occurred between time t and $t+\Delta t$
$r^O(t)$	Total number of Type 2 appointment requests received between time t and $t+\Delta t$
t_p^O	Arrival time the p^{th} appointment request by Type 2 patients, $p=1,2,\cdots P$
t_q^W	Arrival time of the q^{th} walk-in patient, $q=1,2,\cdots Q$
W_q^W	Time that the q^{th} walk-in patient is willing to wait before he/she can get a same-day appointment in the clinic
$W(t)$	Total number of walk-in patients waiting in the clinic at time t
$W^A(t)$	Number of walk-in patient arrivals between time t and $t+\Delta t$
$W^T(t)$	Total number of walk-in patient arrivals by time t
$W^L(t)$	Number of walk-in patients left the clinic without seeing the provider between time t and $t+1$
P_j^D	Consultation time of the j^{th} patient seen by the provider
Decision variables	
A_i	Number of pre-booked patients in slot i, $i=1,2,\cdots N$
$A_i^O(t)=\begin{cases}1, \text{if a patient of Type 2 is scheduled in the } i^{th} \text{ slot by time } t\\0, \text{otherwise}\end{cases}$	
$A_i^W(t)=\begin{cases}1, \text{if the } i^{th} \text{ slot is assigned to a patient of Type 3 by time } t\\0, \text{otherwise}\end{cases}$	
$r_i^O(t)$	Number of Type 2 appointment requests accepted and scheduled in slot i between time t and $t+\Delta t$
$r_i^W(t)$	Number of walk-in patients accepted and assigned to slot i between time t and $t+\Delta t$
$a_{iq}(t)=\begin{cases}1, \text{if the } q^{th} \text{ walk-in patient is scheduled into the } i^{th} \text{ slot between time } t \text{ and } t+\Delta t\\0, \text{otherwise}\end{cases}$	
P_j^A	The appointment index of the j^{th} patient seen by the provider
P_j^S	Consultation starting time of the j^{th} patient seen by the provider
P_j^E	Consultation ending time of the j^{th} patient seen by the provider
P_j^W	Waiting time of the j^{th} patient seen by the provider

Given the notation in Table 4.1 and Table 4.2, the assumption of independent no-shows implies that the number of Type 1 patient no-shows of the i^{th} slot follows a binomial distribution

with parameters A_i and β_R, i.e. $\sum_{t\le T} N_i^R(t) \sim Binomial(A_i, \beta_R)$. Similarly, Type 1 patient cancellations, and Type 2 patient no-shows and cancellations of the i^{th} slot also follow a binomial distribution, i.e., $\sum_{t\le T} C_i^R(t) \sim Binomial(A_i, \beta_R)$, $\sum_{t\le T} N_i^O(t) \sim Binomial(\sum_{t\le T} r_i^O(t), \beta_O)$, and $\sum_{t\le T} C_i^O(t) \sim Binomial(\sum_{t\le T} r_i^O(t), \alpha_O)$, respectively. On the other hand, the independent arrivals of Type 2 and Type 3 patient requests indicate that the inter-arrival time of both requests follow exponential distributions, i.e. $t_{p+1}^O - t_p^O \sim \exp(\lambda)$ and $t_{q+1}^W - t_q^W \sim \exp(\theta)$. By applying the indices, parameters, random variables, and decision variables shown in Table 4.1 and Table 4.2, the problem can be formulated as follows:

Min

$$f = \left(\eta_W \sum_{j=1}^{J} P_j^W + \eta_I (T + \max\{0, P_J^E - T\} - \sum_{j=1}^{J} P_j^D) + \eta_O \max\{0, P_J^E - T\} \right) \Big/ \sum_{i=1}^{N} (A_i^O(T) + A_i^W(T)) \tag{4.1}$$

s.t.

$$P_j^W = P_j^S - P_j^A T_\Delta \quad \forall j \tag{4.2}$$

$$P_j^S = \begin{cases} P_{j-1}^E, \text{ if } P_{j-1}^E \ge P_j^A T_\Delta \\ P_j^A L, \text{otherwise} \end{cases} \quad \forall j, \tag{4.3}$$

$$P_j^E = P_j^S + P_j^D \quad \forall j \ , \tag{4.4}$$

$$0 \le A_i \le 2 \quad \forall i, \tag{4.5}$$

$$A_i \ge \sum_{t\le T} [C_i^R(t) + N_i^R(t)] \quad \forall i, \tag{4.6}$$

$$A_i^O(t) \ge C_i^O(t) + N_i^O(t) \quad \forall i, \tag{4.7}$$

$$\sum_{t \geq iT_\Delta} C_i^R(t) = 0 \quad \forall i \,, \tag{4.8}$$

$$\sum_{t \geq iT_\Delta} C_i^O(t) = 0 \quad \forall i \,, \tag{4.9}$$

$$\sum_{t \neq iT_\Delta} N_i^R(t) = 0 \quad \forall i \,, \tag{4.10}$$

$$\sum_{t \neq iT_\Delta} N_i^O(t) = 0 \quad \forall i \,, \tag{4.11}$$

$$C_i^O(t), N_i^O(t), C_i^R(t), N_i^R(t) \geq 0 \quad \forall i \,, \tag{4.12}$$

$$A_i^O(t + \Delta t) = A_i^O(t) - C_i^O(t) - N_i^O(t) + r_i^O(t) \quad \forall i \,, \tag{4.13}$$

$$A_i^W(t + \Delta t) = A_i^W(t) + r_i^W(t) \quad \forall i \,, \tag{4.14}$$

$$A_i^O(t) + A_i^W(t) \leq 1 \quad \forall i \,, \tag{4.15}$$

$$0 \leq A_i^O(t) \leq \max(0, [\sum_{t' < t} C_i^R(t')] - A_i - A_i^W(t) + 1) \quad \forall i \,, \tag{4.16}$$

$$0 \leq A_i^W(t) \leq \max(0, [\sum_{t' < t} C_i^R(t') + N_i^R(t')] - A_i - A_i^O(t) + 1) \quad \forall i \,, \tag{4.17}$$

$$I_i(t) = \begin{cases} 1, \text{if } iT_\Delta \geq t \\ 0, \text{otherwise} \end{cases} \quad \forall i \,, \tag{4.18}$$

$$0 \leq r_i^O(t) \leq \max(0, \{[\sum_{t' < t} C_i^R(t')] - A_i - A_i^O(t) - A_i^W(t) + 1\} \cdot I_i(t)) \quad \forall i \,, \tag{4.19}$$

$$0 \leq r_i^W(t) \leq \max(0, \{[\sum_{t' \leq t} C_i^R(t') + N_i^R(t')] - A_i - [A_i^O(t) - N_i^O(t)] - A_i^W(t) + 1\} \cdot I_i(t)) \quad \forall i \,, \tag{4.20}$$

$$r^O(t) \geq \sum_{i=1}^{N} r_i^O(t) \,, \tag{4.21}$$

$$W(t) + W^A(t) \geq \sum_{i=1}^{N} r_i^W(t) \,, \tag{4.22}$$

$$W(t) + W^A(t) \geq W^L(t) + \sum_{i=1}^{N} r_i^W(t) \,, \tag{4.23}$$

$$W(t+\Delta t) = W(t) - W^L(t) - \sum_{i=1}^{N} r_i^W(t) + W^A(t), \tag{4.24}$$

$$W^L(t) = \sum_{q=1}^{W^T(t)} [Z_q^W(t) - Z_q^W(t) \sum_{t' \le t} \sum_{i=1}^{N} a_{iq}(t')], \text{ where } Z_q^W(t) = \begin{cases} 1, \text{if } t_q^W + W_q^W = t \\ 0, \text{otherwise} \end{cases} \forall q, \tag{4.25}$$

$$W^A(t) = \sum_{q=1}^{Q} x_q(t), \text{ where } x_q(t) = \begin{cases} 1, \text{if } t_q^W = t \\ 0, \text{otherwise} \end{cases} \forall q, \tag{4.26}$$

$$\sum_{j=1}^{J} x_{ij} = A_i - [\sum_{t \le T} C_i^R(t)] - [\sum_{t \le T} N_i^R(t)] + A_i^O(T) + A_i^W(T), \text{ where } x_{ij} = \begin{cases} 1, \text{if } P_j^A = i \\ 0, \text{otherwise} \end{cases} \forall i, j, \tag{4.27}$$

$$r^O(t) = \sum_{p=1}^{P} y_p(t), \text{ where } y_p(t) = \begin{cases} 1, \text{if } t_p^O = t \\ 0, \text{otherwise} \end{cases} \forall k, \tag{4.28}$$

$$\sum_{i=1}^{N} a_{iq}(t) \le 1 \quad \forall q, \tag{4.29}$$

$$\sum_{q=1}^{W^T(t)} a_{iq}(t) = r_i^W(t), \text{ where } W^T(t) = \sum_{t' \le t} W^A(t'), \tag{4.30}$$

$$\sum_{t' \le t} \sum_{i=1}^{N} a_{iq}(t') \ge \sum_{t' \le t} \sum_{i=1}^{N} a_{iq'}(t'), \text{ where } \forall q < q', t_q^W + W_q^W \le t \text{ and } t_{q'}^W + W_{q'}^W \le t, \tag{4.31}$$

$$\sum_{t' \le t} \sum_{i=1}^{N} a_{iq}(t') \le 1 \quad \forall q, \tag{4.32}$$

$$Y_i = \{[\sum_{t' \le t} C_i^R(t') + N_i^R(t')] - A_i - [A_i^O(t) - N_i^O(t)] - A_i^W(t) + 1\} \cdot I_i(t) \quad \forall i, \tag{4.33}$$

$$\sum_{q=1}^{W^T(t)} a_{iq}(t) \ge \sum_{q=1}^{W^T(t)} a_{i'q}(t), \text{ where } \forall i < i', Y_i = 1 \text{ and } Y_{i'} = 1, \tag{4.34}$$

$$P_j^A \le P_{j+1}^A \quad \forall j, \tag{4.35}$$

$$0 \le t_p^O \le t_{p+1}^O \le T \quad \forall p, \tag{4.36}$$

$$0 \le t_q^W \le t_{q+1}^W \le T \ \forall l, \tag{4.37}$$

The objective function 4.1 is to minimize the average cost per unit of patient accessibility. The total cost is weighted sum of patient waiting time, provider idle time, and provider overtime, while the total patient accessibility number is the total number of Type 2 and Type 3 patients admitted. Eq. 4.2 – Eq. 4.37 are the constraints. Specifically, constraints 4.2 – 4.4 are the definitions of patient waiting time, real appointment start time, and real appointment ending time; constraints 4.5 ensure that a slot can be pre-booked with at most 2 patients; constraints 4.6 – 4.7 ensure that the number of no shows and cancellations cannot exceed the number of scheduled appointment; constraints 4.8 – 4.12 enforce the definition of cancellation and no-show; constraints 4.13 – 4.14 are the definition of $A_i^O(t)$ and $A_i^W(t)$; constrains 4.15 – 4.17 enforce the relationship between $A_i^O(t)$ and $A_i^W(t)$; constraints 4.18 – 4.20 enforce the condition of Type 2 and Type 3 patient admission; constraint 4.21 comes from the definition of $r^O(t)$; constraints 4.22 – 4.24 enforce the relationships among $W(t)$, $W^L(t)$, $r_i^W(t)$ and $W^A(t)$; constraints 4.25 – 4.26 are the definition of $W^L(t)$ and $W^A(t)$; constraints 4.27 indicate that the number of patient seen by the provider in slot *i* equals the number of patients scheduled slot *i* minus the number of no-shows and cancellations; constraints 4.28 is the definition of $r^O(t)$; constraints 4.29 – 4.30 are the definition of $a_{iq}(t)$; constraints 4.31 – 4.34 enforce the rules for walk-in patient admission, i.e., walk-in patients should be scheduled into the first available slots at the time of arrival, while the re-opened slots should be given to the walk-in patient with the longest waiting in the clinic; constraints 4.35 – 4.36 are the definitions of P_j^A, t_p^O and t_q^W, respectively. They generally specify the sequence of patients seen by the provider, the sequence of Type 2 patient request arrivals and the sequence of Type 3 patient request arrivals.

4.2. Solution Approach

As indicated above, the solution to the model is the optimal scheduling template. Given an appointment schedule, the corresponding objective value (average cost per unit of patient accessibility) is determined by a complex stochastic process, which includes innumerous uncertainties, such as random service time, patient no shows, and appointment cancellations. Hence, it is challenging to develop an analytical solution, which establishes the relationship between a scheduling template and the corresponding objective value. However, the discrete event simulation (DES) can provide us with the benefit of mimicking the clinic process, once a scheduling template is given. Thus, it provides us the way to estimate the objective value for a given appointment schedule.

On the other hand, the solution space increases exponentially with the number of the appointment slots in the template. Unfortunately, even a small clinic usually has a scheduling template of more than 10 appointment slots on a clinical day. This makes it impossible to run an exhausted search on the solution space with the objective value of each candidate solution determined by the discrete event simulation. To address this, the genetic algorithm (GA) can be adopted to efficiently guide the solution searching process.

Based on the above understanding, we develop an approach combining discrete event simulation and genetic algorithm in this study. The framework of the DES-GA approach is illustrated in Fig. 4.2. The genetic algorithm is used to efficiently search the solution space, while the discrete event simulation is used to estimate the objective value, which is the average cost per unit of patient accessibility, for a given clinic scheduling template. The entire DES-GA procedure is coded in-house in Matlab^{TM} environment. In the following, the DES-GA approach is discussed in detail.

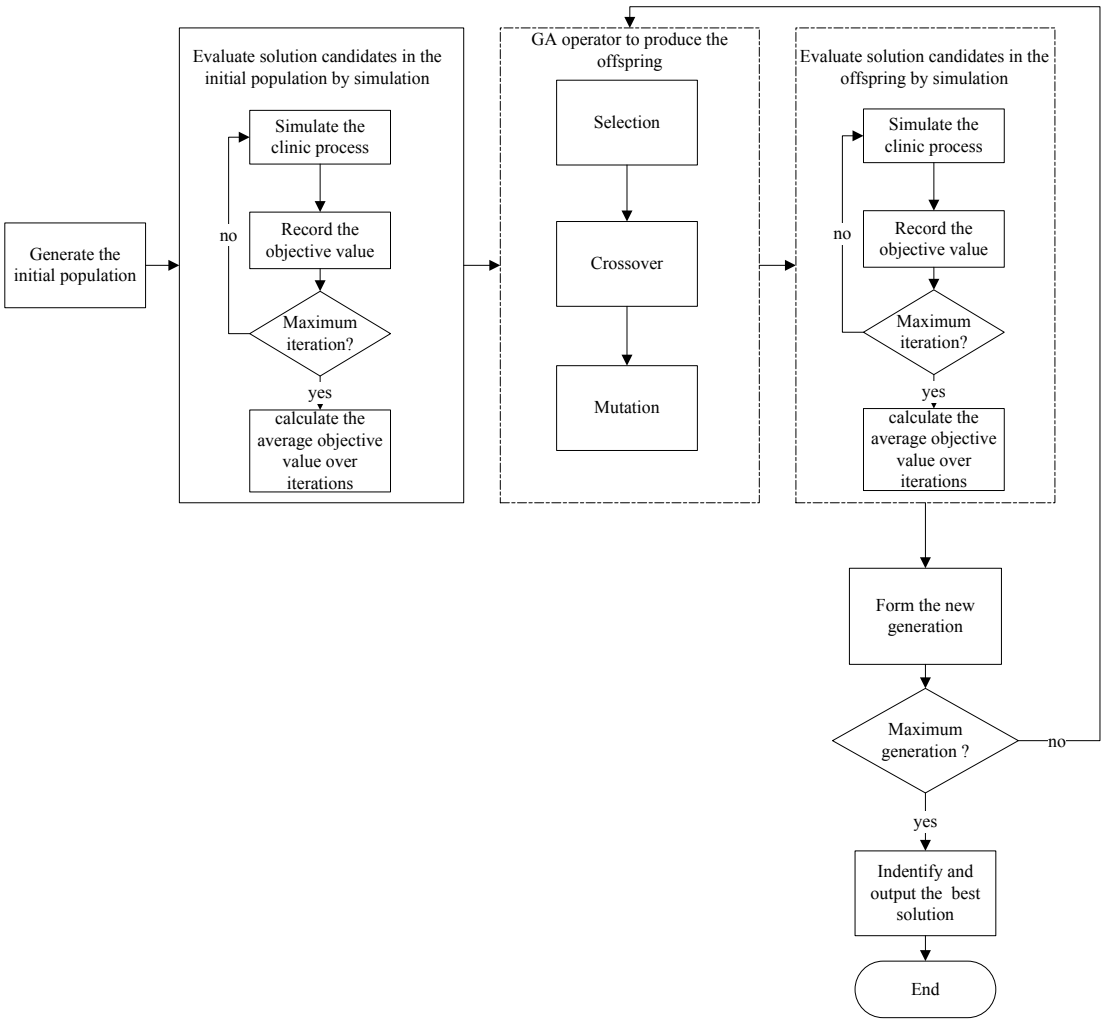

Fig. 4.2: Framework of the DES-GA approach

4.2.1. Representation

The candidate solution is represented as a chromosome, which is a string consisting of N integer numbers. The values of the integer numbers can only be 0, 1 or 2, which indicate the number of pre-booked Type 1 patients in the corresponding slot. Fig. 4.3 shows an example of the candidate solution, represented as a chromosome. In this example, an appointment scheduling template of 8 slots is represented as the chromosome. The 1^{st}, 7^{th}, and 8^{th} slots are

single pre-booked with Type 1 patients; the 3rd and 6th slot are double pre-booked with Type 1 patients; the 2nd, 4th and 5th slots are reserved for Type 2 and Type 3 patients.

1	0	2	0	0	2	1	1

Fig. 4.3: A chromosome example

4.2.2. Initialization

In initialization, the initial population of p chromosomes is generated, namely, $\mathbf{x}^{(0)}(1)$, …, $\mathbf{x}^{(0)}(p)$. To generate a chromosome that properly represents an appointment scheduling template, we randomly assign 0, 1, and 2 to the slots which together make up the chromosome. Furthermore, the parameters for the genetic algorithm are also initialized, which include p_c, the subpopulation size for crossover, β, the point mutation probability, n_1 and n_2, the sample sizes of discrete event simulation, and $g_{\max}$, the generation limit. Note that n_1 is the small sample size used in simulation before the maximum generation is reached, which n_2 is the large sample size used at the last generation for identifying the best solution.

4.2.3. Evaluation of the solution candidates by discrete event simulation (DES)

According to the approach for advancing the simulation clock, the discrete event simulation can be divided into two types, which are time-driven clock simulation and event-driven simulation. The time-driven simulation advances the time by a fixed time increment, while the event-driven simulation uses the next event time to advance the simulation clock. The advantage of the event-driven simulation is that the inactive period between two closest events can be skipped and the causality is guaranteed in the simulation. In this paper, the event-driven simulation is programmed due to such benefits. A flowchart of the discrete event simulation is shown in Fig. 4.4.

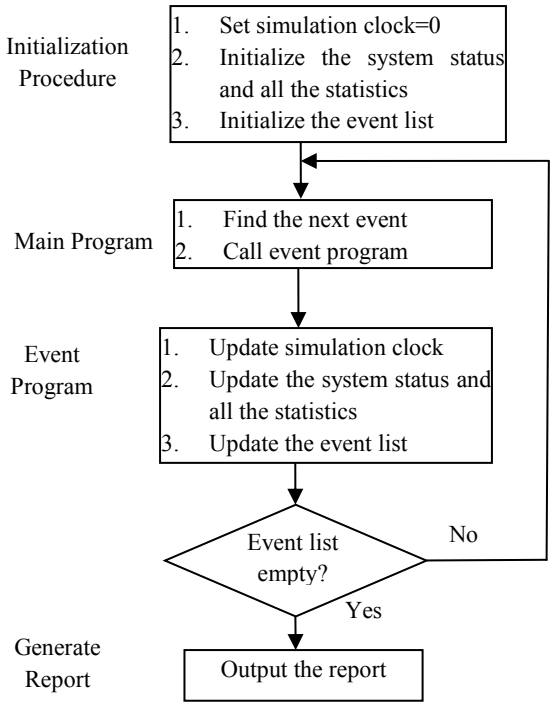

Fig. 4.4: Simulation flowchart

In the initialization procedure of simulation, Type 1 patients will be generated for a given schedule. The status of patients (on time, cancellation and no show) will be randomly assigned based on the given probability. Type 2 and Type 3 patients will also be created based on the given arrival rates. After this, an event list will be formed with all the events arranged in time ascending order. In the following we provide the details for the initialization procedure:

1) Obtain A_i from the given appointment schedule and set $p=1$, $q=1$ and j=1

2) Generate P, Q, t_p^O, t_q^W and W_q^W.

3) Create no-show and cancellation events for both Type 1 and Type 2 patients.

4) Create the event list

5) Set system clock to 0

In this study, the events on the event list are classified into 12 categories, including Type 1 patient on time, cancellation, no show, starting consultation and ending consultation, Type 2 patient request, on time, cancellation, no show, starting consultation and ending consultation, Type 3 patient arrival, left, admission, starting consultation and ending consultation. Note that the starting and ending consultations of Type 1, Type 2 and Type 3 patient belong to the same category. Each type of event would trigger the corresponding activities. To better illustrate how the event is processed in our discrete event simulation program, we provide the details of procedure for handling events from each category in the APPENDIX A.

4.2.4. Selection, crossover and mutation operations

Selection, crossover and mutation are the GA operators for generating offspring. The selection operator selects p_c chromosomes for crossover operation from the population of current generation by using the roulette-wheel rule. The probability that a chromosome will be chosen is proportional to the fitness value of the chromosome. The fitness value of the chromosome is determined of its corresponding rank in the population, where all chromosomes are ranked in a descending order by their estimated expected objective function value from the simulation procedure. Given the rank of chromosome in the population, the corresponding fitness value is defined by Eq. (4.38), where p is the population size (Pohlheim, 2006).

$$fitness(rank) = 2 \times \frac{rank - 1}{p - 1} \tag{4.38}$$

In the crossover operation, the selected p_c chromosomes, which are also known as parents, are divided into $p_c/2$ pairs, and a two-point crossover is executed on each pair to generate p_c new chromosomes. After crossover, a mutation operator will be executed on the newly generated chromosomes and form the offspring. In the mutation operation, each number on the chromosome may change with probability β, also known as the mutation rate. Note that the

numbers constituting chromosome can only be 0, 1, or 2. Hence, if the mutation happens, the changed number cannot take the values other than 0, 1, or 2.

4.2.5. Formation of new generation

The new generation is formed by replacing the chromosomes in the current population with the newly generated offspring. To be specific, chromosomes in the current population are ranked by their fitness values in a descending order. Then the last p_c chromosomes will be replaced by the p_c offspring, i.e., the chromosomes with low fitness value in the current population will be replaced by the offspring. The surviving chromosomes from the current population and the offspring together constitute the population of a new generation.

4.2.6. Identification of the best solution

After the maximum generation is reached, the discrete event simulation procedure is run again to better estimate the expected objective values of all chromosomes in the final generation. Unlike the small sample size n_1 used in the previous generations, the large sample size n_2 will be applied for indentify the best solution in the final generation. Note that n_2 is selected to ensure that the standard error on the expected objective value of each chromosome is small enough to realize statistically significant separations among the chromosomes, while n_1 is chosen to balance the computation efficiency and the solution quality, since as the sample size n_1 increases, the procedure is more computationally expensive but is likely to converge to better solutions.

4.3. Case Study

In this section, we report a case study to demonstrate how the proposed solution approach works in finding the heuristic optimal scheduling template for open access clinics that admit walk-in patients. The results of the case study are analyzed to identify the patterns of heuristic optimal scheduling templates. Furthermore, a sensitive analysis is conducted to study how clinic settings influence the optimal scheduling templates.

4.3.1. Experimental design

In order to verify the effectiveness of our proposed solution approach, we design 9 different scenarios for the case study. In each scenario, we consider a clinic with one provider offering sixteen 30-minute appointments per day. The total length of a clinic day is 480 minutes (8 hours). The parameters used in the case study are chosen based on the data collected in an outpatient clinic in a local hospital and the statistical data in the literature. In the following, we summarize the related literature, as well as the parameter choices in our case study:

First of all, regarding the non-attendance rate of Type 1 patients (which include no-shows and cancellations), the literature report the following,

- Johnson et al. (2007) indicate that the no-show rate vary from 3% to 42%, with an average of 17%.
- 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%.
- Al-Shammari (1992) and Hermoni et al. (1990) report non-attendance rates of 29.5% and 36%, respectively.
- Moore et al. (2001) suggest that no-shows and cancelled appointments combined amount 31.1% of appointments.

In the case study, we consider three levels of patient attendance rate for Type 1 patients, namely, high attendance rate (on-time arrival: 95%, no-show: 3%, cancellation: 2%), medium attendance rate (on-time arrival: 70%, no-show: 17%, cancellation: 13%), and low attendance rate (on time: 45%, no-show: 42%, cancellation: 13%). Note that in the case of medium attendance rate, we considered 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.

Secondly, regarding the non-attendance rate of Type 2 patients, the literature report the following,

- Open access scheduling has the ability to reduce patient no-show rate as well as the cancellation rate, when compared with the traditional scheduling system (Affiliated Computer Services, 2003, Lee and Yih, 2010).
- Kopach et al. (2007) provide an estimate of the no-show rate of open access patients, which is 50% less than that of the pre-booked patients.

In the case study, we assume that both the no-show rate and cancellation rate of Type 2 patients are 50% less than those of Type 1 patients. Based on this assumption, we also consider three levels of attendance rate for Type 2 patients, namely, high attendance rate (on-time: 97.5%, no-show: 1.5%, cancellation: 1%), medium attendance rate (on-time: 85%, no-show: 8.5%, cancellation: 6.5%), and low attendance rate (on-time: 72.5%, no-show: 21%, cancellation: 6.5%).

Thirdly, regarding the arrival rates of Type 2 and Type 3 patients, the literature reports the following,

- LaGanga and Lawrence (2009) propose to model the mean arrival rates of open access requests and walk-in patients as some fraction of the clinic capacity (e.g., 50%)

In the case study, the capacity of the clinic is assumed to be two patients per hour and we consider three levels of arrival rate for both Type 2 and Type 3 patients, namely, low arrival rate (1 per 2 hour, 25% capacity), medium arrival rate (1 per hour, 50% capacity) and high arrival rate (2 per hour, 100% capacity).

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.

Other parameters need to be determined as well, which include consultation time distribution, cancellation time distribution and the time that a walk-in patient is willing to wait before he/she obtains a same-day appointment. The parameter values are obtained based on our observations by working with a local clinic. Also, the parameters for genetic algorithm, shown in Table 4.3, are selected through trial runs.

Table 4.3: Parameters for the DES-GA approach

Notation	Description	Value
β	Mutation probability	0.01
$g_{\max}$	Iteration limit	150
n_1	Sample size for interior sampling	200
n_2	Sample size for the comparison of the last-generation chromosomes	1000
p	Population size	100
p_c	Population size of subdivision for crossovers	50

Based on the above discussion, we design 9 scenarios, namely Case 0 – Case 8, with Case 0 being the base case. In the base case, we generally adopt the moderate values for the parameters. To be specific, we consider medium attendance rates for both Type 1 patients (on-time: 70%, no-show: 17%, cancellation: 13%) and Type 2 patients (on-time: 85%, no-show: 8.5%, cancellation: 6.5%), and medium arrival rate (1 patient per hour on average) for both open access appointment requests and walk-in patients. Note that Type 2 and Type 3 patient arrivals

are modeled by the inter-arrival time. In the case of the medium arrival rate, the inter-arrival time (minutes) follows an exponential distribution with the rate of 1/60. In addition, the cost coefficients 1:5.2:7.8 are used in the model, which correspond to the 50th percentiles of national hourly wage as mentioned above. Other parameters in Case 0, such as cancellation time, consultation time, and "willing-to-wait time", are chosen based on our observations in a local clinic. To be specific, the cancellation time distribution suggests that a patient may cancel appointment at any time before the scheduled appointment time with equal chance; the consultation time distribution shows that a provider could finish seeing a patient within 20-30 minutes; the parameter "willing to wait" explains how long a walk-in patient is willing to wait before he/she can obtain a same-day appointment. In Case 0, the distribution of "willing to wait" suggests a patient is willing to wait 0-120 minutes in order to get a same-day appointment. All parameters for the base case are in shown in Table 4.4.

Table 4.4: Model parameters for the base case

<table>
<tr><th colspan="2">Parameters</th><th>Rate/Distribution</th><th>Parameters</th><th>Rate/Distribution</th></tr>
<tr><td rowspan="2">On time rate</td><td>Type 1 patient</td><td>70%</td><td rowspan="2">Inter-arrival time of type2 patient (in minutes)</td><td rowspan="2">Exponential(1/60)</td></tr>
<tr><td>Type 2 patient</td><td>85%</td></tr>
<tr><td rowspan="2">No show rate</td><td>Type 1 patient</td><td>17%</td><td rowspan="2">Inter-arrival time of type3 patient (in minutes)</td><td rowspan="2">Exponential(1/60)</td></tr>
<tr><td>Type 2 patient</td><td>8.5%</td></tr>
<tr><td rowspan="2">Cancellation rate</td><td>Type 1 patient</td><td>13%</td><td rowspan="2">Willing-to-wait time of type 3 patients (in minutes)</td><td rowspan="2">Uniform(0,120)</td></tr>
<tr><td>Type 2 patient</td><td>6.5%</td></tr>
<tr><td rowspan="2">Cancellation time of i^{th} appointment</td><td>Type 1 patient</td><td rowspan="2">Uniform(0,(i-1)T_{Δ})</td><td>c_1</td><td>1</td></tr>
<tr><td>Type 2 patient</td><td>c_2</td><td>5.2</td></tr>
<tr><td colspan="2">Consultation time per appointment (in minutes)</td><td>Uniform(20,30)</td><td>c_3</td><td>7.8</td></tr>
</table>

As the base case (i.e., Case 0) represents the moderate situation, we develop 8 other cases in order to show the proposed model would work under various situations and how the parameter selection influences the optimal scheduling templates. Compared with 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.

- Cases 1 & 2 represent the situations of high attendance rate and low attendance rate, respectively, by altering the on-time rate, no-show rate and cancellation rate, simultaneously.
- Cases 3 & 4 represent the situations of high arrival rate and low arrival rate of Type 2 patients, respectively, by altering the inter-arrival time distribution of Type 2 patients.
- Cases 5 & 6 represent the situations of high arrival rate and low arrival rate of walk-in patients, respectively, by altering the inter-arrival time distribution of walk-in patients.
- Cases 7 & 8 illustrate the situation of provider seeing low-income patients and high-income patients, respectively, by altering the cost coefficient.

The altered parameters for Cases 1-8 are shown in Table 4.5. 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 90^{th} percentile of national hourly wage. However, all other parameters remain the same as Case 0.

4.3.2. Case study results

The proposed solution approach is used to find the best heuristic scheduling template for each case presented in Section 5.1. The DES-GA procedure is run on a personal computer with an Intel 2.67GHz i5 dual-core processor and 2.9GB RAM. It takes 4.5 hours to find the heuristic optimal scheduling templates for all 9 cases using one CPU core, and thus the average computation time for one case is 30 minutes. The best scheduling templates found for the nine

cases are shown in Table 4.6, and the descriptive performance statistics of each scheduling template are also presented.

Table 4.5: Parameter adjustments for Cases 1-8 compared with Case 0

Case number	Parameter		Rate/Distribution
Case 1 (high attendance rate)	On time rate	Type 1 patient	95%
		Type 2 patient	97.5%
	No show rate	Type 1 patient	3%
		Type 2 patient	1.5%
	Cancellation rate	Type 1 patient	2%
		Type 2 patient	1%
Case 2 (low attendance rate)	On time rate	Type 1 patient	45%
		Type 2 patient	72.5%
	No show rate	Type 1 patient	42%
		Type 2 patient	21%
	Cancellation rate	Type 1 patient	13%
		Type 2 patient	6.5%
Case 3 (high requesting rate from type 2 patient)	Inter-arrival time of type2 patient/min		Exponential(2/60)
Case 4 (low requesting rate from type 2 patient)	Inter-arrival time of type2 patient/min		Exponential(1/120)
Case 5 (high requesting rate from type 3 patient)	Inter-arrival time of type3 patient / min		Exponential(2/60)
Case 6 (low requesting rate from type 3 patient)	Inter-arrival time of type3 patient / min		Exponential(1/120)
Case 7 (low-income patients)	c2		10
	c3		15
Case 8 (high-income patients)	c2		2.1
	c3		3.1

The heuristic optimal scheduling templates in Table 4.6 show that for each case, a portion of slots are reserved for Type 2 and Type 3 patients (i.e., open access and walk-ins) in order to achieve the best objective value. Furthermore, it is also shown that early appointment slots in a day should be scheduled with Type 1 patients, while the late appointment slots should be reserved for Type 2 and Type 3 patients. Table 4.6 indicates that the 1st slot is pre-booked with Type 1 patient in all cases, while the last 7 slots are always reserved for Type 2 and Type 3 patients. This is because early reservations are more likely to be wasted, since appointment requests from Type 2 and Type 3 patients may not arrive during the early reservation times.

However, the late reservations for Type 2 and Type 3 patients could possibly accommodate the appointment requests, as long as the patients arrive before the starting time of these slots.

Table 4.6: Best scheduling templates found for Cases 0 –8

slot index	Case 0	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8
1	2	2	2	2	2	2	2	2	2
2	2	1	2	1	1	0	1	2	1
3	1	1	2	2	2	2	2	2	1
4	2	1	0	1	2	0	2	1	1
5	1	1	2	0	1	0	1	0	2
6	0	0	0	0	0	0	0	2	0
7	0	2	2	0	1	0	1	0	0
8	0	0	2	2	1	0	2	2	0
9	2	0	0	0	2	2	2	0	0
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0
# of reserved appt.	10	10	10	11	8	13	8	10	11
# of double booking	4	2	6	3	4	3	5	5	2
# of single booking	2	4	0	2	4	0	3	2	3

Meanwhile, it is clear that the heuristic optimal scheduling templates in Table 4.6 are different among the nine cases. For instance, Cases 0, 1, and 2 have a medium patient attendance rate, high patient attendance rate and low patient attendance rate, respectively. The heuristic optimal scheduling templates of the three cases have the same number of slots reserved for Type 2 and Type 3 patients. However, the numbers of double booking are different: the least double booking slots in Case 1 versus the most double booking slots in Case 2. This observation validates that the double booking policy can be used as a solution to mitigating the adverse effect of low attendance rate (i.e., high patient no-show and cancellation rates). However, the attendance rate will not affect the number of slots reserved for Type 2 and Type 3 patients.

Similarly, analysis can be conducted regarding how Type 2 patient request arrival rate, Type 3 patient request arrival rate, and cost coefficients affect the number of slot reservations

and double booking in a scheduling template. The results are shown in Figs. 4.5 and 4.6. Fig. 4.5 indicates that the number of slots reserved for Type 2 and Type 3 patients decreases with the decrease of Type 2 and Type 3 request rates, while the attendance level and cost coefficient ratios do not significantly affect the number of reserved appointments. On the other hand, Fig. 4.6 shows the number of double booking slots is negative linearly correlated with the patient attendance level and positively correlated with the cost coefficient ratio. When the attendance rate is high, double booking is likely to result in enormous patient waiting. Thus, less double booking is preferred in this situation. When the attendance rate is low, double booking can increase the probability that an appointment slot will not be wasted due to patient no-shows and cancellations. In addition, high Type 3 patient arrival rates might mitigate the adverse effect of no-shows and cancellations, since Type 3 patients usually wait in clinic for a while for the slots re-opened due to cancellations or no-shows. As such, high Type 3 patient arrival rates could reduce the number of double booking slots. At last, the cost coefficient ratio reflects the relative difference of time values between the patients and the provider. When the ratio is lower, the patients' time is relatively more valuable. Therefore, they 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.

In Table 4.7, the objective value measures the average cost per unit of patient accessibility. The term "cost" indicates the total cost, which is the weighted summation of patient waiting time, provider idle time and provider, under the corresponding heuristic optimal scheduling template.

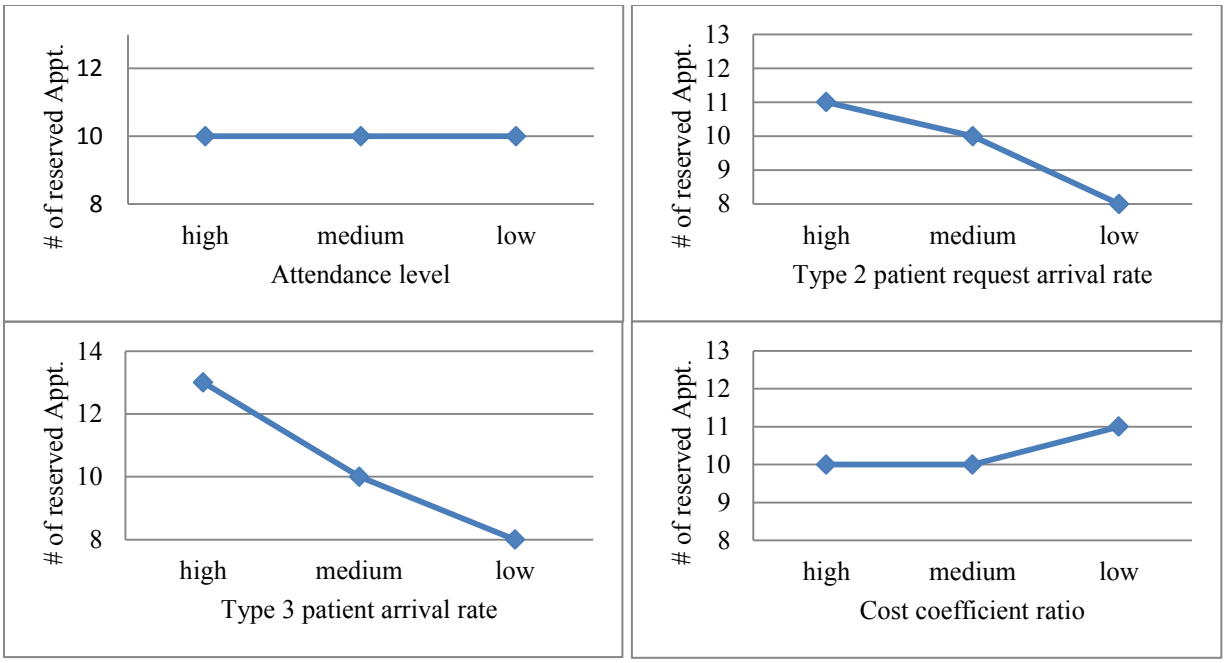

Fig. 4.5: Number of reserved appointment slots with respect to parameter selections

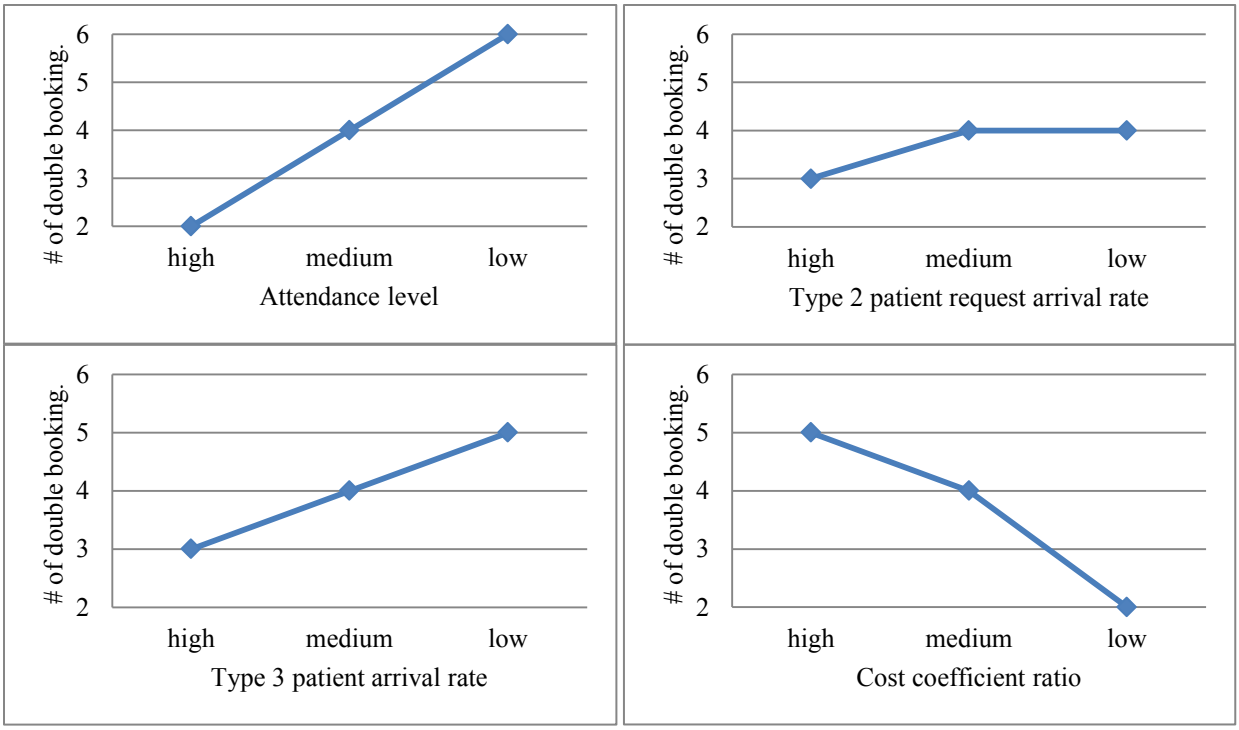

Fig. 4.6: Number of double booking slots with respect to parameter selections

The term “same-day appt & walk-in” represents the number of Type 2 and Type 3 patients seen during a clinic day. “Patient waiting”, “provider idle” and “provider overtime” record the total patient waiting time, provider idle time and provider overtime, respectively. In this paper, each heuristic optimal scheduling template is simulated 1,000 times to estimate its performance. The mean of performance metrics as well as the standard deviation of mean are presented in Table 4.7. 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 settings except for the requesting rate (demand) from Type 2 patients, where the requesting rate is higher for Case 3. As a result, Case 3 has a lower objective value than Case 4. This supports the general concept that high same-day appointment demand is preferred for open access clinics. A similar observation can be made by comparing Case 5 and Case 6, where Case 5 has a lower objective value and a higher arrival rate of Type 3 patients (demand) compared with Case 6. In addition, 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 4.7: Summary of descriptive performance statistics for Cases 0- 8

		Case0	Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case8
Objective	mean	59.8	42.1	83.9	48.7	76.1	37.2	91.2	94.4	30.2
	std of mean	0.94	0.57	1.27	0.63	1.35	0.47	1.56	1.40	0.49
Cost	mean	540.2	393.8	723.5	501.3	570.0	458.0	618.0	866.3	286.2
	std of mean	5.60	3.58	6.92	4.77	5.81	4.27	6.61	10.16	2.58
Same-day appt & walk-in	mean	9.49	9.67	9.14	10.60	8.00	12.61	7.29	9.51	10.14
	std of mean	0.04	0.03	0.04	0.03	0.04	0.03	0.04	0.03	0.04
Patient waiting/min	mean	172.7	143.4	84.5	116.5	177.2	129.8	219.5	284.8	65.8
	std of mean	4.98	1.80	3.67	3.20	4.74	3.64	5.80	7.77	1.89
Provider idle/min	mean	69.8	48.1	122.5	73.8	74.6	62.7	73.9	54.1	105.0
	std of mean	1.40	0.88	1.66	1.24	1.43	1.17	1.56	1.29	1.45
Provider overtime/min	mean	0.55	0.02	0.29	0.14	0.65	0.24	1.85	2.69	0.00
	std of mean	0.10	0.01	0.07	0.04	0.11	0.05	0.20	0.26	0.00

5. CONCLUSION

In this study, we implement the approach of integrated simulation and genetic algorithm to develop optimal scheduling template for both "traditional" and "open access" scheduling systems with unique features that have not been well addressed in literature. This effort indicates that this approach is a promising and powerful solution methodology for complex stochastic programing problems. The potential of commercializing this approach for appointment scheduling optimization cannot be underestimated.

Under the setting of traditional scheduling system, we propose a two-phase approach for designing a weekly scheduling template in an outpatient specialty clinic providing services of multiple types. These service types are clustered into several categories so that no substantial changeover time is incurred between any services in the same category. In the first phase of our approach, an MILP model is formulated to assign service categories to clinic sessions and determine the optimal number of appointments reserved for each service type in each clinic session. In the second phase, an SMIP is formulated to allocate appointments into equal-length time slots in each clinic session. To solve the SMIP, we develop a genetic algorithm embedded with Monte Carlo sampling. In our future research, we are interested in developing exact SMIP solution methods and sophisticated simulation optimization approaches with controllable performance guarantee.

For the design of outpatient scheduling template, it is likely to achieve higher-quality solutions when integrating the MILP model and the two-stage SMIP model. However, such an integrated model would lead to a stochastic optimization problem of much larger scale and the solution approach would limit the interactions with clinic managers in real-world practice. In our

future research, we will investigate whether such an integrated model could find weekly scheduling templates that significantly improve the performance metrics.

In the case study, we test the proposed two-phase approach for scenarios with different levels of patient demand and different weighting coefficients on patient waiting time, provider idle time, and provider overtime. Our results demonstrate that the proposed two-phase approach can efficiently identify promising weekly scheduling templates for an outpatient clinic on an average personal computer. The best weekly scheduling templates found can significantly reduce the total patient waiting time while maintaining the provider idle time. Meanwhile, our results suggest that the sampling based solutions to the SIMP in Phase II become more sensitive to the weighting coefficients as the provider workload increases. Moreover, our results suggest that the patterns of the best weekly scheduling templates found are different between cases with low and high levels of workload. This observation implies that in order to improve the performance of their appointment scheduling systems, individual clinics need to design scheduling templates based on their service processing characteristics. The two-phase approach proposed in this study provides a quantitative tool for outpatient specialty clinics to design better scheduling templates. In the future, we plan to test our approach in other outpatient clinics around the nation to investigate the regional differences and the differences among medical specialties.

Under the setting of "open access" scheduling system, we propose a DES-GA approach to find the heuristic optimal scheduling templates for open access clinics that admit walk-in patients. The costs of patient waiting time, provider idle time, and provider overtime are adopted to form the cost function. The patient accessibility is measured by the number of same-day appointment patients and work-in patient admissions. The objective is to minimize the cost per unit of patient accessibility. The DES-GA approach employs discrete event simulation to

compare the solution candidates, while using genetic algorithm to guide the solution searching. Hence, it inherits the advantages of both simulation and genetic algorithm, which is capable of finding the heurist optimal solution among the huge solution space for complex non-linear programming problem. By using this approach, not only the optimal number/percentage of reserved appointments can be found, but also the optimal allocation of these reserved slots can be obtained.

To demonstrate the effectiveness of this approach, a case study is conducted. The model parameters are collected from literature or hospital observations, and the heuristic optimal scheduling templates are determined for a variety of cases that adjust model parameters within reasonable ranges. In the case study, it is demonstrated that the heuristic optimal scheduling templates could change under different clinic settings. The level of demands for same-day appointments and walk-in admissions significantly impact the number of slot reservations in the heuristic optimal scheduling templates, while the number of double booking slots are greatly affected by patient attendance rate, cost coefficient as well as the level of demands for walk-in admissions. Our results also reveal that the first and last pre-booked slots should be double-booked.

REFERENCES

Amiri, M., & Mohtashami, A. (2012). Buffer allocation in unreliable production lines based on design of experiments, simulation, and genetic algorithm. International Journal of Advanced Manufacturing Technology, 62(1-4), 371-383.

Affiliated Computer Services. (2003). Commander's Guide to Acccess Success. Retrieved from http://www.tricare.mil/tma/tai/downloads/CDRGuide4Jun03.pdf

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

Armstrong, B., Levesque, O., Perlin, J. B., Rick, C., Schectman, G., & Zalucki, P. M. (2005). Reinventing Veterans Health Administration: focus on primary care. Journal of Healthcare Management, 50(6), 399-408.

Ahmadizar, F., Ghazanfari, M., & Fatemi Ghomi, S. M. T. (2010). Group shops scheduling with makespan criterion subject to random release dates and processing times. Computers & Operations Research, 37(1), 152-162.

Al-Khamis, T., & M'Hallah, R. (2011). A two-stage stochastic programming model for the parallel machine scheduling problem with machine capacity. Computers & Operations Research, 38(12), 1747-1759.

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

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.

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.

Batun, S., Denton, B. T., Huschka, T. R., & Schaefer, A. J. (2011). Operating room pooling and parallel surgery processing under uncertainty. INFORMS Journal on Computing, 23(2), 220-237.

Begen, M. A., & Queyranne, M. (2011). Appointment scheduling with discrete random durations. Mathematics of Operations Research, 36(2), 240-257.

Begen, M. A., Levi, R., & Queyranne, M. (2012). Technical Note—A Sampling-Based Approach to Appointment Scheduling. Operations Research, 60(3), 675-681.

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

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., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. Production and Operations Management, 12(4), 519-549.

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.

Cheu, R.L., Wang, Y. & Fwa, T.F. (2004). Genetic algorithm-simulation methodology for pavement maintenance scheduling. Computer-aided Civil and Infrastructure Engineering, 19(6), 446-455.

Centers for Medicare & Medicaid Services (2012) National Health Expenditure data. http://www.cms.hhs.gov/NationalHealthExpendData/downloads/highlights.pdf. Accessed March 2012

Dyke. (2010). Open access: Same-day appointments help patients and practices. Retrieved from http://www.bizjournals.com/kansascity/stories/2010/09/06/focus1.html

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

Denton, B., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. IIE Transactions, 35(11), 1003-1016.

Denton, B., Viapiano, J., & Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. Health care management science, 10(1), 13-24.

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

Francois, S. (2004). Health Care Delivery System in the United States. Retrieved from www.isye.gatech.edu/eemag/pdfs/20042Summer.pdf

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.

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.

Gholami, M., & Zandieh, M. (2009). Integrating simulation and genetic algorithm to schedule a dynamic flexible job shop. Journal of Intelligent Manufacturing, 20(4), 481-498.

Gul, S., Denton, B.T., Fowler, J. W., & Huschka, T. (2011). Bi-criteria scheduling of surgical services for an outpatient procedure center. Production and Operations Management Society, 20(3), 406-417.

Gupta, D., & Denton, B. (2008) Appointment scheduling in health care: challenges and opportunities. IIE Transactions, 40(9), 800-819.

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

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

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(2), 178-184.

Hartman, M., Martin, A. B., Benson, J., & Catlin, A. (2013). National Health Spending In 2011: Overall growth remains low, but some payers and services show signs of acceleration. Health Affairs, 32(1), 87-99.

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.

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

Huang, H., Kato, S., & Hu, R. (2012). Optimum design for indoor humidity by coupling Genetic Algorithm with transient simulation based on Contribution Ratio of Indoor Humidity and Climate analysis. Energy and Buildings, 47, 208-216.

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.

Izard, T. (2005). Improving patient care: Managing the habitual no-show patient. Farmily Practice Managment, 12(2), 65-66.

Institute of Medicine (2001) Crossing the quality chasm: A new health system for the 21st century, National Academy Press, Washington, DC

Jeong, S.J., Lim, S.J., & Kim, K.S. (2006). Hybrid approach to production scheduling using genetic algorithm and simulation. The International Journal of Advanced Manufacturing Technology, 28(1-2), 129-136.

Johnson, B.J., Mold, J.W. & Pontious, J.M. (2007). Reduction and management of no-shows by family medicine residency practice exemplars. Annals of Family Medicine, 5(6), 534-539.

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

Klein Haneveld, W. K., & van der Vlerk, M. H. (1999). Stochastic integer programming: General models and algorithms. Annals of Operations Research, 85, 39-57.

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

Klassen, K. J., & Yoogalingam, R. (2009). Improving performance in outpatient appointment services with a simulation optimization approach. Production and Operations Management, 18(4), 447-458.

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.

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., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., 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.

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

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.

Lin, C., Hsie, M., Hsiao, W., Wu, H., & Cheng, T. (2012). Optimizing the schedule of dispatching earthmoving trucks through genetic algorithms and simulation. Journal of performance of constructucted Facilities, 26(2), 203-211.

Mitchell, M. (1996). An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA

MathWorks, Inc. (2012) R2012a Documentation – Version 7.12 (R2011a) MATLAB Software. http://www.mathworks.com/help/techdoc/rn/bsru49a.html. Accessed August 2012

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., & 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.

Manbachi, M., Mahdloo, F., Ataei, A. & Haghifam, M.R. (2011). New interface for coordinated maintenance scheduling of generating units in neighbor countries applying genetic

algorithm and Monte-Carlo simulation. International Review of Electrical Engineering, 6(4), 1748-1756.

Ma, W.T., Cheu, R.L., & Lee, D.H. (2004). Scheduling of lane closures using genetic algorithms with traffic assignments and distributed simulations. Journal of Transportation Engineering, 130(3), 322-329.

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.

Medical Resource Group, 2011. Physician Salaries – Salary Survey Results. http://www.studentdoc.com/salaries.html

Nicoara, E.S., Filip, F.G., & Paraschiv, N. (2011). Simulation-based Optimization Using Genetic Algorithms for Multi-objective Flexible JSSP. Studies in Informatics and Control, 20(4), 333-344.

O'Keefe, R. M. (1985). Investigating outpatient departments: implementable policies and qualitative approaches. Journal of the Operational Research Society, 705-712.

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.

Oh, H.C., & Chow, W.L. (2011). Scientific Evaluation of Polyclinic Operating Strategies with Discrete-Event Simulation. International Journal of simulation model, 10(4), 165-176.

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.

Pohlheim, H. (2006). Version 3.8 of the GEATbx: Genetic and Evolutionary Algorithm Toolbox for use with Matlab. Retrieved at http://www.geatbx.com/docu/algindex-02.html.

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

Potter, B. & Seavecki, M. (2010). New Scheduling System Helps Wingra Clinic Reduce Patient No-Shows. Retrived from http://www.fammed.wisc.edu/our-department/newsletter/winter-2010/wingra-clinic-reduce-patient-noshows

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., 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.

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

Robinson, L. W., & Chen, R. R. (2003). Scheduling doctors' appointments: optimal and empirically-based heuristic policies. IIE Transactions, 35(3), 295-307.

Rohleder, T. R., Lewkonia, P., Bischak, D. P., Duffy, P., & Hendijani, R. (2011). Using simulation modeling to improve patient flow at an outpatient orthopedic clinic. Health Care Management Science, 14(2), 135-145.

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.

Santibáñez, P., Chow, V. S., French, J., Puterman, M. L., & Tyldesley, S. (2009). Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency's ambulatory care unit through simulation. Health Care Management Science, 12(4), 392-407.

Schultz, R. (2003). Stochastic programming with integer variables. Mathematical Programming, 97(1-2), 285-309.

Schimpff, S. (2012). Can You Get a Prompt Appointment With Your Doctor? Retrieved from http://medcitynews.com/2012/06/can-you-get-a-prompt-appointment-with-your-doctor/#ixzz2L6pn9VZX

Steinbauer, J.R., Korell, K., Erdin, J., & Spann, S.J. (2006). Implementing open-access scheduling in an academic practice. Family Practice Mangement, 13(3), 59-64.

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

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.

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.

US Department of Labor (2012) National Occupational Employment and Wage Estimates in the United States. http://www.bls.gov/oes/current/oes_nat.htm. Accessed June 2012.

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

World Health Organization (2000).The World Health Report 2000: Health Systems: Improving Performance. Retrieved from www.who.int/whr/2000/index.htm.

Yeh, J., & Lin, W. (2007). Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. Expert Systems with Applications, 32, 1073-1083.

Yen, J. W., & Birge, J. R. (2006). A stochastic programming approach to the airline crew scheduling problem. Transportation Science, 40(1), 3-14.

APPENDIX. 12 CATEGORIES OF EVENTS IN SIMULATION

***Category 1**: Type 1 patient on time*

1. update system time to the event time

2. check provider availability

if provider is free, then

mark the corresponding provider as seized

insert a corresponding "Type 1 patient starting consultation" event into the event list, with event time=system time

else

put the patient in waiting list

re-order the waiting list by patient appointment time in ascending order

end

3. remove this Type 1 patient on time event from the event list

4. re-order the event list by event time in ascending order.

***Category 2**: Type 1 patient cancellation*

1. update system time to the event time

2. check the availability of the slot after the cancellation

if available, then

check walk-in patient queue

if empty, then

mark the slot as open for scheduling

else

drop the walk-in patient with the longest waiting time from walk-in patient queue

delete the corresponding "Type 3 patient left" event from the event list

insert a corresponding "Type 3 patient on time" event into the event list, with event time = scheduled appointment starting time of the cancelled appointment

end

else

do nothing

end

3. remove this Type 1 patient cancellation event from the event list

4. re-order the event list by event time in ascending order.

***Category 3**: Type 1 patient no show*

1. update system time to the event time

2. check the availability of the slot after the no show

if available, then

check walk-in patient queue

if empty, then

do nothing

else

drop the walk-in patient with the longest waiting time from walk-in patient queue
delete the corresponding "Type 3 patient left" event from the event list
insert a corresponding "Type 3 patient on time" event into the event list, with event time = scheduled appointment starting time of the cancelled appointment
end
else
do nothing
end
3. remove this Type 1 patient no show event from the event list
4. re-order the event list by event time in ascending order.

***Category 4**: Type 2 patient request*
1. update system time to the event time
2. check the number of available appointment
if zero, then
do nothing
else
randomly select an available slot for the patient
randomly generate patient status (on time, cancellation or no show)
if on time, then
insert a corresponding "Type 2 patient on time" event into the event list, with event time = start time of the selected slots
else if cancellation, then
randomly generate a "cancellation time"
insert a corresponding "Type 2 patient cancellation" event into the event list, with event time = "cancellation time"
else
insert a corresponding "Type 2 patient no show" event into the event list, with event time = start time of the selected slots
end
end
3. $k = k + 1$
4. remove this "Type 2 patient request" event from the event list
5. re-order the event list by event time in ascending order.

***Category 5**: Type 2 patient on time*
1. update system time to the event time
2. check provider availability
if provider is free, then
mark the corresponding provider as seized
insert a corresponding "Type 2 patient starting consultation" event into the event list, with event time=system time

else
put the patient in waiting list
re-order the waiting list by patient appointment time in ascending order
end
3. remove this Type 2 patient on time event from the event list
4. re-order the event list by event time in ascending order.

***Category 6**: Type 2 patient cancellation*
1. update system time to the event time
2. check walk-in patient queue
if empty, then
mark the slot as open for scheduling
else
drop the walk-in patient with the longest waiting time from walk-in patient queue
delete the corresponding "Type 3 patient left" event from the event list
insert a corresponding "Type 3 patient on time" event into the event list, with event time = scheduled appointment starting time of the cancelled appointment
end
3. remove this Type 1 patient cancellation event from the event list
4. re-order the event list by event time in ascending order.

***Category 7**: Type 2 patient no show*
1. update system time to the event time
2. check walk-in patient queue
if empty, then
do nothing
else
drop the walk-in patient with the longest waiting time from walk-in patient queue
delete the corresponding "Type 3 patient left" event from the event list
insert a corresponding "Type 3 patient on time" event into the event list, with event time = scheduled appointment starting time of the cancelled appointment
end
3. remove this Type 1 patient no show event from the event list
4. re-order the event list by event time in ascending order.

***Category 8**: Type 3 patient arrival*
1. update system time to the event time
2. check the number of available appointment
if zero, then
put patient in the walk-in patient queue
insert a corresponding "Type 3 patient left" event to the event list, with event time = $W_l^W + t_l^W$
else

select the 1st available slot for the patient
insert a corresponding "Type 3 patient admission" event into the event list, with event time = start time of the selected slots
end
3. $l = l + 1$
4. remove this "Type 3 patient arrival" event from the event list
5. re-order the event list by event time in ascending order

***Category 9**: Type 3 patient left*
1. update system time to the event time
2. drop the patient from the walk-in patient queue
3. remove this "Type 3 patient left" event from the event list
4. re-order the event list by event time in ascending order.

***Category 10**: Type 3 patient admission*
1. update system time to the event time
2. check provider availability
if provider is free, then
mark the corresponding provider as seized
insert a corresponding "Type 3 patient starting consultation" event into the event list, with event time=system time
else
put the patient in waiting list
re-order the waiting list by patient appointment time in ascending order
end
3. remove this Type 3 patient admission event from the event list
4. re-order the event list by event time in ascending order.

***Category 11**: patient starting consultation*
1. update system time to the event time
2.mark the corresponding provider as seized
3. get the appointment information of the corresponding event and assign it to P_j^A
4. a*ssign current system time to* P_j^S
5. random generate service time P_j^D *according to the distribution.*
6. assign $P_j^S + P_j^D$ to P_j^E
7. insert a corresponding "patient ending service" event into the event list, with event time = P_j^E
8. j=j+1
9. remove this "patient starting consultation" event from the event list
10. re-order the event list by event time in ascending order.

***Category 12**: patient ending consultation*
1. update system time to the event time
2. mark the corresponding provider as free

3.check patient waiting list

If empty, then

do nothing

else

drop the patient ranked 1st from the patient waiting list

insert a corresponding "patient starting service" event into the event list, with event time=system time

end

5. remove this event "patient ending service" event list

6. re-order the event list by event time in ascending order.