Appointment overbooking with different time slot structures

Yan Chen\textsuperscript{a,} Yong-Hong Kuo\textsuperscript{b,}\textsuperscript{*,} Ping Fan\textsuperscript{c}, Hari Balasubramanian\textsuperscript{d}

\textsuperscript{a} Department of Decision Sciences, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau
\textsuperscript{b} Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong
\textsuperscript{c} Department of Logistics Management and Engineering, Zhuhai College of Jilin University, Zhuhai, Guangdong, China
\textsuperscript{d} Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst, 160 Governors Drive, Amherst, MA 01003, United States

\textbf{A R T I C L E  I N F O}

Keywords:
Appointment scheduling
Overbooking
Stochastic programming
Deterministic equivalent
Time slot structure

\textbf{A B S T R A C T}

Unattended appointments result in underperformance of a healthcare service provider. Such uncertainty in the appointment process not only lowers healthcare utilization and productivity but also hinders patients from having timely access to healthcare services and extends waiting time for receiving medical examinations or treatments. In this study, we formulate this stochastic optimization problem for appointment scheduling as its two-stage deterministic equivalent to simultaneously optimize overbooking and scheduling decisions to compensate patient no-shows with different time slot structures. We examine the impacts of three types of time slot structures, which are of fixed-length slot intervals, dome-pattern slot intervals and flexible appointment start times, on the efficiency of the system. With the optimal solutions found, we investigate the interaction between the time slot structure and the optimal overbooking solution. We found that the flexibility in appointment start times result in a “dome-dome-dome pattern with alternate long and short time slots” and could achieve a better patient experience (regarding the patient waiting time) while maintaining the same service provider efficiency (regarding the resource overtime and idle time), compared with the pre-defined time intervals. While there has been a large body of literature on appointment scheduling, to the best of our knowledge, this “dome-dome-dome” pattern has not been reported in the existing literature. Sensitivity analysis further shows that the flexible slot scheme is more effective when the number of overbooked patients is relatively low and the service duration is relatively long.

1. Introduction

The problem of patient no-shows, which is scheduled patients not showing up for their appointments, is prevalent in many health care settings. The no-show rates reported by Rust, Gallups, Clark, Jones, and Wilcox (1995), Sharp and Hamilton (2001), Dreher, Weitzman, Davdovici, Shapiro, and Cohen (2008), Defife, Conklin, Smith, and Poole (2010) exhibit high variability among different medical specialties and geographic regions, ranging from as low as 3\% to as high as 80\%. Unattended time slots result in medical resource idleness, leading to lower utilization and productivity. Therefore, improving the operational efficiency of service providers in the presence of patient no-shows attracts broad interest from researchers and practitioners.

Our research is motivated by a realistic appointment scheduling problem in a medical imaging center of a major hospital. Due to the increasing patient demand for medical diagnostic services, our project team was consulted by the hospital administrators to improve its operational efficiency. In addition to anticipating the growing demand, the center has also been suffering from patient no-shows and is looking for solutions to reduce unnecessary staff overtime and resource idle time. After analyzing their operational data, we were surprised to find out that the no-fasting rate (the percentage of patients who do not fast as instructed) is much higher than the no-show rate. In this case, these patients simply cannot proceed to the scheduled diagnostic tests, resulting in a low resource utilization. Although these patients were physically present at the center, they were also regarded as no-show patients as their scheduled appointment slots were wasted. Consequently, this high no-fasting rate has resulted in over 50\% of the total time slots remaining unattended. The management team also mentioned to us that text messages were issued to remind patients of fasting before attending their appointments. However, they revealed that this solution is still not effective.

Unattended time slots not only lead to low resource utilization and productivity but also prevent patients from receiving timely medical examinations and treatments (due to the long waiting list). Different from the time for physically waiting to receive medical examinations at
a healthcare facility, this waiting time between the referral date and the appointment date is kind of “indirect” and sometimes is ignored in related academic studies. However, in practice, this “indirect” waiting time is of particular importance and can be critical to an early diagnosis of diseases, which could result in more effective medical treatment. Therefore, such no-show phenomenon not only is an issue regarding the profit and cost for the healthcare organization and patient experience when attending their appointments, but also has a crucial impact on the effectiveness of health services provided to patients. To mitigate this adverse impact of patient no-shows, a popular method is overbooking schemes, where the booked patients are more than the number of available time slots. Overbooking is a regular practice in the airline industry adopted to increase its revenue. Muthuraman and Lawley (2008) point out that overbooking in healthcare is different from overbooking flights in the way that overserving patients must be treated by service providers, which will result in excessive workload and impaired patient experience.

Intuitively, clinics may want to counteract the unattended time slots by adopting double/triple booking. However, according to Kim and Giachetti (2006), this strategy may lead to unnecessary clinic overcrowding, increased staff overtime and prolonged patient waiting. This is particularly true if time slots are fixed in length. For example, patients may be assigned to 15-minute slots throughout the day. Fixed length slots are convenient from an administrative viewpoint, and patients are more likely to remember appointment times that are multiples of, say, 15-minute intervals. For example, a 9:15 am appointment is easier to remember than a 9:41 am appointment. However, the downside is that overbooking of time slots creates uneven appointment clusters increasing the chance of long waits.

An alternative is allowing time slots to vary in length. This provides greater flexibility in appointment start times and allows a practice to avoid overbooked clusters. Instead, appointments are distributed more evenly during a session. Flexible appointment start times are also better suited to situations where patients can be informed and reminded of more precise appointment times (for example 9:41 am) by apps installed on smart phones for appointment purposes.

With this tradeoff in mind, we first consider two common pre-defined slot length structures. One has fixed-length intervals, the most popular scheme in practice. The other has varying-length intervals shaped as a dome pattern, which is widely reported to work well and be robust under various conditions in Wang (1997), Robinson and Chen (2003), and Denton and Gupta (2003). Finally, we consider an entirely flexible time slot structure without a pre-defined pattern. The three structures can be regarded as a progressive relaxation of the fixed slot length constraint.

The main objective of this study is to identify optimal overbooking assignment and slot interval structure in response to the uncertainty caused by service time and patient no-show behavior. We develop two-stage stochastic mixed-integer linear programming models to identify the best overbooking solutions so as to minimize a weighted combination of three performance measures: staff overtime, resource idle time and patient waiting time. These performance metrics corresponding to the three optimal overbooking solutions are examined and compared for the medical imaging center which motivated this study. Finally, using a more general experimental design, we provide a detailed analysis of slot interval structures and overbooking.

The appointment scheduling literature is extensive, and the methods we use in this paper are similar to those that other researchers have used. However, as our literature review in the next section suggests, the research on the joint consideration of patient overbooking and optimal appointment slot structure appears to be inadequate. The scope of this paper is to investigate it in detail.

2. Literature review

This study lies in the research area of outpatient appointment scheduling. Appointment scheduling in healthcare is a popular research area and there have been extensive studies conducted to improve the productivity and service quality. For a comprehensive review, we refer readers to Cayirli and Veral (2003), Gupta and Denton (2008), and Ahmadi-Javid, Jalali, and Klasse (2017). In this section, we review the most related literature which considers patient no-shows and/or overbooking.

There are two streams of literature related to patient no-shows and overbooking: (i) studies accessing impacts of patient no-shows or different overbooking schemes; (ii) studies proposing specific overbooking schemes, such as overbooking levels, overbooking assignments, slot interval structures (by determining appointment times), and so on.

The first stream of literature includes studies examining the effects of patient no-shows or overbooking schemes. The initial efforts of Bailey (1952) investigate the effect of patient no-shows on patient waiting time and doctor idle time and conclude that patient no-shows impact the suitable choice of the appointment system. More recent studies also draw consistent conclusions. With a simulation study, Ho and Lai (1992) report that among three environmental factors – clinic size, no-show probability and variation of service times – no-show probability has the most significant impact on the performance of an appointment system. With a simulation model, Klassen and Rohleder (1996) compare different scheduling rules under various medical environments with patient no-shows. They conclude that the best scheduling rules depend on particular clinic goals and environmental conditions. In a later paper, Cayirli, Veral, and Rosen (2006) demonstrate, also with a simulation approach, that patient no-shows is an important factor influencing the performance and ultimate selection of an appointment system. With queuing models, Hassan and Mendel (2008) study the effects of patient no-shows on the performance of a single-server healthcare system and conclude that no-shows greatly affect the optimal schedule structure and should be taken into account when designing an appointment system. LaGanga and Lawrence (2007) examine the problem of no-shows and propose appointment overbooking as one means of reducing the negative impact of no-shows. By conducting simulation experiments, they show that patient access and provider productivity are significantly improved with overbooking, but that overbooking causes increases in both patient waiting time and provider overtime. Chen, Kuo, Balasubramanian, and Wen (2015) use simulation to examine the effect of different overbooking schemes and conclude that overbooking may not necessarily lead to the increase in the resource overtime and the way of assigning overbooked patients does have a significant impact on the operational efficiency of service providers.

The second stream of literature consists of studies proposing specific schemes/strategies of overbooking to improve system performance. Some studies focus on determining the optimal overbooking level, i.e. the optimal number of patients overbooked or booked for a given service session/day. Shonick and Klein (1977) estimate patient no-show probabilities based on patient characteristics and use such information to overbook appointments until the expected number of arrivals reaches a specified level. Kim and Giachetti (2006) use a stochastic overbooking model to determine the optimal number of appointments to be scheduled to maximize the total expected profits for the healthcare service provider under the assumption of stochastic arrivals and deterministic service times. Green, Savin, and Wang (2006) and Kolisch and Sickinger (2008) schedule different types of patients (outpatients with a fixed no-show probability, inpatients and emergency patients) to a given number of time slots of identical length dynamically using Markov decision process. Muthuraman and Lawley (2008) consider a sequential appointment scheduling problem with exponential service time and develop a myopic policy for scheduling call-ins, i.e., the service provider must schedule an appointment time (without considering the future requests) with the patient when an appointment request is received. Each calling patient has his own no-show probability and overbooking is used to compensate for patient no-shows. A stopping
rule is developed to determine when to stop booking more patients under the condition that the profit function is unimodal. Chakraborty, Muthuraman, and Lawley (2010) extend the sequential scheduling model with heterogeneous patients in Muthuraman and Lawley (2008) to the general service time setting. A later paper presented by Zeng, Turkcan, Lin, and Lawley (2010) proves that the profit function of Muthuraman and Lawley (2008) is multi-modular with homogeneous patients; but this may not be true for heterogeneous patients in a static setting. A search algorithm to find local optimal schedules is proposed for the latter case. Lin, Muthuraman, and Lawley (2011) continue the discussion in Muthuraman and Lawley (2008). Instead of a myopic policy, an approximate dynamic programming approach is proposed for this sequential scheduling problem. Liu, Ziya, and Kulkarni (2010) consider patient no-shows and cancellation when they propose dynamic policies for assigning an appointment date to each patient sequentially. With their simulation results, they conclude that the open access policy works better than other strategies when patient load is low. With the assumption that the service times are deterministic and the show-up rates vary according to the appointment time slot, LaGanga and Lawrence (2012) develop an analytic appointment scheduling model to investigate the near-optimal overbooked schedule obtained by complete enumeration and heuristic search. Luo, Kulkarni, and Ziya (2012) study an appointment scheduling problem with patient no-shows and emergent interruptions. They report that appointment scheduling policies can have a very poor performance if such interruptions are not considered. Liu and Ziya (2014) consider a panel size and overbooking decision where whether or not a patient shows up depends on the appointment queue length. The authors conclude that in addition to the magnitudes of the no-show probability, the patient sensitivity to delays is also important in this decision-making process.

Some studies are interested in finding out good approaches of implementing overbooking after the overbooking level is determined. Cayirli, Yang, and Quek (2012) summarized that there are two general approaches in literature to accommodate overbooking in an appointment system. One is to assign two or more patients into the same slot. The other one is to adjust appointment intervals proportionally. Kaandorp and Koole (2007) develop a local search procedure for outpatient appointment scheduling problem where no-shows may occur. They find that the optimal overbooking appointment schedule is similar to Baily-Welch rule reported in Bailey (1952) and Welch and Bailey (1952) in certain cases. A later paper by Cayirli et al. (2012) propose a universal appointment rule making use of parameterized dome patterns for general types of clinics, and an adjustment procedure to minimize the impacts of patient no-shows and walk-ins on the efficiency of the clinic. Later, Cayirli and Yang (2014) extend the work of Cayirli et al. (2012) to consider different classes of services further. With a multi-stage stochastic linear program, Erdogan and Denton (2013) extend the two-stage stochastic linear programming model proposed by Gupta and Denton (2008) to incorporate patient no-shows for determining optimal appointment times. Tsai and Teng (2014) develop a stochastic overbooking model to assign patients to time slots with equal length in one session with the goal of increasing the utilization of multiple resources in a medical facility such as a physical therapy room.

There are some other discussions on overbooking in the literature. For example, Robinson and Chen (2010) compare the performance of traditional appointment scheduling systems to open-access policies with naïve overbooking. They conclude that overbooking with open-access will significantly outperform traditional schedules in terms of a weighted average of waiting time, idle time and overtime. Later, Patrick (2012) also demonstrate with simulation experiments that combining open access with overbooking can mitigate the impact of no-shows. Alaeddini, Yang, Reddy, and Yu (2011) develop an integrated model of logistic regression and Bayesian inference to predict the no-show probability of a patient. Their procedure can be used for selective overbooking strategy and mitigation of patient no-show risk. Motivated by the increasingly prevalent electronic appointment booking systems, Feldman, Liu, Topaloglu, and Ziya (2014) propose a model for the service provider to dynamically decide which appointment days to make available for the patients taking into account patient preferences, cancellation of appointments, and no-shows. Liu (2016) apply queueing models to determine the optimal appointment scheduling window with the consideration of patient no-shows.

Although there has been a large amount of work done in examining patient no-shows and overbooking schemes, our research differs from the earlier studies from multiple aspects. First, different from the previous studies in overbooking, we propose a mathematical programming approach to consider both the overbooked appointments and slot interval length (by determining appointment time). Second, the overbooking pattern, i.e. where and when to book the overbooked patients, is explored based on the optimal solution attained by solving the stochastic optimization model. Varying time slot lengths and blocking/clustering features (number of patients booked for each time slot) are identified from the optimal solutions. Third, we incorporate the widely recognized dome pattern into the overbooking strategy. The interaction between the slot structure (i.e. fixed interval and variable interval) and the overbooking pattern are investigated in this paper. Finally, we conduct a comprehensive study with a sensitivity analysis to examine the effectiveness of the joint decisions of overbooking and scheduling schemes under different scenarios. Table 1 compares the characteristics of the most relevant papers about overbooking schemes with our research.

3. Mathematical modeling

3.1. Problem description

In this section, we present our proposed stochastic optimization models for scheduling patients in a specific session. In this paper, a session is a period of time (typically, a morning, an afternoon, or a day), during which the healthcare service provider is in continuous operation, until which the service provider temporarily suspends operations (a lunch break, end of the day, etc.). For the sake of administrative convenience, in practice, a service session is very often divided into multiple pre-defined time slots for appointment booking. For this reason, in Section 3.3, we first consider this appointment scheduling problem in this setting and propose a deterministic equivalent of a stochastic mixed-integer linear programming model, to assign a fixed number of patients to the pre-defined time slots of the session. A deterministic equivalent of a stochastic model considers a large number of realizations of the random events due to the stochastic nature of the problem. Each realization generated is called a scenario. In Section 3.4, with a modified version of this model, we examine a completely flexible session structure, where the intervals are not fixed, to determine the appointment times in another deterministic equivalent model. In our paper, each time slot is differentiated with its appointment start time. For each time slot, there can be one or more patients to be assigned; however, there is only one resource throughout the whole day and thus at most one patient can be seen at any moment. Thus, upon arrival, each patient has to wait to start the required examinations until the previous patient, if any, is done with his/her tasks. The time to finish all the required examinations for each patient follows a certain probability distribution, which is assumed to be known from historical data. For each patient assigned to a time slot, with pre-determined probabilities, he/she will either arrive at the facility at the scheduled appointment time or not show up. Hence, there are two dimensions of uncertainty in our model: the presence of patients, and the service durations. If the scheduled appointments cannot be finished within the regular hours of the session, the staff will still continue to perform all the required tasks and the working time beyond regular hours is regarded as overtime.

In this work, we propose deterministic equivalent models of the problem where a set of independent and identically distributed (i.i.d.)
Before we present our mathematical models, we first introduce the notation used in this paper. Our mathematical models require the following sets and parameters.

**Sets:**

- \( N = \) set of patients \([1, 2, \ldots, n]\) to be scheduled for the session, indexed by \(i\)
- \( J = \) set of time slots in the session, indexed by \(j\)
- \( S = \) set of scenarios, indexed by \(s\)
- \( N_i^s = \) set of patients who show up at the healthcare unit under scenario \(s\)
- \( N_i^c = \) set of patients who do not show up at the healthcare unit under scenario \(s\)

The total number of patients scheduled for one session, i.e., \(n\), is fixed in our models. In Section 4.3, we will discuss how we determine this number. In our models, the two dimensions of uncertainty – the presence of the patients and the service durations – are modeled by the realizations of their outcomes under different scenarios. In our model, we consider a finite number of scenarios, denoted by \(S\), to approximate those stochastic components. These realizations of these outcomes are assumed to be i.i.d. among the scenarios. Under each scenario, each patient either shows up or does not, according to the no-show rate.

**Parameters:**

- \( b_j = \) beginning time of time slot \(j\)
- \( d_i = \) service duration for the \(i^{th}\) patient under scenario \(s\)
- \( w^r_i = \) penalty for each unit of resource overtime
- \( w^{wait}_i = \) penalty for each unit of patient waiting time
- \( w^{idle}_i = \) penalty for each unit of resource idle time
- \( E = \) close time of the healthcare facility

When a service session is divided into pre-defined time slots, the beginning time \(b_j\) of each slot \(j\) is determined. For example, in the current practice of our motivating application, the slot interval is set to 20 min, which is approximately the total of the average setup time and the average examination time. Therefore, \(b_1 = 0, b_2 = 20, b_3 = 40\), and so on. Prior to solving the optimization problem, \(|S|\) scenarios will be randomly and independently realized to capture the stochastic components of the original problem, whose distributions are estimated from the empirical data. Specifically, the random scenario realization procedure will first sample the outcome whether a patient will show up or not with the empirical no-show probability. If patient \(i\) shows up, the service duration of the patient, denoted by \(d_i\), for the \(i^{th}\) patient under scenario \(s\), is then realized via random sampling from its probability distribution; otherwise \(d_i\) will be set to 0.

### 3.3. Model for overbooking with pre-defined time slots

Slot scheduling is ingrained in different healthcare environments. In this section, we first consider for a given service session with pre-defined time slots, how to assign patients to different time slots in order to minimize a weighted sum of multiple performance metrics: resource overtime and idle time and patient waiting time. We first introduce the
decision variables in this model.

**Decision Variables:**

\[
X_{ij} = \begin{cases} 1, & \text{if the } i^{th} \text{ patient in the schedule is assigned to the } j^{th} \text{ time slot;} \\ 0, & \text{otherwise} \end{cases}
\]

\[a_i = \text{appointment time of the } i^{th} \text{ patient in the schedule}\]

\[z_{i}^{\text{start}} = \text{start time of the medical service provided to the } i^{th} \text{ patient under scenario } s\]

\[z_{i}^{\text{end}} = \text{end time of the medical service provided to the } i^{th} \text{ patient under scenario } s\]

\[\text{wait}_{i} = \text{waiting time of the } i^{th} \text{ patient for receiving the medical service under scenario } s\]

\[\text{idle}_{i} = \text{idle time of the resource between the } (i)^{th} \text{ and the } (i + 1)^{th} \text{ services under scenario } s\]

\[\text{ot}_{i} = \text{overtime of the resource under scenario } s\]

Our goal is to determine the optimal number of patients assigned for each specified time slot in a session, which is given by the \(X_{ij}\) assignments to counteract the uncertainty and inefficiency caused by varying service time and patient no-shows. Appointment time of the \(i^{th}\) patient \(a_i\) is then determined by \(x_{ij}\) and \(b_j\). If patient \(i\) is assigned to time slot \(j\), the appointment time \(a_i\) is simply the beginning time \(b_j\) of time slot \(j\). The rest of the decision variables are to determine the time that each event takes place and the performance indicators that we wish to optimize. We propose a two-stage deterministic equivalent of the problem, formulated as a mixed-integer linear program, for this single server appointment overbooking problem with the consideration of patient no-shows. Here, we note that all the patients, including those no-show patients, would enter the model under each scenario; we will discuss in Section 3.5 how we model those no-shows.

**Mathematical Model:**

\[
\text{minimize} \sum_{s \in S} \sum_{j \in J} \alpha_j \text{idle}_j + \sum_{i \in N, j \in S} \text{wait}_i + \sum_{i \in N, s \in S} \text{idle}_i
\]

subject to

\[a_i = \sum_{j \in J} b_j X_{ij} \quad \forall i \in N \quad (1)\]

\[\sum_{j \in J} x_{ij} = 1 \quad \forall i \in N \quad (2)\]

\[a_{i+1} \geq a_i \quad \forall i \in N \setminus \{n\} \quad (3)\]

\[z_{i}^{\text{start}} = 0 \quad \forall s \in S \quad (4)\]

\[z_{i}^{\text{start}} = a_i + \text{wait}_i \quad \forall i \in N, \ s \in S \quad (5)\]

\[z_{i}^{\text{end}} = z_{i}^{\text{end}} + \text{idle}_i \quad \forall i \in N \setminus \{1\}, \ s \in S \quad (6)\]

\[z_{i}^{\text{end}} = \text{idle}_i - \text{ot}_i \quad \forall i \in N, \ s \in S \quad (7)\]

\[z_{i}^{\text{end}} + \text{idle}_i - \text{ot}_i = E \quad \forall i \in N, \ j \in J \quad (8)\]

\[z_{i}^{\text{end}}, z_{i}^{\text{end}}, \text{wait}_i, \text{idle}_i, \text{ot}_i \geq 0 \quad \forall i \in N, \ s \in S \quad (9)\]

\[x_{ij} \in \{0, 1\} \quad \forall i \in N, \ j \in J \quad (10)\]

In a standard two-stage stochastic programming model, decision variables can be divided into two types, the first stage and the second stage variables (Al-Qahtani & Elkamem, 2011). The first stage variables are determined with the consideration of the realizations of the random parameters. Specifically, for the mathematical model described above, the first stage variables are the assignments of the patients to different time slots, determined by \(x_{ij}; i \in N; j \in J\). Because the beginning time of each time slot is given by parameter \(b_i; j \in J\), the appointment time for each patient can be determined by Constraints (1), (2) and (10) altogether. Furthermore, because of the assumption that the patients are punctual, the arrival time of each patient is determined as well. All the second stage variables (which are all associated with a scenario index \(s\)) are determined by the values of \(a_i; i \in N\) chosen in the first stage, after parameters \(\{a_i; i \in N, s \in S\}\) are realized. The proposed integrated model simultaneously searches for the optimal values of all decision variables of the two stages.

Our objective is to minimize a weighted sum of the resource over-time (OT) and idle time (IT) and average patient waiting time (WT) among all the scenarios, which is equivalent to minimizing the average over all scenarios. Constraint (1) acts to assign the correct arrival time for the \(i^{th}\) patient, who is assigned to the \(j^{th}\) time slot in the schedule. Constraint (2) ensures that exactly one time slot is scheduled for each patient.

Constraint (3) enforces the correct sequence of the scheduled appointments. All the remaining constraints except for (10) in the model are associated with a scenario. Constraint (4) initializes start time of service for the first patient, which is 0. Constraints (5) and (7) calculate the patient waiting time, service start time and service end time for each patient. Constraint (6) computes the idle time of the resource between the \((i-1)^{th}\) and the \(i^{th}\) diagnostic tests. Constraint (8) measures the overtime of the resource. Constraint (9) imposes non-negativity condition on the rest of the variables. Finally, Constraint (10) states that the assignment of patients \(x_{ij}; i \in N, j \in J\) are binary.

### 3.4. Model for overbooking without pre-defined time slots

Recent advances in information and communication technologies have revolutionized many working procedures in different industries, such as the medical and healthcare industries. In particular, web-based applications provide a convenient and flexible medium of communicating and sharing information between patients, administrative staff and care providers. In recent years, online appointment booking systems have become more and more popular among healthcare professionals. We have observed that the adoption of varying-length time slots is more common when scheduling patients on online appointment systems. Such trend motivates us to consider the completely flexible session structure, which we define as appointment time slots that are not restricted to fixed appointment times and interval lengths. In this section, we propose a related two-stage deterministic equivalent model of the problem for scheduling patients with this flexible session structure.

Different from the model presented in Section 3.3, the time slot is no longer pre-defined, but flexible with variable length. The decision variables in this model are identical to those in the model with pre-defined appointment times introduced in Section 3.3, except that \(x_{ij}\) is no longer required as a patient is not assigned to a fixed appointment time in this setting. Instead, the decision variable \(a_i\) acts to determine the appointment time for patient \(i\). Our goal is to determine the optimal appointment time \(a_i\) for each patient \(i\) in a session to minimize the objective function. To make a fair comparison with the performance of the model with pre-defined appointment times, we consider the same objective as in Section 3.3.

The first stage variables are the appointment times of each patient, determined by \(\{a_i; i \in N\}\). Again, by the assumption that the patients who show up are punctual, the appointment time of each patient is the same as the arrival time. The second stage variables (which are all associated with a scenario index \(s\)) are determined by the values of \(\{a_i; i \in N\}\) chosen in the first stage, after the parameters \(\{a_i; i \in N, s \in S\}\) are realized. The proposed integrated model simultaneously searches for the optimal values of all decision variables of the two stages. The constraints in this model include Constraints (3)–(9). In other words, the decision variables \(x_{ij}; i \in N, j \in J\) and Constraints (1), (2) and (10) are dropped in this model. In addition, Constraint (11) is imposed to ensure that no appointment is scheduled.
beyond the session period.

\[ 0 \leq a_i \leq E \quad \forall i \in N \quad (1) \]

3.5. Modelling patient no-shows

In the models that we presented in Sections 3.3 and 3.4, we include both the patients who show up and do not in the model under each scenario. To model the effect of patient no-shows, we employ the following procedure. Let \( \pi \) be the probability of a patient not showing up for his/her scheduled appointment.

1. \( \forall i \in N \) and \( s \in S \), generate a random number \( U_i \) which is uniformly distributed in the interval \((0, 1)\).

2. \( N_s^1 = \{i \in N: U_i > \pi\} \) and \( N_s^0 = \{i \in N: U_i \leq \pi\} \). i.e., \( (N_s^1, N_s^0) \) is a partition of \( N \), where \( N_s^1 \) and \( N_s^0 \) respectively denote the sets of patients who show up for the appointment and do not.

3. For \( s \in S \) and \( i \in N_s^1 \), \( d_{si} \) is generated according to their empirical probability distributions. For \( s \in S \) and \( i \in N_s^0 \), \( d_{si} = 0 \).

The above procedure can be interpreted as follows. When a patient does not show up, he/she is still assumed to physically arrive at the appointment time and has to wait until the resource is available. Although waiting time may be incurred for this “no-show” patient, this value is not included in the objective, and thus the performance is not affected. As the service duration for this patient is zero, thus the presence of this patient would not impact the waiting times of subsequent patients. Furthermore, the idle time of the resource is still the sum of the time differences between start times and end times of consecutive “valid” tasks.

4. Computational study

In this section, we conduct a computational study in the setting of our motivating application, i.e., the medical imaging center that provides ultrasound diagnostic services to both inpatients and outpatients. We first provide the background about their current practice of appointment scheduling. On each working day, there are 12 time slots available per morning (9 am to 1 pm) / afternoon (2 pm to 6 pm) session. i.e., \( E = 240 \). Each time slot is of 20 min, which is approximately the total of the average examination time and the average setup time for a single medical diagnostic test. In most cases, an appointment has to be made for a patient to guarantee the required tests can be scheduled. Surprisingly, we observed from the historical data provided by the center that there was a high proportion of patients who showed up for their appointment but could not undergo the scheduled examinations. After a discussion with the staff at the center, we realized that some diagnostic examinations required patients to fast a certain period until after the examination time, and the other for the setup time. After that, we adopt different overbooking levels and multiple sets of objective weights for the deterministic equivalent model presented in Section3.3. The resulting average waiting time, idle time and overtime per service session are reported in Table 3. Based on these reported performance measures, a suitable combination of weights which can strike a good balance among different measures across most conditions is then chosen. In this case, 500 scenarios are considered to be sufficient because models with more than 500 scenarios lead to consistent computational results. In addition, to indicate the relative importance of each measure, we normalize the weights such that their sum is equal to 1.

As expected, from Table 3, we observe that, as the overbooking rate increases, the resource idle time decreases while the resource overtime and the patient waiting time increase. This trend illustrates the tradeoff between the operational cost and the patient experience at the facility when appointment overbooking is adopted. We also observe that overbooking can reduce system inefficiency due to patient no-shows. As an example, for the set of objective weights \((0.63, 0.30, 0.07)\), when the number of patients increases from 14 to 16, the resource idle time

| Table 2
<table>
<thead>
<tr>
<th>Summary statistics from empirical data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random variable or parameter</td>
</tr>
<tr>
<td>---------------------------------</td>
</tr>
<tr>
<td>Examination time per test (min)</td>
</tr>
<tr>
<td>Setup time for each test (min)</td>
</tr>
<tr>
<td>No-show probability</td>
</tr>
<tr>
<td>No-fasting probability</td>
</tr>
</tbody>
</table>

\( p \)-value from Chi-Square test is greater than 0.10.

4.1. Data collection and summary

In this sub-section, we present the empirical study that motivated this study. We collected data related to the ultrasound imaging service during the period of January 2015 to March 2015. In total, 341 observations were recorded. The data contains the information about the dates and times that patients called to make appointments, their appointment dates and times, their arrival times at the center (if showed up), the start times and end times of medical diagnostic tests, and patient departure times. The empirical data are summarized in Table 2. In our study, we consider that each patient may or may not show up for his/her appointment based on the probability estimated from the data, as shown in Table 2. If a patient attends the appointment, we assume he/she arrives at the scheduled appointment time and the service time and the required set up time follows the probability distributions estimated from the historical data, as shown in Table 2. Note that the examination time and setup time are fitted separately because they were collected from different sources. The examination durations were recorded automatically by the Hospital Information System (HIS), while the setup times were recorded manually. Table 2 shows that the examination time per test follows a beta distribution with parameters \( (\alpha = 2.3, \beta = 12.7) \) and an offset of 0.5. And the setup time follows a lognormal distribution with parameters \((\mu = 7.01, \sigma = 6.43)\) and an offset of -0.5. The \( p \)-value of both Chi-Square tests are greater than 10%.

4.2. Determining objective weights

We determine the set of weights in the objective function in a similar way as that in Oh, Muriel, Balasubramanian, Atkinson, and Ptaszkiewicz (2013). Specifically, we first generate 500 scenarios according to the empirical panel characteristics listed in Table 2 to represent the stochastic environment under the study, where the total service time are calculated based on two random samplings, one for the examination time, and the other for the setup time. After that, we adopt different overbooking levels and multiple sets of objective weights for the deterministic equivalent model presented in Section 3.3. The resulting average waiting time, idle time and overtime per service session are reported in Table 3. Based on these reported performance measures, a suitable combination of weights which can strike a good balance among different measures across most conditions is then chosen. In this case, 500 scenarios are considered to be sufficient because models with more than 500 scenarios lead to consistent computational results.
decreases by 16 min while the resource overtime and the patient waiting time respectively increase by only 0.9 min and 2.1 min. Moreover, from Table 3, we observe that the weight combination \{0.63, 0.30, 0.07\}, compared with the other weight sets, can better balance among over time, idle time, and waiting time for varying overbooking levels. Therefore, we adopt this set of weight combination for the rest of our computational experiments.

### 4.3. Determining overbooking level

In this subsection, the overbooking level, i.e. the number of overbooked patients, is further determined by the desired patient “indirect” waiting time. Both the waiting time at the facility and the “indirect” waiting time to access the appointment are important performance measures in appointment scheduling. (Boenzi, 2014; Boenzi, Mummolo, & Rooda, 2013) For this medical imaging center under study, unattended time slots not only lower service utilization but also prevent timely access of potential patients. The recorded data show that the current average accepted appointment request rate is 26.38 per day, which exceeds the number of time slots available each day (24) in the current practice. Theoretically, the appointment system is unable to accommodate this request rate. However, because of arrangement of extra working time (i.e., over time), patients are still able to schedule their appointments but the average waiting time for patients to receive medical diagnostic tests is as long as 83.4 h, which is 3.47 days, shown by the collected data. One important goal for adopting overbooking under this situation is to enable patients to receive necessary medical services at an earlier stage.

Considering the current request rate, 26.38 per day, if four patients can be overbooked each day (2 more patients per session), i.e., about 28 patients scheduled each day, then, by using an $M/D/1$ queueing model, we can evaluate the average “indirect” waiting time between the request date and the actual appointment date:

\[
\text{Waiting time} = \frac{26.38}{28} \times 2 \times (1-26.38/28) = 0.29 \text{ days}
\]

That is, on average, the patient can receive medical diagnostic tests on the next day after the request is made. This reduction in “indirect” waiting time would be beneficial to patients, for example, due to early diagnosis and medical treatments. Therefore, a sensible overbooking level of 2 patients, i.e. totally 14 patients for each session, is adopted as the baseline.

---

### Table 3

<table>
<thead>
<tr>
<th>Average WT (minutes/patient), IT and OT (minutes/session) with different overbooking levels and sets of objective weights.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overbooking rate (%)</td>
</tr>
<tr>
<td>Total number of patients (N)</td>
</tr>
<tr>
<td>Performance metric</td>
</tr>
<tr>
<td>OT</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>WT</td>
</tr>
<tr>
<td>OT</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>WT</td>
</tr>
<tr>
<td>OT</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>WT</td>
</tr>
<tr>
<td>OT</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>WT</td>
</tr>
<tr>
<td>OT</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>WT</td>
</tr>
<tr>
<td>OT</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>WT</td>
</tr>
<tr>
<td>OT</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>WT</td>
</tr>
</tbody>
</table>

### Table 4

Pre-defined time slot structures.

<table>
<thead>
<tr>
<th>Start times of pre-defined time slots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time slot</td>
</tr>
<tr>
<td>Fixed-length</td>
</tr>
</tbody>
</table>
4.4. Optimal solution structures

In this section, we first investigate the optimal overbooking structures resulting from the three different slot structures suggested by the models introduced in Section 3. With our computational analysis, we observed that the optimal overbooking structures become consistent after around 500 i.i.d. scenarios are generated. Therefore, in this paper, the optimal solutions obtained from computational instances of 1000 i.i.d. scenarios are reported. Then we compare the performance measures achieved by these solutions.

4.4.1. Optimal assignment of patients to fixed-length slots

We first consider a service session with a fixed-length time slot interval, which is currently adopted by the medical imaging center of our study. In the experiments, we consider the morning session, which starts and ends at 9:00 and 13:00 respectively. The start time of each fixed-length slot in a typical session is listed in Table 4. Based on the discussion in Section 4.3, 14 patients will be assigned to these 12 time slots. The model presented in Section 3.3 with fixed pre-defined time slots is employed to find the optimal assignment of patients to the time slots so as to minimize the objective function.

Fig. 1 shows the optimal number of patients assigned to each time slot for the fixed-length time slot structure: (3, 1, 2, 1, 1, 1, 1, 1, 0, 0). Specifically, the optimal solution assigns three patients to the very first time slot to avoid idleness caused by an empty queue plus patient no-shows, and double books time slots 3 and 5 to counteract the patient no-show rate of 53.7%. On the other hand, the solution schedules zero patient to time slot 11 and 12 (which is at the end of the session) to avoid staff overtime. An interesting “frontloaded” pattern is observed in Fig. 1. This optimal overbooking assignment shows a pattern that is not exactly the same as those previously reported in the literature by other researchers and practitioners, but similarities can be observed in Fig. 1, such as “double booking” in LaGanga and Lawrence (2012) and Zacharias and Pinedo (2014).

4.4.2. Optimal assignment of patients to dome-pattern time slots

Now we consider an uneven time slot structure in a dome pattern, which is widely reported in the healthcare literature for its robust performance in many different environments. However, this dome-pattern time slot structure has not been studied for appointment overbooking yet. This motivates us to examine its optimal overbooking assignment and study how it differs from the one of the fixed time interval setting. The start time of each time slot is listed in Table 4 and its dome pattern is illustrated in Fig. 2. This dome pattern is generated using the formula presented on page 50 of Cayirli et al. (2006) with parameters \( \beta_i \) (\( \beta_1 = 0.1, \beta_2 = 0.2, \beta_3 = 0.01 \)) and \( k_i \) (\( k_1 = 6, k_2 = 12 \)). The mean and the standard deviation of the total service time (including the examination time and the setup time) for a patient are 20 min and 13 min, respectively, as obtained from the data. The parameter values of \( \beta_i \) and \( k_i \) are selected to form a dome pattern with the lengths of these time slots. As we observe in Fig. 2, the dome pattern has a shorter interval length for the first and the second last time slots and a longer interval length in the middle. The last time slot has a relatively longer interval length compared to the first and the second last time slots to mitigate the risk of staff overtime. The model presented in Section 3.3 with fixed pre-defined time slots of this dome pattern is solved to find the optimal assignment of patients to the time slots.

Fig. 3 shows that for this dome-pattern time slot structure, a relatively even patient assignment (2, 1, 2, 2, 1, 2, 1, 2, 1, 0, 0) is found in the optimal solution. Compared to that for the fixed-length time slot structure, there is less variability in the number of patients assigned for each time slot. As observed in Fig. 3, overall, the optimal numbers of patients scheduled to the slots appear to form a “2-1-2-1” pattern to balance the adverse outcomes caused by the possible accumulated waiting line and the resource idle time due to patient no-shows. Compared with the patient assignment for the fixed-length time slots, this assignment is relatively uniform because the dome-pattern structure can reduce the impact of a possible accumulated waiting line. Finally, no patient is scheduled for the last three time slots to avoid staff overtime due to the relatively short interval of these time slots in the dome pattern.

4.4.3. Optimal assignment of patients to completely flexible time slots

In this subsection, we take a step further considering a completely
flexible time slot structure, where we allow appointments to be scheduled for any time point within the service session, to overbook patients. The model presented in Section 3.4 is used to obtain the optimal appointment times for all the 14 patients. The optimal appointment times are shown in Fig. 4 and the time slot lengths are illustrated in Fig. 5.

First of all, double booking is observed for the first time slot only to avoid staff idleness at the beginning of the session. Apart from that, only one patient is assigned for each time slot. Second, in terms of the time slot length, a “dome-dome-dome pattern with alternate long and short time slots” is observed through the whole session in Fig. 5: there are three small dome-patterns except for the last two time slots which have a longer slot length to avoid staff overtime. This finding is consistent with our intuition. Instead of managing the crowd due to the accumulated patients in the waiting line and the resource idleness resulting from patient no-shows all at once by a single dome, such manipulation should be made at different periods within the session to alleviate the negative outcomes that may occur in different time slots.

In addition, alternate short and long time slot lengths within domes act to balance resource idle time and patient waiting time.

4.5. Comparison of performances resulting from optimal solutions for overbooking in different settings

In this section, we will examine unique features of the optimal solutions suggested by the three previously examined slot structures and compare their performances, measured by the staff overtime and idle time and the patient waiting time. Fig. 4 visualizes the structure of these three optimal overbooking solutions for a morning session. We observe that the number of patients scheduled for each time slot and the time slot length are manipulated to counteract the uncertainty of service time and patient attendance.

When the time slot length is fixed due to administrative convenience, the number of patients overbooked per slot has a higher variation. When the time slot length is adjustable, the overbooking level tends to be more stable. However, the best strategy to deal with uncertainty is the adoption of alternate short and long time slots in a series of dome patterns, as we illustrated in Fig. 5. By adopting this, the service provider can resolve the accumulated waiting line of patients and utilize any possible idle time within the regular working hours as much as possible.

An interesting and important question is the following: how much will the service provider lose by keeping fixed-length time slots to gain administrative convenience? To answer this question, we compare the performance of these three optimal overbooking solutions, which were previously illustrated in Fig. 4, regarding patient experience and provider efficiency. We apply a simulation approach to assess the solution quality. One thousand scenarios are generated using the empirical distributions presented in Table 2. Then we schedule the patients with the three overbooking solutions and measure the three performance metrics – staff overtime and idle time and patient waiting time – under these one thousand scenarios, as reported in Table 5. The minimum value (i.e., the best) of each measure among the three schemes is highlighted in blue.
First of all, Table 5 shows that the dome-dome-dome pattern resulting from the flexible time slots model has a better performance than the other two in terms of all the three performance measures. Compared to the fixed-length time slot structure and the dome-shaped scheme, the dome-dome-dome scheme leads to 14% and 12% reductions, respectively, in the average patient waiting time. The average waiting time resulting from the dome-dome-dome schedule is statistically lower than those of the other two (p < 0.001), while the overtimes (and idle times) are not statistically different when comparing the dome-dome-dome scheme with the other two (p > 0.05). Note that the benchmark for comparison is the optimal overbooking solutions resulting from the fixed-length and dome-shaped time slot structure.

Fig. 6 shows the percent differences in average waiting time, overtime, and idle time per session resulting from three overbooking solutions. The computational results above suggest that the flexible time slot structure can improve patient waiting time without sacrificing staff overtime and idle time. This is because the dome-dome-dome pattern with alternate long and short time allocation can better utilize the medical resources with uncertainty in both of the service time and patient no-shows. Mathematically, given that $b_j \leq E$ $\forall j \in J$, the flexible appointment time model is a relaxation of the pre-defined time slots model projected onto the subspace defined by those variables without $\{x_{ij}; i \in N, j \in J\}$. Therefore, the flexible appointment time model can always achieve a lower optimal value than the pre-defined time slots model for the same set of scenarios. For practical implementation, the flexible appointment time model also takes advantage of its convexity (without integer variables) so that the problem can be solved efficiently for real-time solutions on online appointment systems. We note that the improvement in the performance of the flexible appointment time scheme depends on the problem characteristics. In the next section, we conduct experiments to identify the situations where flexible appointment times could provide more significant benefit.

### 5. Sensitivity analysis

To examine thoroughly the benefits of this flexible time slot structure, we conduct a sensitivity analysis with a set of computational experiments. We conduct a full factorial experiment while considering two factors: the overbooking rate and the average service duration. There are 3 levels for each factor. Therefore, there are, in total, 9 combinations of different factor levels, which are listed in Table 6. In this set of experiments, we set the overbooking rate as the closest realistic equivalent for the no-show probability. For the fixed time slot length considered, the slot duration is set to the average service time. Other problem settings are identical to those presented in Section 4.

![Fig. 6](image-url)  
Fig. 6. Percent differences in WT, OT, and IT per session resulting from the three solutions with fixed-length, dome-pattern, and dome-dome-dome pattern time slot structures.

sensitivity analysis more robust, instead of using the empirical service times observed in our motivating case, we make a more general assumption about the service time: we assume it follows a normal distribution with a standard deviation of 10 min, truncated at 5 min (i.e., the service lasts for at least 5 min for each patient who shows up).

For each factor combination mentioned above, 1000 scenarios are generated and the fixed-length, dome-pattern and flexible slot structures are adopted to overbook patients. Table 6 lists the absolute and percentage performance improvement of each of the three schemes under different experimental settings with fixed-length slot structure as the benchmark. The computation results in Table 6 show results similar to what we have observed in Section 4.5: the flexible time slot structure reduces the patient waiting time with almost no compromise in staff overtime and idle time. This benefit is significantly larger when:

(1) The number of patients to be overbooked is lower.
(2) The service duration is longer.

The reason for better performance under the two scenarios above can be rationalized as follows. Compared with the fixed-length and single dome schemes, the schedule generated from the flexible slot has more choices. First, a lower number of patients to be scheduled implies that in the flexible slot setting, for each patient, there are more possible appointment start times. While for the fixed-length and single dome schemes, there are fewer appointment start times. Second, the longer service duration also means fewer appointment start times and larger granularity of time slots for pre-defined structures. In particular, when the appointment time slots are longer, moving one patient from one slot to the previous or subsequent one can have a larger absolute change in the performance. On the other hand, the flexible slot scheme allows the provider to schedule the patients for any time point within the session and this flexibility can compensate for the adverse effects due to the above-mentioned reasons. Thus, when considering the tradeoff between a higher level of flexibility in appointment time and administrative convenience, the provider may consider the number of patients to be scheduled for each session and the average service duration. Such flexibility in appointment time is more beneficial when there are fewer patients to be scheduled and the service time is longer.

6. Conclusions

In this paper, we study the problem of medical appointment scheduling with the consideration of patient no-shows and allow overbooking to compensate for the negative outcomes. We propose a two-stage deterministic equivalent of the stochastic optimization problem to obtain the optimal overbooking solutions with three different time slot settings. Among the three different slot structures we investigated, the fixed-length time slot setting is the most popular due to its administrative convenience, the dome-pattern is widely recognized in the healthcare literature for its ability to mitigate the downside effects of accumulated waiting and the completely flexible time slot setting is not restricted to fixed appointment times for the provision of additional improvement in performance.

We have a number of key findings in this study. The most important message to practitioners is that a single dome-pattern does not always perform very well when patients may not show up. A single dome-pattern does help mitigate the accumulated congestion and avoid overtime. However, it may not reduce the staff idle time caused by patient no-shows from time to time. Under these circumstances, a dome-dome-dome pattern with alternate short and long time slots works most efficiently by relieving the congestion and utilize the idle-ness in time. Our second key result is that our sensitivity analysis shows that the flexible slot scheme is more effective when the number of overbooked patients is lower and the service duration is longer. Furthermore, the benefit brought out by a dome-pattern time slot can be substituted by a frontloaded-pattern in the number of patients booked for each time slot. A frontloaded-pattern in the number of patients is similar in performance to the dome-pattern under overbooking, but it allows practitioners to stick to fixed-length time slots for administrative convenience.

Ho and Lau (1992), Cayirli and Veral (2003), and Gupta and Denton (2008) have reported that the patient no-shows have the most pronounced impact on the service provider performance than any other patient behavior characteristics, such as walk-ins and punctuality. Our findings presented in this paper provide hospital administrators with useful insights and guidance as to how to design an overbooking system when patient absence is one of their major concerns.

There are certain research directions worth exploring in the future. First, the model presented in this paper only considers the overbooking schedule for a given overbooking level. While in practice, how to determine the appropriate overbooking level is another important management issue. We are currently working on extending our model to consider decision making about both the overbooking level and overbooking to compensate for the negative outcomes. We propose a two-stage deterministic equivalent of the stochastic optimization problem to obtain the optimal overbooking solutions with three different time slot settings. Among the three different slot structures we investigated, the fixed-length time slot setting is the most popular due to its administrative convenience, the dome-pattern is widely recognized in the healthcare literature for its ability to mitigate the downside effects of accumulated waiting and the completely flexible time slot setting is not restricted to fixed appointment times for the provision of additional improvement in performance.

We have a number of key findings in this study. The most important message to practitioners is that a single dome-pattern does not always perform very well when patients may not show up. A single dome-pattern does help mitigate the accumulated congestion and avoid overtime. However, it may not reduce the staff idle time caused by patient no-shows from time to time. Under these circumstances, a dome-dome-dome pattern with alternate short and long time slots works most efficiently by relieving the congestion and utilize the idle-ness in time. Our second key result is that our sensitivity analysis shows that the flexible slot scheme is more effective when the number of overbooked patients is lower and the service duration is longer. Furthermore, the benefit brought out by a dome-pattern time slot can be substituted by a frontloaded-pattern in the number of patients booked for each time slot. A frontloaded-pattern in the number of patients is similar in performance to the dome-pattern under overbooking, but it allows practitioners to stick to fixed-length time slots for administrative convenience.

Ho and Lau (1992), Cayirli and Veral (2003), and Gupta and Denton (2008) have reported that the patient no-shows have the most pronounced impact on the service provider performance than any other patient behavior characteristics, such as walk-ins and punctuality. Our findings presented in this paper provide hospital administrators with useful insights and guidance as to how to design an overbooking system when patient absence is one of their major concerns.

There are certain research directions worth exploring in the future. First, the model presented in this paper only considers the overbooking schedule for a given overbooking level. While in practice, how to determine the appropriate overbooking level is another important management issue. We are currently working on extending our model to consider decision making about both the overbooking level and overbooking schedule in a rolling-horizon setting (Chen, Balasubramanian,
& Kuo, 2017). An obvious way to tackle this issue is by doing an iterative search with the current model to determine the best overbooking level leading to the minimum objective value, i.e. the weighted sum of the idle times, waiting times and overtimes. More sophisticated ways may involve other considerations, such as the total revenue of the service provider or the indirect waiting times of the patients.

Apart from that, our model assumes that the patients are punctual. However, for most of the healthcare systems, this may not always be true. As the second most disruptive patient behavior only next to patient no-shows, unpunctuality may play a crucial part in clinic schedules. Incorporating consideration about patient unpunctuality is another possible extension of our current model.

Another area worth studying is the effect of multiple assignments. The service provider may wish to limit the maximum number of patients assigned to the same time slot to avoid conflicts that may occur among patients arriving simultaneously. Examination on the impact of such limitation on the overbooking schedule and consequent performance measures, such as overtime, will be interesting.

Acknowledgments

This research is supported by Macau Science and Technology Development Fund under Grant No. 088/2013/A3 and Hong Kong Research Grants Council under General Research Fund Grant No. 14202115. The authors would like to thank Ms. Valencia Chang, the Administrative Director of the Macau University of Science and Technology Hospital, and Professor Chuan Pang, Vice President of the Macau University of Science and Technology, for providing us with their concerns about appointment scheduling for medical diagnostic tests from a hospital administration point of view.

References


