Primary-Care Clinic Overbooking and Its Impact on Patient No-shows

Bo Zeng\textsuperscript{1}, Hui Zhao\textsuperscript{2} and Mark Lawley\textsuperscript{3}

\textsuperscript{1}Industrial and Management Systems Engineering, University of South Florida
\textsuperscript{2}Krannert School of Management, Purdue University
\textsuperscript{3}Weldon School of Biomedical Engineering, Purdue University
\textsuperscript{1}bzeng@usf.edu, \textsuperscript{2}zhaoh@purdue.edu, \textsuperscript{3}malawley@purdue.edu

Following the successful stories in the airline industry, many primary-care clinics have adopted overbooking to deal with their prevalent patient no-show problem. However, there has been very limited research, to the best of our knowledge, that analyzes the impact of overbooking on the major causes/factors of patient no-show and its implications. While overbooking has little impact on many random factors that affect no-show, it does impact two important factors - appointment delay (the time between a patient requesting an appointment and his actually seeing the doctor) and office delay (the amount of time patients wait in the office to see the doctor). In this paper, we develop a general framework to explore the impact of overbooking on these two important factors. While overbooking increases office delay (which negatively affects patient no-show rates), it appears to reduce appointment delay (which positively affects patient no-show rates). Our results show, considering both impacts, while overbooking increases clinic’s expected profit most of the time, patient no-show rates always increase after overbooking! Further, there exists a critical range of the patient panel size within which overbooking may also reduce the clinic’s expected profit. However, we propose two easy-to-implement strategies, overbooking with \textit{controlled appointment queue} and \textit{selective dynamic overbooking}. Both strategies can increase the clinic’s expected profit and improve no-show rates at the same time.

Key words: patient no-shows, overbooking, scheduling.

1. Introduction

Patient no-show is one of the most serious operational issues facing nearly all primary-care clinics (Cayirli and Veral, 2003; Gupta and Denton, 2008) due to its multi-facet damage. On the one hand, it wastes critical resources and causes interruptions in the scheduling process and patient flow; on the other hand, it limits clinic accessibility to other patients, leading to lower staff productivity and reduced revenues for healthcare providers. Unfortunately, clinic no-show rates have often been significant in practice, ranging from 5\% to 60\% (Woodcock, 2003). For some healthcare settings such as public pediatric, no-show rate may be as high as 80\% (Rust et al., 1995). Further, these statistics generally do not include the \textit{late appointment cancelations}, which have nearly the same damage as no-shows and can be dealt similarly.

Customer no-show is not unique in the healthcare industry. The most well-known is the airline industry where passengers miss flights for various reasons. To improve revenues, the airline industry
has successfully applied overbooking strategy to deal with passenger no-shows; see Rothstein (1985); Gang (1998). As stated in Smith et al. (1992), in 1990, more than $225 millions, about 40% of American Airline’s total benefit obtained through revenue management, resulted from overbooking.

Following the successful stories in the airline industry, many clinics have implemented overbooking to stabilize revenue streams and improve healthcare access; see Keir et al. (2002); Kim and Giachetti (2006). In fact, because it is easy to implement and has nearly zero operational cost, the option of overbooking has been provided in many commercial scheduling software, e.g., Encore2008 and Spectrasoft. To fully utilize the power of overbooking, advanced overbooking methods have also been developed by healthcare engineering researchers, e.g., Kim and Giachetti (2006), Laganga and Lawrence (2007), Muthuraman and Lawley (2008), Liu et al. (2009), Zeng et al. (2010) and Robinson and Chen (2009).

Despite its widespread use to counter patient no-show, the impact of overbooking on the major causes/factors of primary-care patient no-show and its implications have not been analyzed.

There has been lots of literature across different disciplines studying the major causes/factors of patient no-show (Dervin et al., 1978; Goldman et al., 1982; Bean and Talaga, 1992; Garuda et al., 1998; Lacy et al., 2004). Each study reveals a list of reasons for patient no-show. When combining many of these lists, we find some prominent factors such as appointment delay (the time between a patient requesting an appointment and his actually seeing a doctor) and many other factors such as patients’ dissatisfaction, forgetfulness, time constraints (concerns of long waiting time), transportation, and patients’ anxiety. Notice that many of the factors are random or uncontrollable, yet some change with the administration’s booking/scheduling strategy. It is the purpose of this paper to examine these factors such that through this analysis we can better evaluate the effect of overbooking and propose easy-to-implement strategies that may improve the clinics’ performance.

We will focus on two important factors - appointment delay and office delay (the amount of time patients wait in the office to see the doctor) - and capture other random factors using a random variable. We choose these two factors for a couple of reasons: (1) they are two prominent controllable factors that the literature has shown to affect patient no-show (e.g. Green and Savin (2008), Daggy et al. (2010), Camacho et al. (2006) and Bean and Talaga (1992)). (2) The impact of overbooking on these two factors (in fact, any factors) of primary-care patient no-show and its implications have not been analyzed.

1Primary-care is different from acute-care for which patients usually have little choice in their show-up decisions. In addition, primary-care patients are usually recommended for future appointments long before the appointment date, e.g., follow-up checks. Since there is no fee for making an appointment, even if patients are not sure whether they will show up for the appointment, they tend to make an appointment anyway “just in case”, having the idea that they can later cancel the appointment. However, due to the long appointment delay and the fact that there is little penalty for no-show, many of them forget to cancel, causing real no-show.
Specifically, overbooking may have positive impact on patient no-show rate by reducing the appointment delay because of its potential of scheduling more patients. For example, in Bibi et al. (2007), the authors report that the clinic used managed overbooking as an intervention tool to reduce appointment delay and observed reduced no-show rate. On the other hand, overbooking may have a negative impact on patient no-show rate due to the increased office delay. As pointed out by Camacho et al. (2006) and Bean and Talaga (1992), patients make their show-up decisions by trading off the benefit (utility) from a particular visit with their perceived cost for this visit. For patients who are covered by third party payments, the only cost is the travel and waiting time (office delay) involved in keeping the appointment. As the office delay increases, the costs to the patient correspondingly increase. For some people this (spending much time waiting) is impossible, and for others the difficulty may be enough to tip the balance towards not attending (Sharp and Hamilton (2001) and Bibi et al. (2007)). Indeed, it is observed that patients with bad waiting time experiences frequently have no-show behavior in their future appointments (Lowes, 2005). Dyer (2005); Lacy et al. (2004); Garuda et al. (1998); van Baar et al. (2006) also report the linkage between patient no-show and their long waiting time or patients’ dissatisfaction (for which long waiting time is a major factor). Sharp and Hamilton (2001) also point out that overbooking may be counterproductive due to the increased no-show caused by the longer waiting time. Further, the negative impact of longer office delay on patient no-show can be particularly true when the utility of an office visit decreases from the time of making the appointment to the time of the appointment (due to the long appointment delay). Indeed, Lacy et al. (2004) has identified significantly improved symptoms as one of the important reasons for patient no-show. To summarize, longer office delay caused by overbooking may negatively affect showing up decisions for patients, especially those who are very sensitive to office delay.

Note that for the patients, one way to counter long waiting time is to arrive late. However, to deal with late arrivals, many clinics impose policies such as rescheduling patients’ appointments upon late arrival up to a certain amount (typically 10-15 mins) or seeing a late patient after all on-time patients, which causes extremely long office delay. Due to the high cost of being cancelled/rescheduled upon arrival or the potentially extremely long waiting time, late arrivals are not common among patients. In fact, empirical study indicates that patients arrive, on average, 3-16 minutes before their appointments, see Cayirli et al. (2006) and the references therein. In addition, for some patients, another way to counter consistently long office waiting time is to switch healthcare providers. However, since many issues (e.g., insurance network, patient-provider relationship, and provider specialty) are involved in switching providers, this is not always an option for many patients, nor is this discussed in much literature. In fact, in the limited literature that does discuss provider switching, it is shown that only a small percentage of patients voluntarily switch providers (involuntary switching includes those caused by moving, patient referrals, etc.), e.g., 5.4% as reported in Rice et al. (1992).
As we discussed above, overbooking may exert both positive and negative impact on patient no-show rates because of patient responses to the reduced appointment delay and the increased office delay. While most of the operations literature on overbooking focuses on developing overbooking methods to maximize clinics' profits assuming unchanged no-show rates (e.g., Muthuraman and Lawley (2008)), the contribution of our work is to develop a general framework to evaluate overbooking policies considering their impact on patient no-show through their effect on both appointment delay and office delay.

Since the two delays correspond to two queues involved in the problem - one for getting an appointment and the other for seeing the doctor on the appointment day, we progressively build the framework by first considering only the potential impact of office delay and other random factors, referred to as the basic model, and then incorporate the impact of the appointment delay (referred to as the integrated model).

Our framework is a general framework in two aspects. First, we have incorporated a patient office delay tolerance factor, which controls how much impact office delay has on patient no-show decision. By changing the patients' tolerance to office delay, we can adjust the weight on the impact of office delay. When this parameter is set to a very large value, i.e., patients are not sensitive to office waiting time at all, the resulted model does not include the impact of office delay, reflecting only the impact of appointment delay and other random factors. In the numerical study, we particularly studied how this tolerance factor changes patient no-show rate and the clinic's expected profit. Second, for analytical tractability and simple exposition, we develop the framework assuming the clinic adopts a simple overbooking policy used in practice, naive statistical overbooking (NSOB) (Kim and Giachetti, 2006). However, the model framework can be modified to study other overbooking policies proposed in literature. Furthermore, NSOB is a very mild policy without too much overbooking, hence our results will be an underestimate of the impact of office delay (heavier overbooking generally leads to more office delay) compared to many other methods proposed in literature.

Our results show, while overbooking will likely increase the clinic’s expected profits, it may lead to reduced expected profit for the clinic for panel sizes within a critical range. Further, patient no-show rate does not improve after overbooking! Therefore, instead of imposing higher and higher degrees of overbooking, clinics should consider other approaches to more effectively conduct overbooking. In this regard, we propose overbooking with controlled appointment queue length and a selective dynamic overbooking strategy. Both strategies can reduce patient no-show and increase the clinic’s expected profit at the same time.

The rest of the paper is organized as follows. In Section 2, we develop the basic model that only considers the impact of office delay and the random factors, and derive results for both single block scheduling (SBS) and multiple block scheduling (MBS). In Section 3, we further develop the integrated
model by incorporating the queueing model of appointment delay into the basic model for SBS and MBS, and derive results. In section 4, we conduct a comprehensive numerical study to investigate the overall impact of overbooking under different parameters and bring out insights which lead to two easy-to-implement overbooking strategies and other useful suggestions to practitioners. In Section 5, we conclude with summary of results, discussion of managerial insights, and some future research directions.

2. The Basic Model with Office Delay

In this section, we study the interactions between patient no-show rate and office delay caused by overbooking, the basic model. We first introduce the basic elements and conceptually develop the basic model, using a single block scheduling (SBS) model for demonstration. We then characterize the solutions to the basic model under SBS and MBS, respectively. Throughout this paper, service times are assumed to be exponentially distributed with a mean normalized to one.

2.1 The Elements of the Basic Model

In the SBS model, all patients are scheduled to arrive at the beginning of the block to see the physician. Clearly, if there are multiple patients scheduled in the block (a typical practice to reduce physician idle time) and the physician serves one patient at a time, patients will expect to have non-zero waiting time. As pointed out in Bean and Talaga (1992) and Camacho et al. (2006), a patient makes his show-up decision by comparing his perceived utility from a clinic visit and the disutility (loss of utility) from the factors such as the expected waiting time.

![Figure 1: Patients’ Utility of A Clinic Visit and Disutility of Waiting Time](image)

To model the fact that patients have different perceived utility on their clinic visits and that many random factors other than waiting time affect patient no-show, e.g., weather, transportation times, patients’ emotions, we factor all these into a random variable $c$, indicating the patients’ perceived utility
less the random factors, and assume $c$ follows a uniform distribution over $[-C_l, C_u]$ with $C_l, C_u \geq 0$ (see Figure 1(a)), i.e.,

$$g(c) = \begin{cases} \frac{1}{C_l + C_u} & \text{if } c \in [-C_l, C_u] \\ 0 & \text{otherwise,} \end{cases}$$

where $g(c)$ is the probability density function of the distribution. We allow $c$ to take negative values to capture the fact that patients could fail to show up due to other random factors even if their expected waiting time is 0. In particular, the ratio $p_0 \equiv \frac{C_l}{C_u + C_l}$ (referred to as the unassignable no-show rate) represents the percentage of patients who will not show up even if the loss of utility from waiting time is 0. Correspondingly, we define $q_0 \equiv 1 - p_0$, knowing that the patient show-up rate is always bounded up by $q_0$.

Let $w$ be a patient’s waiting time in a clinic visit. Clinic experiences indicate that a patient’s disutility of the waiting time, $L(w)$, is a non-decreasing function in $w$. For tractability, we assume patients are homogeneous in their disutility function which can be approximated by a 2-piece linear function of waiting time, i.e.

$$L(w) = \begin{cases} 0 & \text{if } w \leq w_0 \\ \alpha(w - w_0) & \text{if } w > w_0, \end{cases}$$

where $\alpha$ indicates the patients’ sensitivity to waiting time (see Figure 1(b)). A higher $\alpha$ implies the patients are more sensitive (having a higher disutility) to waiting time. We also assume that there exists a threshold of full tolerance of office delay, i.e., a patient incurs disutility from waiting only if his waiting time exceeds $w_0$. Note that in previous literature in which overbooking strategy is developed without considering its impact on patients’ no-show, $w_0$ is assumed to be $+\infty$.

The net utility of a patient’s clinic visit is the utility from the visit less the disutility from waiting, i.e., $c - L(w)$. Let $E[w]$ denote the expected waiting time for his next clinic visit. We assume that a patient will show up if the net utility of his clinic visit is non-negative, i.e., $c - L(E[w]) \geq 0$. Note that since a patient cannot expect the sequence of patients’ arrivals, he does not know how many patients he has to wait for before seeing the physician. Thus, given the same information, all the patients will have the same expected waiting time, referred to as the expected waiting time for this scheduling block. Therefore, for a given expected waiting time, $E[w]$, the patient no-show rate in this block can be computed as

$$p = \int_{-\infty}^{L(E[w])} g(c) dc = \frac{C_l + L(E[w])}{C_l + C_u} = p_0 + \frac{L(E[w])}{C_l + C_u}$$

(1)

and the patient show-up rate will be

$$q = 1 - p = \frac{C_u - L(E[w])}{C_l + C_u}$$

(2)

It is easy to see that the no-show rate increases as the expected waiting time increases, consistent with the observations in practice. Since $q = 1 - p$, throughout the paper, we will interchangeably use $p$ or $q$ in the calculations, whichever is more convenient.
Now, we are ready to describe the dynamic interactions between the office delay caused by overbooking and patient no-show. Suppose initially \( k \) patients are scheduled in a block, the expected waiting time is \( E[w] \), and the corresponding patient no-show rate is \( p = p(E[w]) \). If the clinic overbooks to \( k' \geq k+1 \) patients for the block, it is clear that \( E[w'] > E[w] \) if the no-show rate remains the same, where \( E[w'] \) is the expected office delay with \( k' \) patients scheduled. Thus, \( L(E[w']) \geq L(E[w]) \). However, with the disutility increased because of the longer waiting time, more patients will not show up (see (1)), causing no-show rate to change to \( p' \geq p \). Such dynamic “evolution” continues until it reaches a Nash Equilibrium (NE). The above dynamics can be captured by a game theoretic model between the clinic and the patient population. In the following, we build this model and derive the NE solutions to the patient no-show rate after overbooking under SBS and MBS.

### 2.2 The Basic Model Under Single Block Scheduling (SBS)

Suppose the capacity of the block is \( S \), i.e. the physician can see \( S \) patients in the block (with no overbooking), where \( S \) can only be integers. Given the current no-show rate, \( p \), a clinic (she) determines how many patients (he) to (over)book in the block.

For analytical tractability and simple exposition, throughout the paper, we assume the clinic adopts a simple overbooking policy used in practice, the *naive statistical overbooking* (NSOB) (Kim and Giachetti, 2006). However, the model framework can be modified to study other overbooking policies proposed in literature. With NSOB, the clinic books this block with \( \hat{S} \) patients with

\[
\hat{S} = \lceil S(1+p) \rceil = S + \lceil Sp \rceil
\]

where “\( \lceil \rceil \)” is the ceiling function to ensure the integer requirement of \( \hat{S} \). To simplify exposition, let \( i \) denote the overbooking amount, i.e., the number of patients added due to overbooking. Given the current no-show rate, \( p \), the clinic’s decision, \( i \), can be written as a function of \( p \) (or \( q \)):

\[
i(q) = \lceil Sp \rceil \equiv \lceil S(1-q) \rceil.
\]

Given a show-up rate \( q \), \( S+i(q)−1 \) is the number of patients (excluding oneself) that are scheduled to come. So, the expected waiting time of any patient in this block can be computed from his expected position in the queue \(^2\), i.e.,

\[
E[w] = \frac{1}{2}(S+i(q)−1)q = \frac{1}{2}(S + \lceil S(1-q) \rceil - 1)q.
\]

Then, given the amount of overbooking \( i \), based on equation (2), the patient show-up rate can be derived as:

\[
q(i) = \frac{C_u - \alpha \left(\frac{q(i)}{2}(S+i-1) - w_0\right)^+}{C_i + C_u},
\]

\(^2\)Although all patients are scheduled to arrive at the beginning of the block, we assume there is an \( \epsilon \) difference in their arrival time, which determines their sequence in the queue.
where $a^+ \equiv \max\{0, a\}$ and the $(\cdot)^+$ term captures whether the expected waiting time is greater than the tolerance $w_0$. Further simplification yields

$$q(i) = \min\{\hat{q}(i), q_0\}, \quad (5)$$

where

$$\hat{q}(i) = \frac{C_u + \alpha w_0}{C_l + C_u + \frac{2\alpha}{S+i-1}} \quad (6)$$

is the patient show-up rate if there is positive disutility from waiting and $q_0 = \frac{C_u}{C_u+C_l} = 1 - p_0$ is the patient show-up rate if there is no disutility from waiting. Clearly, when patient response to office delay is not considered (i.e., $w_0 = +\infty$), $q(i) = q_0$, indicating that the patient show-up rate is overestimated.

Since $q(i) = \min\{\hat{q}(i), q_0\}$ is not well-behaved but $\hat{q}(i)$ is convex decreasing in $i$ (see (6)), to establish the NE, we first study the interactions between $i(q)$ and $\hat{q}(i)$ and then consider the impact of $q_0$. To study $\hat{q}(i)$, we make the following two mild assumptions: $C_l \geq \alpha w_0$ (Assumption 1) and $C_l + C_u \geq \alpha$ (Assumption 2). Assumption 1 indicates that $w_0$ and $\alpha$ cannot take large values simultaneously, i.e., patient’s attitude on waiting time is consistent. This assumption is fairly mild and from our numerical study we observe that when this assumption is violated, $q(i)$ typically reduces to $q_0$ and the NE is then $(q_0, i(q_0))$. The second assumption indicates that the range of the patients’ utility for the clinic visit, $[-C_l, C_u]$, is wide enough such that the disutility from one unit waiting time beyond the tolerance is still within the range. We mention that neither of these assumptions is necessary in our numerical study of the patient response game in that the equilibrium is calculated just based on response functions through computational algorithms. Now we are ready to study the interactions between $\hat{q}(i)$ and $i(q)$.

Without considering the integer restriction (denoted as the “continuous game”), the intersection of $\hat{q}(i)$ and $i(q)$ can be obtained by solving (6) together with the continuous relaxation of (4), i.e.,

$$i_c(q) = S(1 - q). \quad (7)$$

Because of Assumption 2, we see that the only feasible solution, $\hat{q}^c$, is

$$\hat{q}^c = 1 + \frac{C_l + C_u}{\alpha S} - \frac{1}{2S} - \sqrt{(1 + \frac{C_l + C_u}{\alpha S} - \frac{1}{2S})^2 - \frac{2(C_u + \alpha w_0)}{\alpha S}}. \quad (8)$$

Thus, the unique NE for the continuous game is $(\min\{\hat{q}^c, q_0\}, i_c(\min\{\hat{q}^c, q_0\}))$.

When considering the integer restriction, we need to further study the structure of the best response functions. Note that because $i(q)$ is neither concave nor convex, we may have multiple NE. The following proposition identifies the NE for the basic model when $q_0 > \hat{q}(i)$ (i.e., $q(i) = \hat{q}(i)$).
Proposition 1. Let \( i^* = i(q^*) \) and \( q^* = \hat{q}(i^*) \), where \( i(\cdot) \) and \( \hat{q}(\cdot) \) are defined in (4) and (6), respectively. For the basic model with SBS, when \( q_0 > \hat{q}(i) \) (i.e., \( q(i) = \hat{q}(i) \)), if \( \hat{q}(i^* + 1) > \frac{S-i^*}{S} \), there is a unique NE, \((q^*, i^*)\); otherwise, there exist at most two NE: \((q^*, i^*)\) and \((\hat{q}(i^* + 1), i^* + 1)\).

The NE characterized above can be demonstrated in Figure 2(a) and Figure 2(b). In these figures, \( i(q) \) is the set of solid vertical lines. Its continuous relaxation, \( i_c(q) \), is the solid straight line with a slope equal to \(-\frac{1}{S}\).

In the above, we have considered the NE when \( q_0 > \hat{q}(i) \) (i.e., \( q(i) = \hat{q}(i) \)). Since \( q(i) = \min(\hat{q}(i), q_0) \), by comparing the values of \( q^* \) and \( \hat{q}(i^* + 1) \) with \( q_0 \), we can characterize all the NE of the basic model with SBS, specified in the following theorem.

Theorem 2. For the basic model with SBS, given \((q^*, i^*)\) defined in Proposition 1, we have

Case (i): If \( q^* \leq q_0 \), then \((q^*, i^*)\) is a NE Further, if \( \hat{q}(i^* + 1) \geq \frac{S-i^*}{S} \), \((q^*, i^*)\) is the unique NE; otherwise, we have two NE: \((q^*, i^*)\) and \((\hat{q}(i^* + 1), i^* + 1)\).

Case (ii): If \( \hat{q}(i^* + 1) \geq q_0 \), then \((q_0, i(q_0))\) is the unique NE.

Case (iii): If \( q^* > q_0 > \hat{q}(i^* + 1) \) and \( i(q_0) = i^* \), then \((q_0, i(q_0))\) is a NE. Further, if \( \hat{q}(i^* + 1) \geq \frac{S-i^*}{S} \), then \((q_0, i(q_0))\) is the unique NE; otherwise, we have two NE: \((q_0, i(q_0))\) and \((\hat{q}(i^* + 1), i^* + 1)\).

Case (iv): If \( q^* > q_0 > \hat{q}(i^* + 1) \) and \( i(q_0) = i^* + 1 \), then \((\hat{q}(i^* + 1), i^* + 1)\) is the unique NE.

![Figure 2: NE of the Patient Response Game for SBS](image)
As we can see for the basic model, up to two NE may exist, one with lower overbooking amount but higher patient show-up rate while the other with higher overbooking amount but lower show-up rate. Although there is no simple analytical answer as to which one is better for the clinic, by computing the net expected profit, one can compare them. In addition, the multiple equilibria are caused by the integer constraints. As $S$ increases, the impact of integer constraints diminishes, approaching a continuous case. From another perspective, as $S$ increases, the two equilibria get closer, approaching a unique equilibrium. In the next section, we discuss the basic model with multiple block scheduling.

2.3 The Basic Model with Multiple Block Scheduling

Different from the SBS model which treats the whole session as a single block, the multiple block scheduling (MBS) model divides the whole session into several blocks and schedules patients in different blocks. Patients are then arranged to arrive at the beginning of their scheduled block so their waiting time is reduced. We focus on equal-length blocks, i.e. each block originally has the same number of patients, since it is one of the most typical scheduling fashion (Cayirli and Veral, 2003) and assume that the NSOB policy is applied to each block. Nevertheless, as will be seen, since the show-up rate is different for each block, there will be different number of patients in each block after overbooking.

A significant complication in modeling MBS under overbooking is the necessity to consider the impact of overbooking on patient overflow from one block to its following block, i.e., patients that were not able to be seen in an earlier block will flow to the next block; see Muthuraman and Lawley (2008) and Zeng et al. (2010). Specifically, since a physician is treated as a single server and her patients are examined in a sequential fashion, the patients’ expected waiting time is determined by the number of patient arrivals in this block and the overflowing patients from previous block. Therefore, to characterize the NE of the MBS model, we must first capture the overflow between the blocks.

Note that the number of patients completed in a block with stochastic service time is a random variable. Typically, the study of the impact of overflow on the waiting time is very involved because of multiple complicated integrations. However, when the service time is exponentially distributed, due to its memoryless property, it is sufficient to count the number of patients overflowing into the next block to compute the expected waiting time for patients in that block. Next, we introduce the key parameters to describe such dynamics as well as the computing methods to drive their distributions. We direct readers to Muthuraman and Lawley (2008) for more detailed derivation.

Let $X_j, Y_j$ and $L_j$ denote the number of patients arriving in block $j$, the number of patients remain in clinic at the end of block $j$, and the number of patients served in block $j$. It is easy to see that

$$Y_j = (X_j + Y_{j-1} - L_j)^+.$$

To compute the expected waiting time, we define the patient arrival matrix $[Q_{j,t}]$ where $Q_{j,t}$ is the
probability that \( l \) patients arrive at the beginning of block \( j \), and the overflow matrix \([R_{j,k}]\) where \( R_{j,k} \) is the probability that \( k \) patients overflow from block \( j \) to block \( j+1 \). Note that the number of patient arrivals follows a binomial distribution and the number of patients served follows a Poisson distribution if the queue is not empty. So, for block \( j \), these matrices are computed as

\[
Q_{j,l} = \left\{ q_j^* \right\}^l \left\{ p_j^* \right\}^{(S+i_j^* - l)} \left( S + i_j^* \right)
\]

and

\[
R_{j,k} = \begin{cases} 
\sum_l \sum_m (1 - F_{L_j}(l + m))Q_{j,l}R_{j-1,m} & \text{if } k = 0 \\
\sum_l \sum_m f_{L_j}(l + m - k)Q_{j,l}R_{j-1,m} & \text{if } k \geq 1
\end{cases}
\]

where \( f_{L_j}(m) = e^{-\mu t}(\mu t)^m \) is the probability mass function of \( L_j \), following Poisson distribution with \( \mu = 1 \) in our case, and \( F_{L_j}(m) = \sum_{\tilde{m}=0}^{m-1} f_{L_j}(\tilde{m}) \).

Given these equations, especially the derivation for \([R_{j,k}]\), we have

\[
E[Y_j - 1] = \sum_k k R_{j-1,k}
\]

and therefore the expected waiting time for patients in block \( j \) is

\[
E[w_j] = \frac{1}{2}(S + i_j - 1)q_j + E[Y_j - 1].
\]

(9)

Using the same approach as used for SBS, we have the patient show-up rate in block \( j \) in response to the amount of overbooking for block \( j \), \( i \), as

\[
q_j(i) = \min\{\hat{q}_j(i), q_0\},
\]

(10)

where

\[
\hat{q}_j(i) = \frac{C_u + \alpha (w_0 - E[Y_{j-1}])}{C_l + C_u + \frac{\alpha}{2}(S + i - 1)}.
\]

(11)

To ensure \( \hat{q}_j(i) > 0 \), we add a very mild condition \( C_u > \alpha(E[Y_{j-1}] - w_0) \), which is mostly satisfied since \( E[Y_{j-1}] \) is generally small compared to \( C_u \), indicating that there are always some patients need to see doctors regardless of office waiting time.

Given (10), the equilibrium solution, for the case when \( q_0 > \hat{q}_j \), without considering the integer restriction, can be solved from equations (7) and (11) as

\[
\hat{q}_j^c = 1 + \frac{C_l + C_u}{\alpha S} - \frac{1}{2S} - \sqrt{(1 + \frac{C_l + C_u}{\alpha S}) - \frac{1}{2S^2} - \frac{2(C_u + \alpha (w_0 - E[Y_{j-1}]))}{\alpha S}}.
\]

(12)

Comparing (11) with (6) and (12) with (8), we find that \( \hat{q}_j(i) \) as well as \( \hat{q}_j^c \) are different from \( \hat{q}(i) \) and \( \hat{q}_j^c \) in the SBS model only by a constant related to \( E[Y_{j-1}] \). In other words, the geometric properties of the best response functions captured in the SBS model still hold in the MBS model. So, given \( E[Y_{j-1}] \), we can obtain the following results of NE in block \( j \) of the MBS model.
Theorem 3. Given $E[Y_{j-1}]$, denote $i_j^* = i(\hat{q}_j^*)$ and $q_j^* = \hat{q}_j(i_j^*)$ in MBS model, where $\hat{q}_j(\cdot)$ is defined in (11). Then in block $j$, if $C_u > \alpha(E[Y_{j-1}] - w_0)$, there exist at most two NE. Specifically,

**Case (i):** If $q_j^* \leq q_0$, then $(q_j^*, i_j^*)$ is a NE. Further, if $\hat{q}_j(i_j^* + 1) \geq \frac{s - i_j^*}{s}$, $(q_j^*, i_j^*)$ is the unique NE; otherwise, there are two NE: $(q_j^*, i_j^*)$ and $(\hat{q}_j(i_j^* + 1), i_j^* + 1)$.

**Case (ii):** If $\hat{q}_j(i_j^* + 1) \geq q_0$, $(q_0, i(q_0))$ is the unique NE.

**Case (iii):** If $q_j^* > q_0 > \hat{q}_j(i_j^* + 1)$ and $i(q_0) = i_j^*$, $(q_0, i(q_0))$ is a NE. Further, if $\hat{q}_j(i_j^* + 1) \geq \frac{s - i_j^*}{s}$, $(q_0, i(q_0))$ is the unique NE; otherwise, we have two NE: $(q_0, i(q_0))$ and $(\hat{q}_j(i_j^* + 1), i_j^* + 1)$.

**Case (iv):** If $q_j^* > q_0 > \hat{q}_j(i_j^* + 1)$ and $i(q_0) = i_j^* + 1$, $(\hat{q}_j(i_j^* + 1), i_j^* + 1)$ is the unique NE.

As described in the above theorem, given the expected waiting time from the patients overflowing from the previous block, we can calculate the equilibrium of block $j$. Hence, we can solve the NE sequentially. Specifically, start from block 1. Since there is no patient overflow for block 1, we can solve for the NE of this block following Theorem 2 for the SBS model. Then, given the NE of block 1, we can calculate the corresponding $E[Y_1]$. In case there are multiple NE, there is a different $E[Y_1]$ corresponding to each NE. Then, given $E[Y_1]$, we can calculate the NE of block 2 according to Theorem 3. Then we calculate $E[Y_2]$, based on which we can calculate the NE of block 3. We continue this process until we have solved the NE for all blocks.

3. The Integrated Model with Appointment Delay

After building the basic model which captures the impact of overbooking on patient no-show through the increased office delay, we incorporate the impact of overbooking through the decreased appointment delay. To do this, we need to model the queue for obtaining an appointment. There has been little operations literature on appointment delay except Green and Savin (2008), referred to as GS in the rest of this paper. In GS, the authors develop a queuing model to analyze the relationship between panel size (the number of patients in the clinic’s patient population) and appointment delay, considering patient no-show. However, GS does not involve overbooking. In this section, assuming Poisson arrivals of the patient appointment requests and exponential service times, we adapt the $M/M/1/K$ queueing model developed in GS and incorporate it with the basic model to capture the impact of overbooking on both appointment delay and office delay, which in turn affect patient no-show.

Before we describe our model, we first briefly summarize the queueing model in GS for a better understanding of our adaption of that model. In the GS model, the authors conceptualize the appointment system as a single-server queueing system in which patients who will enter service have a state-dependent probability of not being served (no-show) and may rejoin the queue with a re-booking
probability, \( r \). The steady-state distribution of the queue length when a patient requests an appointment (appointment delay) is determined and the impact of panel size on the clinic’s performance (e.g., expected appointment delay) is studied. Specifically, let the state \( k \) be the queue length at the time when a patient requests an appointment (appointment delay). The authors assume an empirical state-dependent no-show rate as a function of \( k \) (equation (1) in GS), which, using our notation, can be rewritten as

\[
p(k) = p_{\text{max}} - (p_{\text{max}} - p_{\text{min}})e^{-k/C},
\]

where \( p_{\text{min}} \) reflects the minimum observed no-show rate when there is no appointment delay (i.e., \( k = 0 \)), \( p_{\text{max}} \in [p_{\text{min}}, 1] \) represents the maximum observed no-show rate, and \( C \) is a no-show appointment delay sensitivity parameter. Further, assuming that a patient who sees an appointment delay \( k \) exceeding a certain limit \( (K) \) will leave for service elsewhere, the authors use equations (17)-(18) in GS to compute the steady-state probabilities of the queue length \( k \), \( \pi(k), k = 0, 1, \ldots K \), for the \( M/M/1/K \) queue with state-dependent no-show rate. Adapting into our model, we use the following equations, (14)-(17), to compute the steady-state probabilities of the appointment delay \( k \), \( \pi(k), k = 0, 1, \ldots K \), with state-dependent no-show rate:

\[
(\lambda N + B \cdot S \cdot (1 - rp(k - 1)))\pi(k) = \pi(k - 1)\lambda N + \pi(k + 1)B \cdot S \cdot (1 - rp(k)),
\]

\[
\lambda N\pi(0) = \pi(1)B \cdot S \cdot (1 - rp(0)),
\]

\[
B \cdot S \cdot (1 - rp(K - 1))\pi(K) = \pi(K - 1)\lambda N,
\]

\[
\sum_{k=0}^{K} \pi(k) = 1,
\]

where \( k = 1, \ldots, K - 1 \), \( N \) is the clinic’s patient panel size, \( r \) is the probability that a no-show patient will make another appointment, \( \lambda \) is the patient arrival rate, \( B \) is the number of blocks in a scheduling session (for single block models, \( B = 1 \)), \( S \) is the scheduling capacity of each block (number of patients that can be seen) without overbooking. Compared with equations (17)-(18) in GS, equations (14)-(17) have two differences: (1) \( B \cdot S \) is used to replace all \( T^{-1} \) in GS equations, where \( T \) is defined in GS as the average service time. This is because we consider single block and multiple block scheduling as well as overbooking. It’s easy to see \( T^{-1} = B \cdot S \). (2) \( \gamma(k) \) in GS is replaced with \( p(k) \) due to different notation used in our paper.

### 3.1 Integrated Model for Single Block Scheduling (I-SBS)

In this section, we describe how we incorporate the adapted GS queueing model into the basic model for single block scheduling. Specifically, we model that no-show comes from three sources: appointment delay, office delay, and other unassignable reasons, which are independent of each other, i.e., \( p = p^a + p^o + p_0 \), where \( p^a \), \( p^o \), and \( p_0 \) represent the no-show rate caused by appointment delay, office delay,
and other unassignable reasons, respectively. Note that \( p_0 = \frac{C_l}{C_l + C_u} \), as we modelled in section 2.1.

Rewriting (13), we get \( p(k) = (1 - \eta)p_{\text{max}} + \eta p_{\text{min}} \), where \( \eta = e^{-k/C} \). Recall \( p_{\text{min}} \) is the minimum observed no-show rate, i.e., when there is no appointment delay \( (k = 0) \). Hence, \( p_{\text{min}} \) is not related to appointment delay. Let \( p_{\text{max}} = p_{\text{max}}^o + p^o + p_0 \) and \( p_{\text{min}} = p_{\text{min}}^o + p_0 \) since the office delay and the unassignable no-show are independent of the appointment delay. Collecting terms, we have

\[
p(k) = (1 - \eta)p_{\text{max}}^o + p^o + p_0
\]

where \( p_{\text{max}}^o = p_{\text{max}} - p_{\text{min}} \). Notice that if we do not specifically model office delay, i.e., lumping \( p^o \) into the unassignable factors \( p_0 \), the above equation reduces to the one in GS model. Hence, our model is an extension to the GS model to consider office delay (together with the appointment delay), particularly to analyze the impact of overbooking.

To incorporate overbooking, from (18) we write the patient no-show rate as

\[
p(k) = (1 - \beta)(p_{\text{max}} - p_{\text{min}}) + p^o + p_0
\]

where \( \beta = e^{-k/A} \) and \( A = \frac{\hat{S}}{S}C \) is the no-show appointment delay sensitivity parameter considering overbooking. The definition of \( A \) captures the benefits of overbooking in reducing the appointment delay. A patient who originally sees an appointment delay of \( k \) will see an appointment delay reduced to \( k \frac{\hat{S}}{S} \) after overbooking. In addition, \( p^o \) in equation (19) is the office delay captured in section 2.2.

To summarize, the patient show-up rate, \( q(k) \), considering the impact of overbooking on both delays, can be written as

\[
q(k) = 1 - p(k) = \frac{C_u}{C_l + C_u} - (1 - e^{-k\frac{\hat{S}}{S}})(p_{\text{max}} - p_{\text{min}}) - \frac{\alpha(E[w] - w_0)^+}{C_l + C_u},
\]

where

\[
E[w] = \frac{1}{2}(\hat{S} - 1)\bar{q},
\]

\[
\bar{q} = \sum_{k=0}^{K} q(k)\pi(k),
\]

\[
\hat{S} = S + [S(1 - \bar{q})].
\]

In the above equations, (21) is the expected office waiting time equation, (22) calculates the expected show-up rate, \( \bar{q} \), based on which we calculate how much to overbook in (23), and \( \pi(k), k = 0, 1, ..., K \) used in (22) are solved from equations (14)-(17) with \( S \) replaced by \( \hat{S} \) and \( B = 1 \). Hence, by solving equations (20)-(23) and (14)-(17), we can obtain the equilibrium solutions to the integrated model with single block scheduling.

\(^3\)Since office delay is the waiting time on the appointment day, it is independent of \( p^o \) or \( k \). Also, unassignable no-show is related to random factors. Therefore, \( p^o \), \( p^o \), and \( p_0 \) are independent of each other.
Before further analysis, we would like to discuss two different operating policies depending on whether the clinic schedules to fill the extra capacity resulted from overbooking. The difference in these two policies is reflected in the value of $K$ in the above derived equations.

Recall that $K$ was defined in GS as the queue length above which a patient who requests an appointment will seek medical service elsewhere. For example, assuming the scheduling capacity is $S$ patients per day and a patient seeing an appointment delay beyond 20 days will seek service elsewhere, then $K = 20S$. As we see, with overbooking, the daily scheduling capacity will become $\hat{S} \geq S$, hence a patient who originally sees an appointment delay of 20 days will see it reduced to $20\hat{S}/S$ days after overbooking. Therefore, patients originally seek medical service elsewhere may stay for appointments, i.e., if we still use 20 days as the threshold of seeking service elsewhere, $K$ is now changed to $20\hat{S}/S$. This is referred to as the adjusted-$K$ scenario. As can be seen, with adjusted-$K$, all extra capacity resulted from overbooking is filled. On the other hand, clinics may control $K$ such that not all extra capacity is used up for more patients. A possible scenario is to keep $K$ the same as that before overbooking by not allowing appointments to be made beyond certain days ($20\hat{S} / \hat{S}$ days), referred to as the constant-$K$ overbooking scenario. Although adjusted-$K$ is what is commonly practiced, we will show in section 4 that constant-$K$ overbooking, an easy change from adjusted-$K$, surpasses adjusted-$K$ in its performance.

The complicated inter-relationship among equations (20)-(23) makes it extremely difficult to analyze the existence and uniqueness of the NE to the I-SBS model. Fortunately, for the constant-$K$ overbooking scenario, we are able to characterize the NE when $r = 0$ (i.e., no-show patients do not make up the appointments they missed), as will be presented next.

### 3.2 Integrated Model with Single Block Scheduling and Constant $K$

In this section, we characterize the NE solutions for constant-$K$ overbooking when re-booking rate, $r$, equals 0, i.e., no-show patients will not reschedule the appointment they skipped. Specifically, with constant-$K$, when $r = 0$, the queuing model in GS reduces to the classical single server queuing model with fixed capacity, $K$. As a result, the steady-state probabilities, i.e., $\pi(k)$, are independent of the show-up rate $q(k)$ for $k = 0, \ldots, K$, i.e., we have $\pi(0) = \frac{1}{\sum_{l=1}^{\infty} \rho^l}$ and $\pi(k) = \frac{\rho^k}{\sum_{l=1}^{\infty} \rho^l}$ where $\rho = \lambda NT$. Furthermore, $\pi(k)$ is independent of $\hat{S} = S + i$ (recall $i$ is the amount of overbooking). Consequently, from (22), we can simplify $\overline{\tau}$ as:

$$
\overline{\tau} = \frac{C_u}{C_l + C_u} - (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} \left(1 - e^{-k\hat{S}/(S+i)}\right) \pi(k) - \frac{\alpha(E[w] - w_0)}{C_l + C_u} + \left(p_{\text{max}} - p_{\text{min}}\right) \sum_{k=0}^{K} e^{-k\hat{S}/(S+i)} \pi(k).
$$

(24)
Since (24) indicates that $\overline{q}$ is a function of $i$ given $E[w] = \frac{1}{2}(S + i - 1)\overline{q}$, we use $\overline{q}(i)$ to denote the expected show-up rate for a given $i$. Next, we analyze the properties of $\overline{q}(i)$ and the NE when considering only appointment delay (i.e., patients are insensitive to office delay) and when both delays are considered, respectively.

### 3.2.1 Constant-$K$ Overbooking with Appointment Delay Only

This case corresponds to the situation where $w_0$ is very large so that patient responses towards office delay can be ignored. In this case, the expected patient show-up rate, $\overline{q}(i)$, is an increasing function in $i$. Therefore, we can easily see that there exists a unique NE for the “continuous game” (the intersection of $\overline{q}(i)$ and $i_c(q)$), see Figure 3.

However, since $q(k)$ and hence $\overline{q}(i)$ involves exponential functions, there is no closed-form expressions to the NE. Next, to help characterize the NE when considering the integer requirement, although concavity/convexity of $\overline{q}(i)$ cannot be obtained, we establish bounds on its first derivative.

**Lemma 4.** $0 < \overline{q}(i)' < \frac{1}{S}.$

Based on Lemma 4, we characterize the NE when considering the integer requirement.

**Proposition 5.** Considering only appointment delay, Let $\overline{q}_c^*$ be the equilibrium show-up rate of the continuous game and $i^* = i(\overline{q}_c^*)$ be the corresponding integer overbooking level. For the constant-$K$ overbooking with $r = 0$, there is either no NE or a unique NE at $(\overline{q}(i^*), i^*)$, where $\overline{q}(i^*)$, calculated using (24), is the equilibrium patient show-up rate given the amount of overbooking, $i^*$. 

![Figure 3: Nash Equilibrium When Considering Only Appointment Delay (S = 5)](image)
Although Proposition 5 states that it is possible to have no NE, it is of a small probability as $S$ is sufficiently large in single block scheduling and hence the model can be treated as the continuous game.

### 3.2.2 Constant-$K$ Overbooking Considering Appointment Delay and Office Delay

When considering the impacts of overbooking on both appointment delay and office delay, $\overline{q}(i)$ becomes the minimum of two differentiable functions, i.e.,

$$\overline{q}(i) = \frac{C_u}{C_l + C_u} - (p_{\text{max}} - p_{\text{min}}) + (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} e^{-\frac{ks}{C(S+i)}} \pi(k) - \frac{\alpha(E[w] - w_0)^+}{C_l + C_u}$$

$$\overline{q}(i) = \min\{\overline{q}_1(i), \overline{q}_2(i)\},$$

where $\overline{q}_1(i) = \frac{C_u}{C_l + C_u} - (p_{\text{max}} - p_{\text{min}}) + (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} e^{-\frac{ks}{C(S+i)}} \pi(k)$

and

$$\overline{q}_2(i) = \frac{C_u}{C_l + C_u} - (p_{\text{max}} - p_{\text{min}}) + (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} e^{-\frac{ks}{C(S+i)}} \pi(k) - \frac{\alpha(E[w] - w_0)}{C_l + C_u}.$$

Since $\overline{q}_1(i)$ is obtained when $(E[w] - w_0)^+ = 0$, it reduces to the expected show-up rate when considering only the appointment delay (analyzed in last subsection). For $\overline{q}_2(i)$, we next derive the bounds on its first derivative which helps us to identify the NE of the continuous game.

**Proposition 6.** $-\frac{1}{2S} < \overline{q}_2(i)' < +\infty$ for $i \in [0, S]$.

Based on Proposition 6, we can easily see that there exists at most one intersection of $\overline{q}_2(i) < 1$ and $i_c(q)$. We use $\overline{q}_{2c}$ to denote the patient show-up rate at the intersections of $\overline{q}_2(i)$ and $i_c(q)$ and set $\overline{q}_{2c} = \infty$ if there is no such intersection. Also, we use $\overline{q}_{1c}$ to denote the intersection of $\overline{q}_1(i)$ and $i_c(q)$, whose existence is guaranteed. Then, we have the following result.

**Proposition 7.** In the continuous game considering both delays, there exists a unique NE where the equilibrium show-up rate is $\overline{q}_c^* = \min\{\overline{q}_{1c}, \overline{q}_{2c}\}$.

Let $i^* = i(\overline{q}_c^*)$, based on Propositions 5 and 6, we can characterize the NE when considering both delays and the integer requirement.

**Proposition 8.** For the game with integer restriction considering both delays, there exist at most two NE. If there exists a unique NE, it is $(\overline{q}(i^*), i^*)$. If two NE exist, they are $(\overline{q}(i^*), i^*)$ and $(\overline{q}(i^*+1), i^*+1)$.

As we mentioned, due to the complexity of the equations, no closed-form solutions to the NE can be obtained. Fortunately, a search algorithm (Algorithm 1) can be developed which offers quick solutions to equations (20)-(23), providing a numerical way to obtain the equilibrium. Algorithm 1 and its convergence test are presented in the Appendix.
3.3 Integrated Model for Multiple Block Scheduling (I-MBS)

When we move on to multiple block scheduling considering appointment delay, the analysis and calculations are much more complicated. The biggest challenge is that some parameters are calculated for specific blocks and other parameters are aggregated over all blocks. In the following, we will discuss this and other different challenges of I-MBS.

First, for multiple blocks, we again have the patient overflow problem and the resolution is similar to what we discussed in Section 2.3. Specifically, given the expected overflow from block \( j - 1 \) to block \( j \), \( E[Y_{j-1}] \), which can be calculated following the same equations in Section 2.3, we can calculate the block-specific parameters for block \( j \) using (26)-(29), corresponding to equations (20)-(23) for the single block model.

\[
q_j(k) = \frac{C_u}{C_l + C_u} - (1 - e^{-k/A'}) (p_{\text{max}} - p_{\text{min}}) - \frac{\alpha (E[w_j] - w_0)^+}{C_l + C_u}, \quad (26)
\]

\[
E[w_j] = \frac{1}{2} (\hat{S}_j - 1) \bar{q}_j + E[Y_{j-1}] \quad (27)
\]

\[
\bar{q}_j = \sum_{k=0}^{K} q_j(k) \pi(k), \quad (28)
\]

\[
\hat{S}_j = S + [S(1 - \bar{q}_j)], \quad (29)
\]

and \( A' = \frac{\sum_{j=1}^{B} \hat{S}_j}{BS} C \). Note that \( \bar{q}_j \) denotes the expected show-up rate for patients in block \( j \) and \( A' \) denotes a weighted average appointment delay sensitivity parameter capturing the overall benefit of overbooking on reducing appointment delay. This is because although each block has different degree of overbooking (due to different office delay and patient show-up rate), appointment delay seen by patients requesting appointments (hence not yet scheduled) should not be block specific.

Due to the same reason, the steady-state probabilities of appointment delay should not be block specific either, yet the calculation of the steady-state probabilities are related to the no-show rates which are different in each block due to different office delay. To resolve this problem, we use a weighted no-show rate over all blocks, i.e.,

\[
p(k) = \frac{\sum_{j=1}^{B} p_j(k) \ast \hat{S}_j}{\sum_{j=1}^{B} \hat{S}_j} \quad (30)
\]

together with equations (14)-(17) where \( B \cdot S \) is replaced by \( \sum_{j=1}^{B} \hat{S}_j \) to calculate the steady-state probabilities of the appointment delay.

Due to the complexity mentioned above, analytical characterization of the NE is intractable, we again develop an algorithm (Algorithm 2) to solve for the block-specific as well as the all-block parameters iteratively to obtain the equilibrium solutions. Comparing with the calculations of the single block model (I-SBS), the complexity of the multiple block models causes much prominent convergence.
problem with I-MBS. In our numerical study, when convergency cannot be obtained within a reasonable time frame, the average of recent values is used. Algorithm 2, its brief outline, more discussion of its convergency, and numerical results of an example of a I-MBS are presented in the Appendix.

4. Numerical Study

In this section, we present results of a numerical study which investigates the overall impact of overbooking (considering its opposing effects on appointment delay and office delay) under different parameters. Based on these results, we obtain insights that lead to easy-to-implement strategies that can improve clinics’ performance.

In the numerical study, we focus on three parameters: panel size ($N$) which directly affects appointment delay, patients’ tolerance of office delay ($w_0$) which directly affects office delay, and patients’ unassignable no-show rate ($p_0$) which captures random factors indicating patients’ no-show tendency that are difficult to model. We also explore two other parameters, the degree of overbooking ($a$) and the no-show patient re-booking rate ($r$) when discussing clinics’ strategies.

We choose to focus on the single block models (although we observe similar insights from the multiple block models) since the figures are much cleaner due to few convergence issues. We do, however, include a sample figure of the multiple block model in the appendix (Figure A-2(b) shows the performance of a 4-block example of I-MBS with adjusted-$K$). When appropriate, we choose parameter values the same as or close to those used in GS which are derived from real data. Specifically, we set $C = 9$, $\lambda = 0.008$, $P_{max} = 0.36$, re-booking rate $r = 1$, $C_I = 15$, $\alpha = 3$, and $S = 20$ for I-SBS. We also set $S_j = 5$, $j = 1, \ldots, 4$ for the 4 blocks of I-MBS. We choose three levels of $w_0$ and two levels of $p_0$, i.e., $w_0 = \{5, 8, 11\}$ and $p_0 = \{\frac{1}{4}, \frac{1}{5}\}$. In addition, the mean service time is 1, the revenue of a regular time unit is 1, and the cost of an overtime unit is 2. Unless explicitly mentioned, overbooking (OB) denotes overbooking with adjusted $K$ in our numerical study.

Figure 4 compares the patients’ expected show-up rate (top) and clinic’s expected profit (bottom) with and without overbooking under different panel sizes and different patient tolerance of office delay ($w_0$). The figure shows that under adjusted-$K$ overbooking (a commonly used strategy in practice), patient show-up rate and clinic’s expected profit stay quite steady as the panel size changes except for a critical range, over which a small increase in the panel size leads to a sharp decrease in patient show-up rate and clinic’s expected profit. GS also observes this critical range when there is no overbooking and suggests that clinics should be careful about choosing their panel sizes because of this. With overbooking, we see that the critical range starts at a smaller panel size and the patient show-up rate and the clinic’s expected profit decrease in a less drastic fashion in the range. This indicates that

\footnote{$w_0 = 5$ means allowing 5 patients to be seen before you without incurring any disutility.}
overbooking mitigates the impact of panel size on patient show-up rate and expected profit (especially when patient tolerance is low).

In addition to the critical range, we make two important observations for overbooking with adjusted-$K$. First, although overbooking will likely increase clinic’s expected profit, it is not guaranteed to do so for panel sizes within the critical range, even when patients are quite tolerant of office delay. Further, overbooking always decreases patient show-up rate! Hence, no other measures taken, overbooking does not address the patient no-show problem. Therefore, instead of imposing higher and higher degrees of overbooking, clinics should think of other approaches to more effectively conduct overbooking. These observations seem somewhat puzzling given the fact that overbooking reduces the appointment delay as we discussed. A closer look provides reasons of the above observations as well as suggestions that can improve the performance after overbooking.

As we mentioned, with adjusted-$K$, clinics fill all the extra capacity made available from overbooking. Therefore, although the original patient group will see reduced appointment delay, the newly added patients, who are at the end of the queue, will still see long appointment delay (close to 20 days if we still use 20 days as the threshold of seeking service elsewhere). To see this more clearly, refer to Figure 5, in which we show expected appointment delay (top) and office delay (bottom) for different panel sizes under regular booking and overbooking with adjusted-$K$ and for different waiting time tolerance. As we can see, the expected appointment delay under adjusted-$K$ overbooking is always no lower than that under regular booking and the office delay is always higher with overbooking. As a result, with adjusted-$K$ overbooking, we will expect worse no-show rates because of the not-much-improved appointment delay and the increased office delay. On the other hand, if we control $K$ so that not all extra capacity resulted from overbooking is used up for more patients, we can better take advantage of the positive effect of overbooking on reducing the appointment delay and hence expect better results. For example, clinics can keep $K$ the same as that before overbooking (the constant-$K$ overbooking scenario) although choosing any $K$ between the original $K$ and the adjusted-$K$ would be more beneficial than using the adjusted-$K$ overbooking.

Figure 6 shows the same cases as in Figures 4, adding constant-$K$ overbooking. While the observations of the critical range are very similar to the adjusted-$K$ case, we observe key differences regarding the expected patient show-up rate and clinic’s expected profit. First, unlike the adjusted-$K$ case where overbooking always decreases patient show-up rate, with constant-$K$, overbooking may increase the patient show-up rate. In fact, comparison of the patient show-up curves under different office delay tolerance beautifully demonstrates the balance of the two opposing effects of overbooking: When patients are very tolerant of office delay ($w_0$ is high), the positive impact of overbooking on appointment delay is significant, exceeding its negative impact on office delay regardless of panel sizes, leading to higher show-up rates with overbooking under all panel sizes. In contrast, when patients are
intolerant of office delay ($w_0$ is small), the negative impact of overbooking on office delay is significant, exceeding its positive impact on appointment delay regardless of panel sizes, leading to lower show-up rates with overbooking under all panel sizes. When patients are moderately tolerant of office delay ($w_0$ is medium), the two effects are comparable and which effect takes the lead depends on the panel sizes, because the positive effect of overbooking on reducing appointment delay is higher with a larger panel size. When the panel size is large, the positive effect of overbooking on appointment delay exceeds its negative effect on office delay, but when the panel size is small, it cannot counter its negative effect on office delay. As a result, we see overbooking reduces patient show-up rates for small panel sizes but increases the patient show-up rates for large panel sizes.

As for the clinic’s expected profit, unlike the adjusted-$K$ case, constant-$K$ overbooking always improves clinic’s profitability due to its improvement on no-show rates. Further, overbooking increases clinic’s expected profit more for larger panel sizes because, as we mentioned, the positive impact of overbooking on the appointment delay is higher with large panel sizes.

The above observations indicate that when adopting overbooking, instead of filling all the extra capacity resulted from overbooking, clinics which impose some control over their appointment queue will see higher expected profits and improved patient no-show at the same time. A practical way of doing this is to not allow appointments made beyond certain days. This has a flavor of the open access scheduling method adopted by some clinics, with which patients call for appointments on the same day or the following day and will be turned down if there is no opening. However, instead of only allowing appointments for the the same day or the following day for open access (which implies little control of their demand on each day, depending on how many patients call for appointments), constant-$K$ is an improvement of the traditional scheduling method through overbooking and controlled queue length. Hence, constant-$K$ overbooking remains control over demand fluctuation, enjoys higher expected profits and potentially improved no-show rate as well.

In addition to the constant-$K$ overbooking with controlled appointment queue, we also explore another variable clinics may control, the degree of overbooking (how heavily clinics overbook). To reflect the degree of overbooking, we revise (3) to become

$$\hat{S} = S + [aSp],$$

where $a \in [0, 1]$ reflects the degree of overbooking with $a = 1$ representing full overbooking using NSOB and $a = 0$ representing no overbooking at all. Figure 7 compares expected patient show-up rate and clinic’s expected daily profit when $a = 0.0, 0.5, \text{ and } 1.0$, under adjusted-$K$ (in order to separate the effects of controlled queue length and the degree of overbooking). We observe a few interesting results. First, although lighter overbooking leads to higher patient show-up rates, overbooking always reduces (or at most keeps the same) patient show-up rates compared to no overbooking. In other words, using
adjusted-$K$, clinics cannot expect to adopt lighter overbooking to fix the problem of higher patient no-show rates. As for clinic’s expected profit, heavier overbooking always leads to higher profits. Further, the degree of overbooking has a bigger effect when the panel size is large (panel sizes beyond the critical range) since overbooking is more effective with higher patient no-show rate (occurring when panel size is large). Combining the above observations, we can see that clinics cannot achieve higher profits and better patient show-up rates at the same time with adjusted-$K$ overbooking by changing the degree of overbooking, but they can achieve both with constant-$K$ overbooking, as we discussed.

Next we explore two other important characteristics for clinics. The first one is the re-booking rate, $r$, defined in both the GS model and our model to be the probability that a no-show patient will reschedule the appointment he missed. For example, $r = 1.0$ indicates that a no-show patient is surely to reschedule his appointment and $r = 0.0$ indicates that a no-show patient will skip this appointment. Figures 8 and 9 explore the impact of the re-booking rate $r$. Specifically, Figure 8 shows the impact $r$ on clinic performance for different panel sizes under overbooking with adjusted $K$ (no control of appointment queue). We make two important observations. First, as $r$ increases, all characteristics of the figures remain the same except that the curves (expected patient show-up rate and clinic’s expected profit) are shifted left (i.e., critical range arrives early). This is because increasing $r$ is equivalent to increasing the demand or the panel size. Second, this shift can be significant, i.e., re-booking rate makes a big difference in the arrival of critical zone, which implies that a clinic may be able to handle a significantly larger panel size if the re-booking rate is low (e.g., as $r$ decrease from 1.0 to 0.5, the panel size may increase by $600 - 800$). Figure 9 reveals some other important results by comparing the impact of $r$ on clinic performance under different booking policies. First, comparing the regular booking with adjusted-$K$ overbooking shows that overbooking reduces the patient show-up rate, and the reduction increases as $r$ increases. This matches our intuition since an increase in $r$ means more no-shows will re-schedule their missed appointments, taking additional capacity, and may lead to no-show again. Second, comparing adjusted-$K$ with constant-$K$ overbooking, we see that constant-$K$ overbooking shows its advantage over adjusted-$K$ by increasing patient show-up rate and clinic’s expected profit, and delaying the arrival of critical zone (implying that clinics can handle larger panel size). In addition, this benefit is larger for clinics with a larger re-booking rate. As we mentioned before, overbooking can be used to stabilize clinic’s expected profit and such stabilization effect is also larger under higher $r$.

Finally, Figures 10(a) and 10(b) show the impact of the unassignable no-show rate, $p_0$, a parameter indicating patients’ no-show tendency, on the expected patient show-up rate and clinic’s expected profit. Both figures have the same parameters except the values of $p_0$ (Figure 10(a) shows a high $p_0$ and Figure 10(b) shows a lower $p_0$). Two important points stand out: (1) With higher $p_0$, the critical panel size range shifts left, i.e., clinics would experience sharp decrease of patient show-up rate at
smaller panel sizes. This is because, with a certain rescheduling rate \( (r) \), a higher no-show tendency translates to a higher demand/panel size as no-show patients reschedule their missed appointments.

(2) Overbooking is more effective and beneficial for patient population with a higher \( p_0 \). This is reflected in two observations: that overbooking brings higher improvements in the clinic’s expected profit when \( p_0 \) is higher and that the decrease of the patient show-up rate is less drastic within the critical range when \( p_0 \) is higher.
(a) \( w_0 = 5 \)
(b) \( w_0 = 8 \)
(c) \( w_0 = 11 \)

Figure 4: Expected Patient Show-up Rate and Profit for Adjusted \( K \) with \( p_0 = \frac{1}{7} \)

(a) \( w = 5 \)
(b) \( w = 8 \)
(c) \( w = 8 \)

Figure 5: Expected Appointment and Office Delays for Different Re-booking Rates when \( p_0 = \frac{1}{7} \)
Figure 6: Expected Patient Show-up Rate and Profits with $p_0 = \frac{1}{7}$

Figure 7: Expected Patient Show-up Rate and Profit for Different OB Levels when $p_0 = \frac{1}{7}$
Figure 8: Expected Patient Show-up Rate and Profit for Different Re-booking Rates when $p_0 = \frac{1}{7}$

(a) $w_0 = 5$
(b) $w_0 = 8$
(c) $w_0 = 11$

Figure 9: Expected Patient Show-up Rate and Profit for Different Re-booking Rates when $p_0 = \frac{1}{7}$ and $w_0 = 8$

(a) re-booking = 0
(b) re-booking = 0.5
(c) re-booking = 1
The second observation indicates that overbooking leads to higher improvement in the expected profit for clinics with higher unassignable no-show rates, indicating that the effectiveness of overbooking is different for patient population with different characteristics (in this case, their no-show tendency). Therefore, instead of the traditional strategy of overbooking all scheduling sessions, we encourage what we call a selective dynamic overbooking strategy with which clinics continuously monitor and classify patients based on their no-show records and determine accordingly whether to impose overbooking on the different classes of patients. For the “well-behaving” patients, since overbooking will cause extended waiting time and overtime, clinics may consider no or only light overbooking. For patients with consistently bad no-show records, e.g., the “habitual” no-show patients, clinics may put them into overbooking sessions with appropriate overbooking methods. By clearly communicating with the patients that different scheduling methods are adopted based on their no-show records to reduce office delay and dynamically monitoring the patients’ records to adjust their scheduling sessions, this strategy may serve as an incentive mechanism to improve patients’ show-up behavior and clinic’s expected profit at the same time. With lower unassignable no-shows, clinics can also potentially handle bigger panel sizes.

Discussion with different practitioners confirms that clinics do observe “habitual” no-show patients and that the selective dynamic overbooking is insightful. A similar approach was documented in a short article by a practitioner, Izard (2005), in which, without considering the motivating impact of this strategy (mentioned above), the authors report a significant reduction in patient no-shows due to this strategy. Similarly, Giachetti (2008) uses a simulation model to study the overbooking policy that is only applied to habitual no-show patients. Based on a real data set that has small portion of habitual no-show patients, he reports that this policy reduces the expected office waiting time while having little impact on appointment delay.
5. Discussion and Conclusion

Overbooking has been widely used in primary-care clinic scheduling to deal with the prevalent patient no-show problem. While overbooking has little impact on many random factors that affect no-show, it does impact two very important factors - appointment delay and office delay, which the management may be able to influence through careful administration of overbooking. In this paper, we develop a general model framework to analyze the impact of overbooking on these two important factors. While overbooking appears to reduce the appointment delay which could positively affect patient show-up rate, it increases the office delay which raises the disutility of patient appointments, hence negatively affects patients’ show-up rate. The overall impact of overbooking depends on the relative magnitude of these two effects. Our analysis has provided a few interesting and important insights which can help clinics better understand overbooking and improve their performance.

First, there exists a critical panel size range (with and without overbooking), in which both the patient show-up rate and the clinic’s expected profit experience sharp decreases as the panel size increases. However, with overbooking, patient show-up rate and clinic’s expected profit decrease in a less drastic fashion in the critical range. Hence, overbooking mitigates the impact of panel size on show-up rate and expected profit (especially when patients are less tolerant to office delay) and can be used to cope with panel size/patient demand fluctuation and stabilize clinic revenues.

Second, although overbooking will likely increase the clinic’s expected profit, it may reduce clinic’s expected profits for panel sizes within the critical range, even when patients are quite tolerant of office delay. Further, no other measures taken, overbooking always increases patient no-show rates (due to the increased office delay and not-much-improved appointment delay because clinics tend to fill the capacity made available from overbooking with more patients). In other words, no other measures taken, no-show problem is worse after overbooking! And, clinics cannot use lighter overbooking to fix the problem either. Therefore, instead of imposing higher and higher degrees of overbooking, clinics should think of other approaches to more effectively conduct overbooking. One strategy we propose is overbooking with controlled appointment queue, with which clinics will not allow appointments made beyond certain days while overbooking. Numerical results show that such simple change in the policy can help clinics achieve higher expected profits and better patient show-up rates at the same time.

Third, overbooking leads to higher improvement in the expected profit for clinics with higher unassignable no-show rates, indicating that the effectiveness of overbooking is different for patient population with different characteristics (in this case, their no-show tendency). Therefore, instead of the traditional strategy of overbooking all scheduling sessions, we encourage what we call a selective dynamic overbooking strategy, which may again help clinics achieve higher expected profits and better patient show-up rates at the same time.
Fourth, the no-shows’ probability to re-schedule their missed appointments \( r \) is an important factor to consider when ministering overbooking because they take additional capacity and may lead to no-show again. An increase in \( r \) may imply a significant reduction in the panel size a clinic may be able to handle. Further, the decrease of patient show-up rate after overbooking (with no control of appointment queue) increases as \( r \) increases. We also see that constant-\( K \) overbooking shows its advantage over adjusted-\( K \) by increasing patient show-up rate and clinic’s expected profit, and delaying the arrival of critical zone (implying that clinics can handle larger panel size). In addition, this benefit is larger for clinics with a larger re-booking rate.

Finally, in the current study, we have focused on a simple overbooking strategy, NSOB, to demonstrate the general framework. We may apply this general framework to other more sophisticated overbooking strategies in the literature. However, analytical results will be very involved, if tractable, because of the complicated structure of these overbooking policies. In addition, in this work, we have proposed some useful strategies for clinics. There remain many interesting tactics in applying these strategies that demand future study. For example, for overbooking with controlled appointment queue, what would be the optimal controlled queue length? For the selective dynamic overbooking strategy, what threshold of no-show records should be used to classify the patients (for no overbooking) in order to optimize the clinic’s objective function? These studies will provide further insights for clinics.

References


Technical Appendices

A-1. Proof of Proposition 1

Proof. First, a more explicit expression of \( i(q) \) given in (4) is

\[
i(q) = \begin{cases} 
0 & \text{if } q = 1 \\
\frac{s-k}{s} & \text{if } \frac{s-k}{s} \leq q < \frac{s-k+1}{s}.
\end{cases}
\]  

(A-1)

In Figure 2(a), \( i(q) \) is the set of solid vertical lines and its continuous relaxation, \( i_c(q) \), is the solid straight line with a slope equal to \(-\frac{1}{s}\).

From Assumption 1, \( \hat{q}(0) = \frac{C_u + \omega_w}{C_l + C_u + \frac{s}{2}(S-1)} \leq 1 \). In addition, \( \hat{q}(S) = \frac{C_u + \omega_w}{C_l + C_u + \frac{s}{2}(2S-1)} > 0 \). Since \( \hat{q}(i) \) is a decreasing function in \( i \), there must be exactly one intersection of \( \hat{q}(i) \) and \( i_c(q) \) at \( \hat{q}(\hat{i}_c(q)) = \frac{s-k}{s} \), \( \forall \hat{i} > \hat{i}_c(q) \). From Figure 2(a), we can see that \( \hat{q}'(\hat{i}_c(q)) \geq -\frac{1}{s} \).

If \( \hat{q}(\hat{q}^c) \in \mathbb{Z}_+ \), we have that \( \hat{i}^* = \hat{i}(\hat{q}^c) \) and \( q^* = \hat{q}^c \) and \((q^*, i^*)\) is a NE. Otherwise, because \( \hat{q}'(i_c(q^c)) \geq -\frac{1}{s}, \forall i > i_c(q^c) \) (due to convexity) and based on (A-1), we can see that an intersection of \( \hat{q}(i) \) and \( i(q) \) is \((q(\hat{i}(\hat{q}^c)), i(\hat{q}^c))\), i.e. \((q^*, i^*)\). It follows that \((q^*, i^*)\) is a NE when \( q(i) = \hat{q}(i) \).

![Figure A-1: Proof of 2 NE](image)

In case of multiple NE, it is easy to see that \((q^*, i^*)\) is the one with the least value on \( i \). Further, because \( \hat{q}(i) \) is convex and \( \hat{q}'(i) > -\frac{1}{s}, \forall i > i_c(q^c) \), we have \( \hat{q}(i) > \frac{s-k}{s} \) for \( i > i_c(q^c) \), i.e., the curve \( \hat{q}(i) \) is above the line \( i_c(q), \forall i > i_c(q^c) \). Therefore, we can conclude that, as demonstrated in Figure A-1, if there exist multiple NE, the equilibrium number of patients to overbook are in the form of \( i^*, i^* + 1, i^* + 2, \ldots \) (no “holes” in between) since otherwise \( \hat{q}(i) \) will not be a convex function.
Next, we prove that there exist at most two NE by contradiction. We first assume that there exist three NE and they are in the form of \((q^*, i^*)\), \((q(i^* + 1), i^* + 1)\) and \((q(i^* + 2), i^* + 2)\). We introduce a linear function \(l(i)\) that connects two points, \((q_a, i^*)\) and \((q_b, i^* + 2)\), with \(q_a = \frac{S-i^*}{S}\) and \(q_b = \frac{S-i^*-2}{S}\). Clearly, the slope of this linear function is \(-\frac{1}{2S}\). Further, to have 3 NE, we must have \(q^* \geq q_a\) and \(q(i^* + 2) < q_b\) (see Figure A-1). Therefore, there exist at least one intersection of \(\hat{q}(i)\) and \(l(i)\). Let \((q_x, i_x)\) be the intersection with the largest value on \(i\). Clearly, we have \(i^* \leq i_x < i^* + 2\). As a consequence, we have \(\hat{q}'(i) < -\frac{1}{2S}\) for \(i > i_x \geq i^* + 1\).

However, by taking the first derivative of \(\hat{q}(i)\) with respect to (w.r.t.) \(i\), we have
\[
\hat{q}'(i) = -\frac{\alpha(C_u + \alpha w_0)}{2(C_i + C_u + \frac{\alpha}{2}(S + i - 1))^2},
\]i.e.
\[
\frac{-1}{\hat{q}'(i)} = \frac{2(C_i + C_u)^2}{\alpha(C_u + \alpha w_0)} + \frac{\alpha(S + i - 1)^2}{2(C_u + \alpha w_0)} + \frac{2(C_i + C_u)(S + i - 1)}{C_u + \alpha w_0}.
\](A-3)

Because of Assumption 1 \((C_i \geq \alpha w_0)\), \(\forall i \geq 1\), we have
\[
\frac{2(C_i + C_u)(S + i - 1)}{C_u + \alpha w_0} \geq 2(S + i - 1) \geq 2S.
\]
Plugging in (A-3), we have \(-\frac{1}{\hat{q}'(i)} > 2S\), i.e., \(\hat{q}'(i) > -\frac{1}{2S}\), \(\forall i \geq 1\). We reach contradiction with an earlier statement.

Note from (A-1) that when \(\hat{q}(i^* + 1) > \frac{S-i^*}{S}\) and convexity of \(\hat{q}(i)\), \(\hat{q}(i)\) cannot have an intersection with \(i(q)\) at \(i = i^* + 1\). Further, because \(\hat{q}'(i) > -\frac{1}{S}\) for any \(i > i_c(\hat{q})\), we can conclude that \(\hat{q}(i) > \frac{S-i^*+1}{S}\) for \(i \geq i^* + 2\). As a consequence, \(\hat{q}(i)\) can only have one intersection with \(i(q)\), which is \((q^*, i^*)\).

\[\square\]

A-2. Proof of Lemma 4

\textit{Proof.} Since \(\overline{f}(i)\) increases in \(i\), it is clear that \(\overline{f}(i)' > 0\). From (24), to obtain \(\overline{f}(i)'\), we need to study the first derivative of \(e^{-\frac{ks}{c(S+i)}}\) with regard to \(i\).
\[
(e^{-\frac{ks}{c(S+i)}})' = \frac{1}{e^{\frac{ks}{c(S+i)}}} \frac{ks}{c(S+i)^2}
\]
\[
= \frac{1}{\frac{ks}{c(S+i)}} + \frac{ks}{c^2(S+i)^2}/2 + \ldots + \frac{ks^n S^n}{c^n c^2(S+i)^2}/(n!) + \ldots C(S+i)^2
\]
\[
< \frac{1}{\frac{ks}{c(S+i)}} \frac{ks}{c^2(S+i)^2}/2 C(S+i)^2
\]
\[
= \frac{1}{\frac{ks}{c(S+i)}} (S+i) + \frac{ks}{2C}
\]
\[
< \frac{1}{2S}
\]
Because \( p_{\text{max}} - p_{\text{min}} < 1 \), we can easily conclude that \( \overline{q}(i)' < \frac{1}{25} \).

\[ \square \]

**A-3. Proof of Proposition 6**

**Proof.** From (24) and the definition of \( E[w] \), we have

\[
\overline{q}_{2} = \frac{C_{u}}{C_{l} + C_{u}} - (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{C(S+i)}} \pi(k)) - \frac{0.5 \alpha (S+i-1) \overline{q}_{2} - \alpha w_{0}}{C_{l} + C_{u}},
\]

which leads to

\[
\overline{q}_{2}(i) = \frac{C_{u} + \alpha w_{0} - (C_{l} + C_{u})(p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{C(S+i)}} \pi(k))}{C_{l} + C_{u} + 0.5 \alpha (S+i-1)}.
\]

Next, we show it is also bounded below. To simplify our exposition, let \( f(i) = (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{C(S+i)}} \pi(k)) \). Then, the first derivative of \( \overline{q}_{2}(i) \) is

\[
\overline{q}_{2}(i)' = \frac{-\alpha(C_{u} + \alpha w_{0}) + \alpha(C_{l} + C_{u}) f(i)}{2(C_{l} + C_{u} + 0.5 \alpha (S+i-1))^2} - \frac{(C_{l} + C_{u}) f'(i)}{C_{l} + C_{u} + 0.5 \alpha (S+i-1)}
\]

Since \( \overline{q}_{2}(i) \) is continuous in \( i \) and is always differentiable for \( i \geq 0 \), we have \( \overline{q}_{2}(i)' < \infty \) for \( i \in [0, S] \).

Next, we show it is also bounded below. To simplify our exposition, let \( f(i) = (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{C(S+i)}} \pi(k)) \). Then, the first derivative of \( \overline{q}_{2}(i) \) is

\[
\overline{q}_{2}(i)' = \frac{-\alpha(C_{u} + \alpha w_{0}) + \alpha(C_{l} + C_{u}) f(i)}{2(C_{l} + C_{u} + 0.5 \alpha (S+i-1))^2} - \frac{(C_{l} + C_{u}) f'(i)}{C_{l} + C_{u} + 0.5 \alpha (S+i-1)}
\]

The first inequality comes from Assumption 1 that \( C_{l} \geq \alpha w_{0} \), the second inequality comes from the facts that \( f(i) > 0 \), and the third inequality comes from Assumption 2 that \( C_{l} + C_{u} > \alpha \). Because it is easy to see that \( f(i) \) decreases in \( i \), i.e. \( f'(i) < 0 \), we can conclude that \( \overline{q}_{2}(i)' \geq -\frac{1}{25} \).

\[ \square \]

**A-4. Algorithm 1 for I-SBS and Its Performance**

**Algorithm 1.** For panel size \( N \), do
1. Randomly select initial values for $\pi_1(k)$ for $k = 0, \ldots, K$;

2. If $\max_{k=0,\ldots,K} \{|\pi_1(k) - \pi_2(k)|\} > \epsilon$ ($\pi_1(k)$ and $\pi_2(k)$ are used to see whether $\pi(k)$ has converged), do
   a. If $\bar{q}$ is not available, set $\bar{q} = \frac{C_p}{C_1 + C_u}$. Then set $\hat{S} = S + \lceil S(1 - \bar{q}) \rceil$;
   b. Compute expected waiting time $E[w]$ (equation (21)).
   c. For $k = 0, \ldots, K$, compute the show-up probability $q(k)$ using (20), hence $p(k) = 1 - q(k)$;
   d. Set $\pi_1(k) = \pi_2(k)$ and apply (14)-(17) to compute $\pi_2(k)$, $k = 0, 1, 2, \ldots, K$;
   e. Calculate the expected show-up rate $\bar{q}$ according to equation (22).

3. Compute the expected daily profit.

The solution convergence problem does not seem to be an issue in I-SBS. In the many cases we tested using the above algorithm for the numerical study, we always find convergency of the equilibrium solutions. In Figure A-2(a), we show an example of the convergence test for a single block model with the block size equal to 20. We randomly generated 100 sets of initial values for $\pi_1(k)$, $k = 0, 1, \ldots, K$ and apply Algorithm 1. Figure A-2(a) shows the expected patient show-up rates obtained from Algorithm 1 from the 100 sets of different initial values for different panel sizes. As seen from the figure, for a particular panel size, the converged patient show-up rates from different initial values are very close to each other with negligible differences. This could also be attributed to the relatively large capacity ($S$) in single block scheduling models, in which the integral impact (one cause of multiple equilibria) is relatively insignificant.

![Figure A-2: Convergence in I-SBS and Performance of a 4-Block I-MBS](image-url)
Algorithm 2. For panel size $N$, do

1. Select random initial values for $\pi_1(k)$ and $\pi_2(k)$, $k = 0, \ldots, K$ ($\pi_1(k)$ and $\pi_2(k)$ are used to see whether $\pi(k)$ has converged);

2. While $\max_{k=0,\ldots,K}\{|\pi_1(k) - \pi_2(k)|\} > \epsilon$, do

   a. For each block $j = 1$ to $B$, do

      I. if $\bar{q}_j$ is not available, set $\bar{q}_j = \frac{C_{u,v}}{C_{l,v}}$ and $\bar{q}_j = \bar{q}_j - 2\epsilon$ ($\bar{q}_j$ is used to see whether $\bar{q}_j$ has converged). Also, set $\hat{S}_j = S_j + \lceil S_j(1 - \bar{q}_j) \rceil$.

      II. While $|q_j - q_j'| > \epsilon$, do

         i. Set $\bar{q}_j = \sigma \bar{q}_j + (1 - \sigma)q_j'$ (where $\sigma \in (0,1)$ is a randomly selected constant) and update $\hat{S}_j = S_j + \lceil S_j(1 - \bar{q}_j) \rceil$;

         ii. Compute the expected waiting time $E[w_j]$ (equation (27)) with consideration of $E[Y_{j-1}]$ if $j \geq 2$, where the calculation of $E[Y_{j-1}]$ follows the same equations in Section 2.3;

         iii. For $k = 0, \ldots, K$, compute the show-up probability $q_j(k)$ using (26);

         iv. Set $q_j' = q_j$ and obtain $q_j$ using $\pi_2(k)$ from (28);

   III Compute overflow $E[Y_j]$;

   b. Use (30) to compute the weighted no-show rate $p(k)$ using $q_j(k)$ ($p_j(k) = 1 - q_j(k)$) and $\hat{S}_j$;

   c. Set $\pi_1(k) = \pi_2(k)$ and apply (14)-(17) to compute $\pi_2(k)$, $k = 0,1,2,\ldots,K$;

3. Compute the expected show-up rate $\overline{q}$ (hence no-show rate $\overline{p}$) and the expected daily profit using $E[Y_B]$ as the overtime.

The outline of the above algorithm is as follows. Recall that all blocks have different expected office waiting time but they share the same appointment delay process since one patient’s appointment delay generally cannot be related to the block to which she will be assigned. Given that, our algorithm will first compute the block-specific show-up rate, $q_j(k)$, for $j = 1, \ldots, B$, using (26), considering the patient overflow and waiting time, in a sequential manner. We then obtain the weighted show-up rate over all blocks, $q(k)$ (hence $p(k) = 1 - q(k)$), which is used in calculating the steady-state probabilities of the appointment delay, $\pi(k)$. Such procedure is executed iteratively until the convergence of $\pi(k)$.

Comparing with the calculations of the single block model (I-SBS), the complexity of the multiple block models is significantly increased. In particular, the convergence problem could be much more prominent in the multiple block models. From the algorithm, we see that two sets of parameters need
to converge. The first set is $q_j(k)$ for each block for a given set of $\pi(k)$, $k = 0, 1, ...K$. The second set is $\pi(k)$ itself. Since the computation of $\pi(k)$ involves many parameters (e.g., the overflow effect $Y_j$ linking the blocks), plus the integer restriction on $\hat{S}_j$ for all blocks, the convergence of $\pi(k)$ may not be easy to achieve. In our numerical study, when convergency cannot be obtained within a reasonable time frame, the average of recent values is used. Figure A-2(b) shows the computation results of the expected patient show-up rates and clinic’s expected profit for a 4-block model with a capacity of $S = 5$ patients in each block. Compared with figures of I-SBS, there is some randomness involved under I-MBS. But the overall behavior is very similar to those of I-SBS.