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

Bo Zeng • Hui Zhao • Mark Lawley

Industrial and Management Systems Engineering, University of South Florida
Krannert School of Management, Purdue University
Weldon School of Biomedical Engineering, Purdue University
bzeng@eng.usf.edu • *zhaoh@purdue.edu* • *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. In this paper, we develop a general framework to explore the impact of overbooking on two most important factors of patient no-show - appointment delay and office delay. While overbooking increases office delay (which negatively affects patient no-show rates), it reduces appointment delay (which positively affects patient no-show rates). Our results show, considering these 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. Based on our results, we propose two easy-to-implement strategies, overbooking with *controlled appointment queue* and *selective dynamic* overbooking. Both strategies can increase the clinic's expected profit and improve no-show rates at the same time.

1. Introduction

Currently, there are approximately 200,000 non-psychiatric outpatient clinics in the United States (US Census Bureau, 2002) which provide 80% to 90% patient care (Bodenheimer and Grumback, 2002) nation wide. Among all the factors that have significant adverse effect on clinic performance, 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, they cause interruptions in the scheduling process and patient flow and waste critical resources; on the other hand, they limit clinic accessibility to other patients, leading to lower staff productivity and reduced revenues for the 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 *late appointment cancelations*, which have nearly the same damage as no-shows and can be dealt similarly.

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). Although with different methods and focusing on different subject groups, the literature converges to two prominent factors/causes of patient no-show: appointment delay (the time between a patient requesting an appointment and his actually seeing a doctor) and patient dissatisfaction (Lacy et al., 2004; Galucci et al., 2005; Dyer, 2005). Many factors contribute to patient dissatisfaction, among which office waiting time (i.e., the amount of time patients wait in the office to see the doctor) is undoubtedly a major element, e.g., Camacho et al. (2006); Bar-dayan et al. (2002).

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.

Because of the successful stories in the airline industry, overbooking has been proposed and implemented in healthcare organizations to stabilize their 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 packages, e.g., *Encore2008* and *Spectrasoft*. To fully utilize the power of overbooking in maximizing profits, advanced overbooking policies, methods, and algorithms have also been developed by healthcare engineering researchers, see Kim and Giachetti (2006), Laganga and Lawrence (2007), Zeng et al. (2008), Muthuraman and Lawley (2008), Liu et al. (2009) 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[1] patient no-show and its implications have not been analyzed. For example, overbooking may have positive impact on patient show-up rate by reducing the appointment delay because of its potential of scheduling more patients. In fact, in Bibi et al. (2007), the authors report that the clinic used *managed* overbooking as an intervention tool to reduce appointment delay and observe that no-show rate decreased because of that.

[1] Primary-care is different from acute-care where patients usually have less choices in their show-up decisions. In addition, primary-care patients are usually recommended for future appointments, e.g., follow-up checks, long before the appointment date. 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 no penalty for no-show, many of them forget to cancel, causing real no-show.

On the other hand, overbooking may have a negative impact on patient no-show rate by increasing the patients' office waiting time (office delay). Specifically, as Bean and Talaga (1992) and Camacho et al. (2006) point out, patients make their show-up decisions by trading off the benefit (the 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 involved in keeping the appointment. As the waiting time increases, the costs to the patient correspondingly increase" (Bean and Talaga (1992)). "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). A similar statement also appears in 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 waiting time on patient no show decision can be particularly true since the utility of an office visit often changes 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. When the utility of the visit has dramatically decreased from the time of making the appointment to the time of the actual appointment, the disutility from office delay easily outweighs the utility of the appointment. 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[2].

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

[2] It is noted that for the patients, one way to counter the 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 waiting time. 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.

Another way to counter the long office waiting time is switching healthcare providers. However, since there are many issues involved in switching providers, e.g., insurance network, patient-provider relationship, and provider specialty, this is not usually an option for many patients, nor is this discussed in much literature. In the limited literature that discusses switching providers, it is shown that only a small percentage of patients voluntarily switch providers (involuntary switching includes those caused by moving, patient referrals, etc.). For example, a survey of Medicare patients with a sample size of 1647 Rice et al. (1992) indicates that only 5.4% of the patients voluntarily switched providers and 2% (voluntary switching) was due to patients' dissatisfaction with the quality of care or the doctor's personality.

delay. While most of the operations literature on overbooking focuses on developing overbooking methods to maximize clinics' profits assuming unchanged show-up rates after overbooking (e.g., Muthuraman and Lawley (2008)), the contribution of our work is to develop a general framework to evaluate overbooking policies considering its impact on patient no-show through its effect on both appointment delay and office delay.

Specifically, 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, referred to as the basic model, and then incorporate into the basic model the impact of the appointment delay (referred to as the integrated model). By changing the patients' tolerance to office delay, we can adjust the weight on the impact of the office delay in the integrated model. When setting the tolerance parameter to an extreme value, the integrated model considers only the impact of the appointment delay. Hence, our model framework can be used to capture the impact of only the office delay (the basic model), only the appointment delay, or both delays.

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 *always* increases 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.

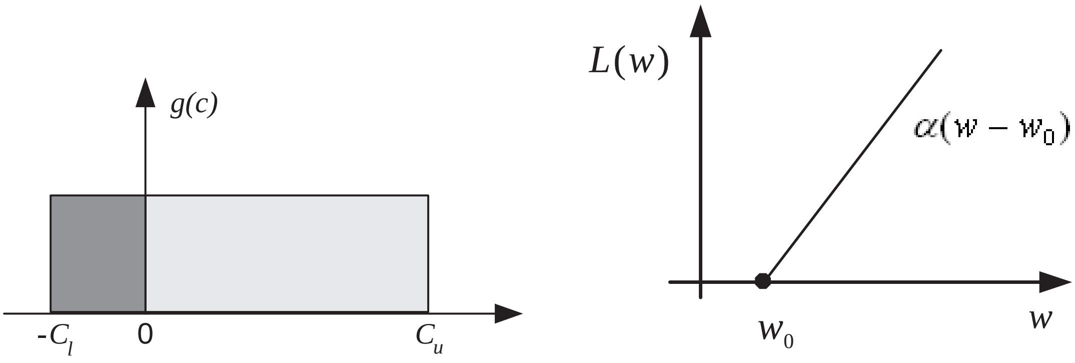
The rest of the paper is organized as follows. In Section 2, we develop the basic model that only considers the impact of office delay. We derive the Nash equilibrium for both single lock scheduling and multiple block scheduling. In Section 3, we incorporate the queueing model of appointment delay into the basic model for the single block and multiple block scheduling. We characterize the Nash equilibrium solutions for some special cases. 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 useful suggestions to the 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 the office delay caused by overbooking and patient no-show, the basic model. We first introduce the basic elements and conceptually develop the basic model, using a single block scheduling model for demonstration. We then characterize the solutions to the basic model under Single Block Scheduling (SBS) and Multiple Block Scheduling (MBS), respectively. Throughout this paper, the service times are exponentially distributed with a mean normalized to one.

2.1 The Elements of the Basic Model

In the single block overbooking model, all patients are scheduled to arrive at the beginning of the block to see the physician, whose service time follows a known distribution with a mean normalized to 1. Clearly, if there are multiple patients scheduled in the block (which is 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), 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.



(a) Distribution of patient utility of a clinic visit

(b) Disutility function of waiting time w

Figure 1: Patients' Utility of A Clinic Visit and Disutility of Waiting Time

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 α indicates the patients' sensitivity to waiting time (see Figure 1(b)). A higher α implies the patients are more sensitive (having a higher disutility) to waiting time. We also assume that there exists a threshold of patients' full tolerance of waiting time, 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 their 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

$$p = \int_{-\infty}^{L(E[w])} g(c)dc = \frac{C_l + L(E[w])}{C_l + C_u} = p_0 + \frac{L(E[w])}{C_l + C_u} \tag{1}$$

and the patient show-up rate will be

$$q = 1 - p = \frac{C_u - L(E[w])}{C_l + C_u} \tag{2}$$

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 waiting time (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 (N.E.). The above dynamics can be captured by a game theoretic model between the clinic and the patient population. In the following, we capture the equilibrium solution to patient no-show rate after overbooking in the single block and multiple block scheduling.

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) policy (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 where

$$\hat{S} = \lceil S(1+p) \rceil = S + \lceil Sp \rceil \tag{3}$$

and "$\lceil \rceil$" is the ceiling function to ensure the integer requirement of $\hat{S}$. To simplify exposition, let i denote the number of patients that are overbooked, i.e., the number of patients slots added as additional capacity. Hence, given the current no-show rate, p, the clinic's decision, i, is written as a function of p (or q):

$$i(q) = \lceil Sp \rceil \equiv \lceil S(1-q) \rceil. \tag{4}$$

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 [3], i.e.,

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

[3] Although all patients are scheduled to arrive at the beginning of the block, we assume there is an ϵ difference in their arrival time, which determines their sequence in the queue.

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

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

where $a^+ \equiv \max\{0, a\}$ and the $()^+$ term captures whether the expected waiting time is greater than the tolerance w_0. Further simplification yields

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

where

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

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.

Equation (4) shows that for a given q, the clinic's overbooking amount, $i(q)$, is uniquely determined. Equation (5) shows that for a given i, the patient population's show-up rate, $q(i)$, is also uniquely determined. The following lemma presents additional properties of $i(q)$ and $q(i)$ functions that will be used in analyzing the existence of the Nash Equilibrium (N.E.) in this patient response game.

Lemma 1. *$\hat{q}(i)$ is convex decreasing in i and the patient show-up rate, $q(i)$, is non-increasing quasi-convex in i for $i \in Z^+$ (non-negative integer). In addition, the clinic's overbooking function $i(q)$ is a discontinuous quasi-concave function with respect to (w.r.t.) q.*

All proofs are provided in the Appendix.

From Lemma 1, we see that although $q(i) = min\{\hat{q}(i), q_0\}$ is not well-behaved, $\hat{q}(i)$ is convex in i. Hence, to establish the N.E., 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 α 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 N.E. 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.,

$$i_c(q) = S(1-q). \tag{7}$$

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

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

Thus, the unique N.E. for the continuous game is $(\min\{\hat{q}^c, q_0\}, i_c(\min\{\hat{q}^c, q_0\}))$, where $i_c(\cdot)$ is given in (7).

When considering the integer restriction, we need to further study the structure of the best response functions. The following proposition identifies a possible N.E. for the basic model.

Proposition 2. *Let $i^* = i(\hat{q}^c)$ and $q^* = \hat{q}(i^*)$, where $i(\cdot)$ and $\hat{q}(\cdot)$ are defined in* (4) *and* (6)*, respectively. Then (q^*, i^*) is a N.E. for the basic model with SBS when $q_0 > \hat{q}(i)$, i.e., $q(i) = \hat{q}(i)$, for all i.*

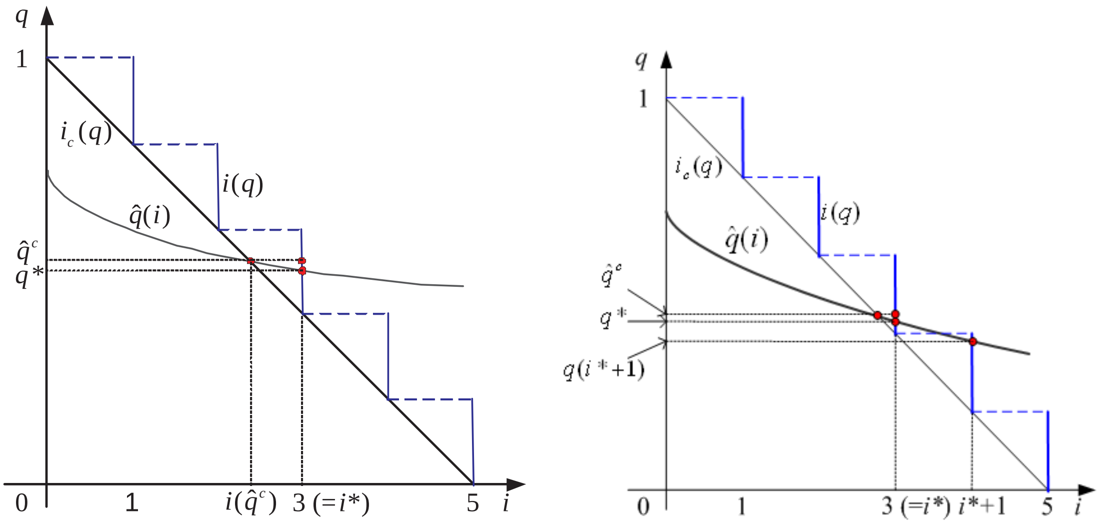
Figure 2(a) depicts (q^*, i^*). In the figure, $i(q)$ is the set of solid vertical lines. Its continuous relaxation, $i_c(q)$, is the solid straight line with a slope equal to $-\frac{1}{S}$.

Further, because $i(q)$ is neither concave nor convex, we may have multiple N.E. as demonstrated in Figure 2(b). The next proposition characterizes the multiple equilibria of the basic model with SBS.

Proposition 3. *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 N.E., (q^*, i^*); otherwise, there exist at most two N.E.: (q^*, i^*) and $(q(i^*+1), i^*+1)$.*

In the above, we have considered the N.E. 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 N.E. of the basic model with SBS, specified in the following theorem.

Theorem 4. *For the basic model with SBS, given (q^*, i^*) defined in Proposition 2, we have*



(a) Unique N.E.

(b) Two N.E.

Figure 2: N.E. of the Patient Response Game for SBS

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

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

As we can see for the basic model, up to two N.E. 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 since it is the most typical scheduling fashion (Cayirli and Veral, 2003) and assume that the NSOB policy is applied to each block.

A significant complication in modeling MBS is the necessity to consider the impact of overbooking on patient *overflow* from one block to its following block. 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 clock and the overflowing patients from previous block. Therefore, to characterize the N.E. 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.

Let L_j denote the number of patients served in block j. It is easy to see that

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

We next write down the expected waiting time following the approach of Muthuraman and Lawley (2008).

We first define the patient arrival matrix $[Q_{j,l}]$ where $Q_{j,l}$ 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 $j+1$. For block j, these matrices are computed as

$$Q_{j,l} = \{q_j^*\}^l \{p_j^*\}^{(S+i_j^*-l)} \binom{S+i_j^*}{l}$$

and

$$R_{j,k} = \begin{cases} \sum\limits_l \sum\limits_m (1 - F_{L_j}(l+m)) Q_{j,l} R_{j-1,m} & \text{if } k = 0 \\ \sum\limits_l \sum\limits_m f_{L_j}(l+m-k) Q_{j,l} R_{j-1,m} & \text{if } k \geq 1 \end{cases}$$

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

Given these equations, we have $E[Y_{j-1}] = \sum_k kR_{j-1,k}$ and therefore the expected waiting time for patients in block j is

$$E[w_j] = \frac{1}{2}(S + i_j - 1)q_j + E[Y_{j-1}]. \tag{9}$$

Having developed the expected waiting time, we are ready to characterize the N.E. of the basic model with MBS. 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

$$q_j(i) = \min\{\hat{q}_j(i), q_0\}, \tag{10}$$

where

$$\hat{q}_j(i) = \frac{C_u + \alpha(w_0 - E[Y_{j-1}])}{C_l + C_u + \frac{\alpha}{2}(S + i - 1)}. \tag{11}$$

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 compare to C_u, as shown in the numerical study.

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

$$\hat{q}_j^c = 1 + \frac{C_l + C_u}{\alpha S} - \frac{1}{2S} - \sqrt{(1 + \frac{C_l + C_u}{\alpha S} - \frac{1}{2S})^2 - \frac{2(C_u + \alpha(w_0 - E[Y_{j-1}]))}{\alpha S}}. \tag{12}$$

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 $\hat{q}^c$ 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. As a result, given $E[Y_{j-1}]$, we can obtain the following results of N.E. in block j of the MBS model.

Theorem 5. *Given $E[Y_{j-1}]$, denote $i_j^* = i(\hat{q}_j^c)$ 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 N.E.. Specifically,*

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

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

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 N.E.. Further, if $\hat{q}_j(i_j^* + 1) \ge \frac{S-i_j^*}{S}$, $(q_0, i(q_0))$ is the unique N.E.; otherwise, we have two N.E.: $(q_0, i(q_0))$ 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 N.E..*

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 N.E. sequentially for the MBS model. Specifically, start from block 1. Since there is no patient overflow for block 1, we can solve for the N.E. of this block following Theorem 4 for the SBS model. Then, given the N.E. of block 1, we can calculate the corresponding $E[Y_1]$. In case there are multiple N.E., there is a different $E[Y_1]$ corresponding to each N.E. Then, given $E[Y_1]$, we can calculate the N.E. of block 2 according to Theorem 5. Then we calculate $E[Y_2]$, based on which we can calculate the N.E. of block 3. We continue this process until we have solved the N.E. 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

$$p(k) = p_{max} - (p_{max} - p_{min})e^{-k/C}, \tag{13}$$

where p_{min} reflects the minimum observed no-show rate when there is no appointment delay (i.e., $k = 0$), $p_{max} \in [p_{min}, 1]$ 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, $\pi(k), k = 0, 1, ...K$, 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, $\pi(k), k = 0, 1, ...K$, with state-dependent no-show rate:

$$\begin{aligned}
(\lambda N + B \cdot S \cdot (1 - rp(k-1)))\pi(k) &= \pi(k-1)\lambda N + \pi(k+1)B \cdot S \cdot (1 - rp(k)), k = 1, \ldots, K-1, && (14)\\
\lambda N \pi(0) &= \pi(1)B \cdot S \cdot (1 - rp(0)), && (15)\\
B \cdot S \cdot (1 - rp(K-1))\pi(K) &= \pi(K-1)\lambda N, && (16)\\
\sum_{k=0}^{K} \pi(k) &= 1, && (17)
\end{aligned}$$

where N is the clinic's patient panel size, r is the probability that a no-show patient will make another appointment, λ 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[4]. Note that $p_0 = \frac{C_l}{C_l+C_u}$, as we modelled in section 2.1. Rewriting (13), we get

$$p(k) = (1-\eta)p_{max} + \eta p_{min},$$

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

$$p(k) = (1-\eta)p^a_{max} + p^o + p_0 \tag{18}$$

where $p^a_{max} = p_{max} - p_{min}$. 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

$$p(k) = (1-\beta)(p_{max} - p_{min}) + p^o + p_0, \tag{19}$$

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 $k\frac{S}{\hat{S}}$ 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

$$q(k) = 1 - p(k) = \frac{C_u}{C_l + C_u} - (1 - e^{-\frac{k}{C}\frac{S}{\hat{S}}})(p_{\max} - p_{\min}) - \frac{\alpha(E[w] - w_0)^+}{C_l + C_u}, \tag{20}$$

where

$$\begin{aligned} E[w] &= \frac{1}{2}(\hat{S} - 1)\overline{q}, & (21)\\ \overline{q} &= \sum_{k=0}^{K} q(k)\pi(k), & (22)\\ \hat{S} &= S + \lceil S(1 - \overline{q})\rceil. & (23) \end{aligned}$$

[4] Since office delay is the waiting time on the appointment day, it is independent of p^a or k. Also, unassignable no-show is related to random factors. Therefore, p^a, p^o, and p_0 are independent of each other.

In the above equations, (21) is the expected office waiting time equation, (22) calculates the expected show-up rate, $\overline{q}$, 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.

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 (i.e., above which a patient will be turned down an appointment when requesting). For example, assuming the scheduling capacity is S patients per day and a patient seeing an appointment delay beyond 20 days will seeking 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 only see it reduced to $20S/\hat{S}$ days after overbooking. Therefore, if allowed, 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 (referred to as the constant-K overbooking scenario) by not allowing appointments to be made beyond certain days ($20S/\hat{S}$ days). Although adjusted-K resembles more the real practice, we will show in section 4 that constant-K overbooking 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 Nash equilibrium to the I-SBS model. Fortunately, for the constant-K overbooking scenario, we are able to characterize the Nash equilibrium for some special cases, as will be presented next.

3.2 Integrated Model with Single Block Scheduling and Constant K

In this section, we characterize the Nash equilibrium solutions for constant-K overbooking when rebooking rate, r, equals 0, i.e., no-show patients skip this appointment. 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 $k = 0, \dots, K$, i.e., we have $\pi(0) = \frac{1}{\sum_{l=1}^{K} \rho^l}$ and $\pi(k) = \frac{\rho^k}{\sum_{l=1}^{K} \rho^l}$ 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 $\overline{q}$ as:

$$
\begin{aligned}
\overline{q} &= \frac{C_u}{C_l + C_u} - (p_{max} - p_{min}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{C(S+i)}}) \pi(k) - \frac{\alpha(E[w] - w_0)^+}{C_l + C_u} \\
&= \frac{C_u}{C_l + C_u} - (p_{max} - p_{min}) - \frac{\alpha(E[w] - w_0)^+}{C_l + C_u} + (p_{max} - p_{min}) \sum_{k=0}^{K} e^{-\frac{kS}{C(S+i)}} \pi(k). \qquad (24)
\end{aligned}
$$

Since (24) indicates $\bar{q}$ 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 Nash equilibrium when considering only appointment delay (i.e., patients are *in*sensitive 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, $\bar{q}(i)$, has the following property.

Lemma 6. *With constant K, $\overline{q}(i)$ is an increasing function in i.*

As in Section 2.2, we first consider the *"continuous game"* in which we ignore the integer requirements of i, Hence, $i(q)$ becomes $i_c(q)$, see equation (7). Note that when $q = 1$, $i_c(q) = 0$, and when $q = 0$, $i_c(q) = S$. Also, $0 < \bar{q}(i) < 1$. Therefore, we can easily see that the two monotonic functions $\overline{q}(i)$ and $i_c(q)$ will have one unique intersection, see Figure 3. As a result, we have the following proposition.

Proposition 7. *With constant K overbooking, for the "continuous game", there exists a unique NE.*

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

Lemma 8. $0 < \overline{q}(i)' < \frac{1}{2S}$.

Based on Lemma 8, we can characterize the Nash equilibrium when considering the integer requirement.

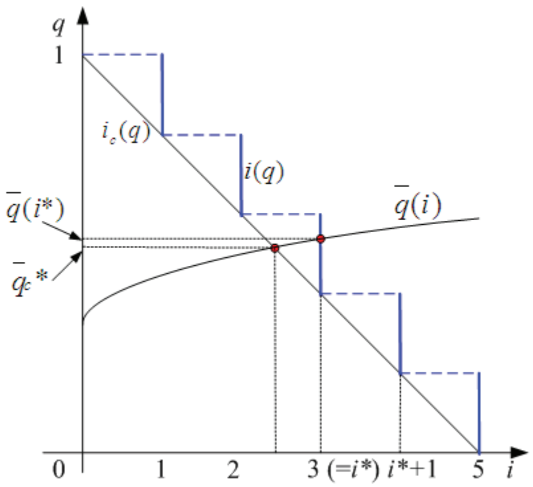


Figure 3: Nash Equilibrium When Considering Only Appointment Delay ($S = 5$)

Proposition 9. *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 ($i(\cdot)$ is shown in* (4)*). For the constant-K overbooking with $r = 0$, there is either no NE or one unique NE at $(\overline{q}(i^*), i^*)$, where $\bar{q}(\cdot)$, calculated using* (24)*, is the equilibrium patient show-up rate given the amount of overbooking, i^*.*

Although Proposition 9 states that it is possible to have no NE, this is of a small probability since 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.,

$$\begin{aligned} \overline{q}(i) &= \frac{C_u}{C_l + C_u} - (p_{max} - p_{min}) + (p_{max} - p_{min}) \sum_{k=0}^{K} e^{-\frac{kS}{C(S+i)}} \pi(k) - \frac{\alpha(E[w] - w_0)^+}{C_l + C_u}, \\ &= min\{\bar{q}_1(i), \bar{q}_2(i)\}, \end{aligned} \tag{25}$$

where

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

and

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

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 derive the bounds on its first derivative in the following proposition, which helps us to identify the N.E. of the continuous game.

Proposition 10. $-\frac{1}{2S} < \overline{q}_2(i)' < +\infty$ *for* $i \in [0, S]$.

Based on Proposition 10, 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 11. *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 in Propositions 9 and 10 and using the similar argument as that in the proof of Proposition 3, we can characterize the N.E when considering both delays and the integer requirement.

Proposition 12. *For the game with integer restriction considering both delays, there exist at most two NE. If a unique NE exists, 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 Nash equilibrium 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.

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

where

$$\begin{aligned} E[w_j] &= \frac{1}{2}(\hat{S}_j - 1)\overline{q}_j + E[Y_{j-1}] &(27)\\ \overline{q}_j &= \sum_{k=0}^{K} q_j(k)\pi(k), &(28)\\ \hat{S}_j &= S + \lceil S(1 - \overline{q}_j)\rceil, &(29) \end{aligned}$$

and $A' = \frac{\sum_{j=1}^{B} \hat{S}_j}{B\dot{S}} C$. Note that $\overline{q}_j$ 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.,

$$p(k) = \frac{\sum_{j=1}^{B} p_j(k) * \hat{S}_j}{\sum_{j=1}^{B} \hat{S}_j} \tag{30}$$

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

Due to the complexity mentioned above, analytical characterization of the N.E. 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: patient panel size (N) which directly affects appointment delay, patients' tolerance of office delay (w_0) which directly affects the 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 probability (r) when discussing clinic's strategies.

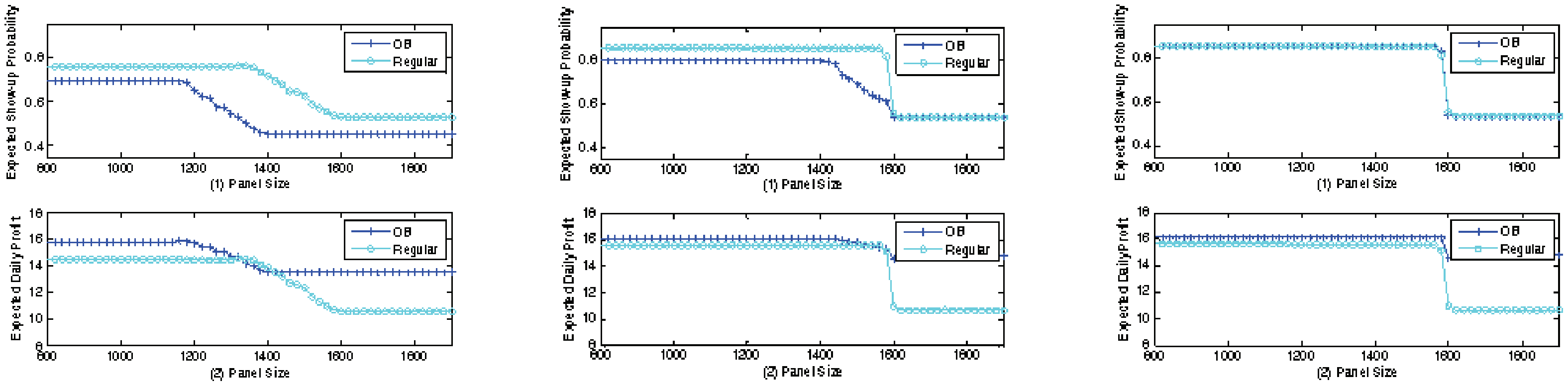
Although we observe similar insights from the multiple block models as in the single block models, we choose to focus on the single block models since the figures are much cleaner due to few convergence issues. We do, however, include a sample figure of the multiple block model in the appendix: Figure A-2(b) in the appendix shows the performance of a 4-block example of I-MBS with adjusted K. Except for some randomness, the overall behavior of the multiple block model is similar to that of a single block model. When appropriate, we choose the same parameter values as or close to the values used in GS which are derived from real data. Specifically, we set $C = 9$, $\lambda = 0.008$, $p^a_{max} = 0.36$, $C_l = 15$, $\alpha = 3$, and $S = 20$ for I-SBS. We also set $S_j = 5$, $j = 1,\ldots, 4$ for the 4 blocks of I-MBS. We look at all integer panel sizes from 700 to 1900 and choose three levels of w_0 and two levels of p_0, i.e., $w_0 = \{5, 8, 11\}$[5] and $p_0 = \{\frac{1}{4}, \frac{1}{7}\}$. 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.

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. Since adjusted-K overbooking requires no other measures and is the usual strategy in practice, it is the scenario we show. The figure shows that 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 ex-

[5] $w_0 = 5$ means allowing 5 patients to be seen before you without incurring any disutility.

pected 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 overbooking mitigates the impact of panel size on patient show-up rate and expected profit (especially when patient tolerance is low) and can be used to cope with panel size/patient demand fluctuation and stabilize clinic revenues.

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 improve 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.

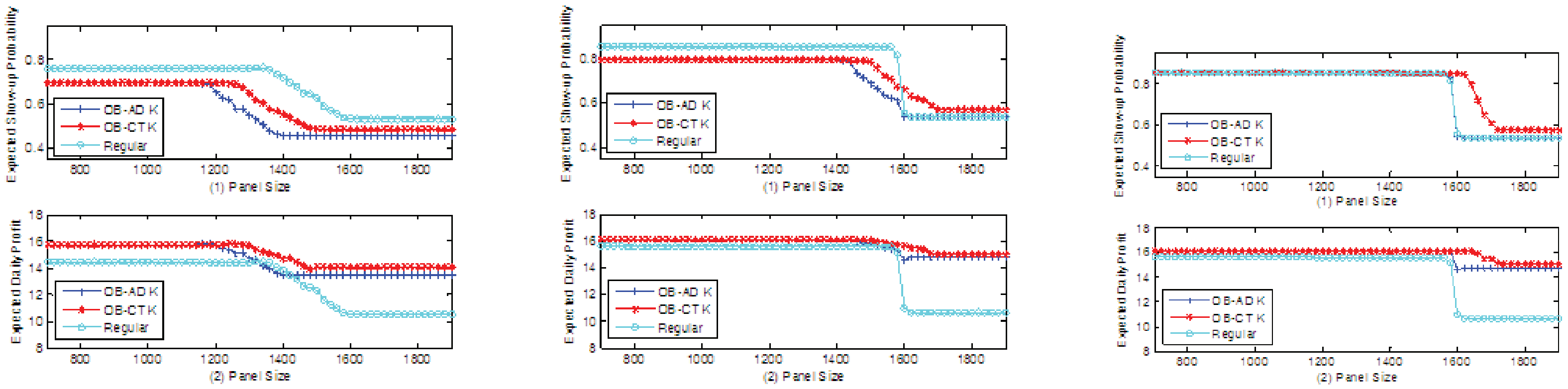


(a) $w_0 = 5$

(b) $w_0 = 8$

(c) $w_0 = 11$

Figure 4: Expected Patient Show-up Rate and Clinic's Expected Profit for Adjusted K with $p_0 = \frac{1}{7}$



(a) $w_0 = 5$

(b) $w_0 = 8$

(c) $w_0 = 11$

Figure 5: Expected Patient Show-up Rate and Clinic's Expected Profits with $p_0 = \frac{1}{7}$

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). 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 will be more beneficial than using the adjusted K.

Figure 5 shows the same cases as in Figures 4, adding constant-K overbooking (the original overbooking scenario in Figure 4 is the adjusted-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 in 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 all the time with overbooking. In contrast, when patients are very *in*tolerant 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 all the time with overbooking. 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.

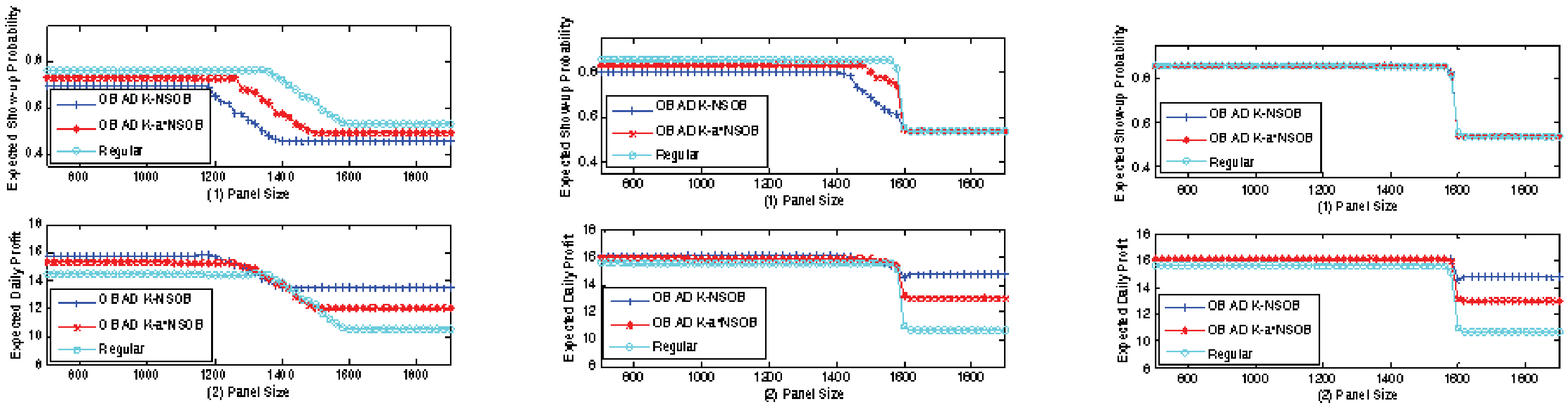
As for the clinic's expected profit, unlike in the adjusted-K case, constant-K overbooking always improves clinic's profitability due to its improvement on no-show rate. Further, overbooking increases clinic's expected profit more for larger panel sizes because, as we mentioned, the positive impact of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

The above observations indicate that when adopting overbooking, instead of filling 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 in practice, where patients call for appointments on the same day or the following day and will be turned down if there is no opening. However, a key difference lies in the fact that open access has little control of their demand on each day (depending on how many patients call for appointments), whereas 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 and enjoys higher expected profits and potentially improved no-show rate as well. It is easy to see that, if imposing overbooking on every day and keeping the queue length to two days, we result in open access as a special case.

While we propose constant-K overbooking with controlled appointment queue, we also explore another variable clinics may control, the degree of overbooking (how heavily clinics overbook), considering again both the appointment delay and office delay. To reflect the degree of overbooking, we revise (3) to become

$$\hat{S} = S + \lceil aSp \rceil,$$

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 6 compares expected patient show-up rate and clinic's expected daily profit when $a = 0$, $a = 0.5$, and $a = 1$, 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 (refer to the panel size 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.

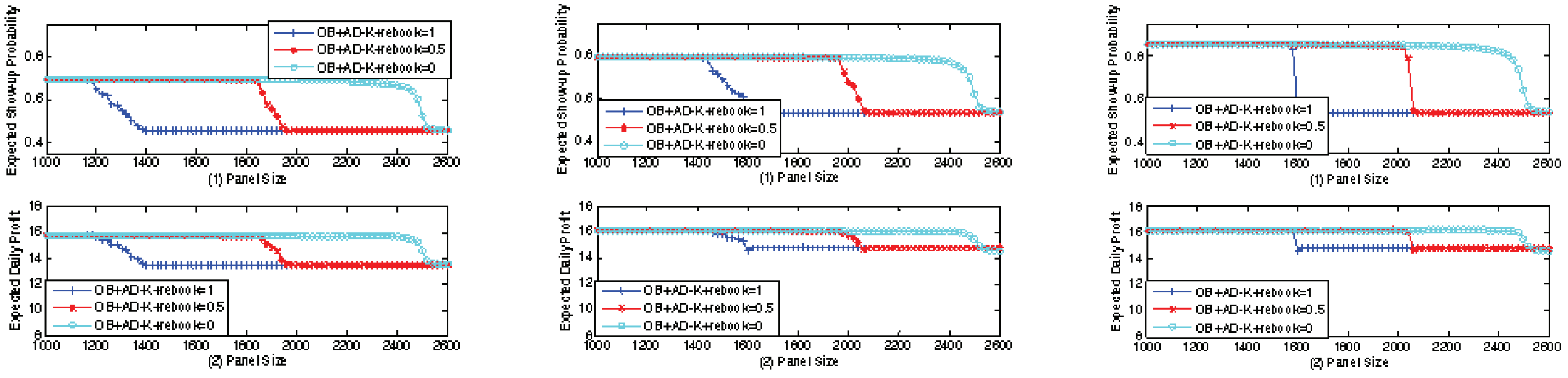


(a) $w_0 = 5$

(b) $w_0 = 8$

(c) $w_0 = 11$

Figure 6: Expected Patient Show-up Rate and Clinic's Expected Profit for Diff. OB Levels when $p_0 = \frac{1}{7}$



(a) $w_0 = 5$

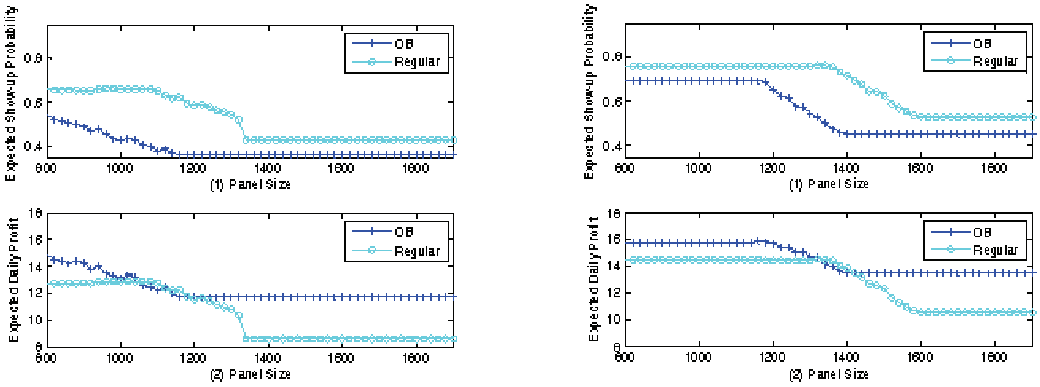
(b) $w_0 = 8$

(c) $w_0 = 11$

Figure 7: Expected Patient Show-up Rate and Clinic's Expected Profit for Diff. Re-scheduling Levels when $p_0 = \frac{1}{7}$

It is worth noting that in the GS model and our model, we used the parameter r to represent the probability that a no-show patient would 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. As we can see from Figure 7, when r is increased, 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 occurs at a smaller panel size) because increasing r is equivalent to increasing the demand or the panel size.

Finally, to see the impact of the unassignable no-show rate, p_0, a parameter indicating patients' no-show tendency, Figure 8(a) demonstrates expected patient show-up rate and clinic's expected profit when p_0 is high. Comparing Figure 8(a) with Figure 8(b), which has all the same parameters except a lower p_0, we notice two main differences: (1) Overbooking is more effective and more beneficial for patient population with a higher no-show tendency, 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. (2) With higher p_0, the critical panel size range shifts left (to lower panel sizes), i.e., clinics would experience sharp decrease of patient show-up rate at smaller panel sizes. This is because, with a certain rescheduling rate, a higher no-show rate translates to a higher demand/panel size as no-show patients reschedule their missed appointments. Understanding of the above differences leads to more effective adoption of overbooking (as will be discussed in the next section).



(a) $p_0 = \frac{1}{4}$

(b) $p_0 = \frac{1}{7}$

Figure 8: Patient Expected Show-up Rate and Clinic's Expected Profit for Diff. p_0 with $w_0 = 8$

5. Discussion and Conclusion

Overbooking has been widely used in primary-care clinic scheduling to deal with the prevalent patient no-show problem. In this paper, we develop a general model framework to analyze the impact of overbooking on the two prominent causes/factors of patient no-show: appointment delay and office delay (as a major element of patient dissatisfaction). While overbooking reduces the appointment delay which could positively affect patient show-up rate, it increases the office delay which raises the disutility of their 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 patient show-up rate and expected profit (especially when patients are not very tolerant of 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, i.e., no-show problem is worse after overbooking! And, clinics cannot use lighter overbooking to fix the problem either. This is because, although appointment delay is reduced by overbooking, the additional patients that are scheduled due to the extra capacity resulted from overbooking still see long appointment delay, canceling out the benefit of the shortened appointment delay for the original patients. 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. Such simple approach can help clinics achieve higher expected profits and better patient show-up rates at the same time, as seen in the numerical results.

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 with which clinics considers different strategies for different patient groups. Specifically, 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 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 show-up 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 no-show rates, 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 has been documented in a short article, 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.

Finally, in the current study, we have focused on a simple overbooking strategy, NSOB, to demonstrate the general framework. Following the same approach, we may study other more sophisticated overbooking strategies, such as those proposed in Kim and Giachetti (2006) and Muthuraman and Lawley (2008), to provide a more comprehensive evaluation of these overbooking strategies. 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 is 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? Further, a natural extension of the two-level *selective dynamic* strategy is to group the patients into multiple classes depending on their no-show records. Such strategy demands more study on the multi-threshold problem.

6. Acknowledgments

The authors thank *Regenstrief Center for Healthcare Engineering* at Purdue University for supporting this work. We also thank Dr. Deanna Ruth Willis and Diann Awrylo at Indiana University Medical Group (IUMG) and other physicians, administrators, and staff of the IUMG and Wishard Primary Care Clinic of Indianapolis, Indiana for the their interactions and comments.

References

Bar-dayan, Y., A. Leiba, Y. Weiss, J. Carroll, P. Benedek. 2002. Waiting time is a major predictor of patient satisfaction in a primary military clinic. *Military medicine* **167** 842–845.

Bean, A., J. Talaga. 1992. Appointment breaking: causes and solutions. *Journal of Health Care Marketing* **12** 14–25.

Bibi, Y., A. Cohen, D. Goldfarb, E. Rubinshtein, D. Vardy. 2007. Intervention program to reduce waiting time of a dermatological visit: Managed overbooking and service centralization as effective management tools. *International Journal of Dermatology* **46** 830–834.

Bodenheimer, T., K. Grumback. 2002. *Understanding Health Policy: A Clinical Approach*. Third edition ed. Lange Medical Books / McGraw-Hill,Medical Publishing Division, New York.

Camacho, F., R. Anderson, A. Safrit, A. Jones, P. Hoffmann. 2006. The relationship between patient's perceived waiting time and office-based practice satisfaction. *North Carolina Medical Journal* **67** 409–413.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: a review of literature. *Production and Operations Management* **12** 519–549.

Cayirli, T., E. Veral, H. Rosen. 2006. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science* **9** 47–58.

Dervin, J. V., D.L. Stone, C. H. Beck. 1978. The no-show patient in the model family practice unit. *Journal of Family Practice* **7**(6) 1177–1180.

Dyer, O. 2005. Sick of getting stood up? no-shows say it's because they need a little respect. *National Review of Medicine* **2**(1).

Galucci, G., W. Swartz, F. Hackerman. 2005. Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services* **56** 344–346.

Gang, Y., ed. 1998. *Operations research in the airline industry*. Springer.

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

Giachetti, R. 2008. A simulation study of interventions to reduce appointment lead-time and patient no-show rate. S. J. Mason, R. R. Hill, L. Mnch, O. Rose, T. Jefferson, J. W. Fowler, eds., *Proceedings of the 2008 Winter Simulation Conference*. 1463–1468.

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

Green, L., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56** 1526–1538.

Gupta, D, B.T. Denton. 2008. Health care appointment systems: Challenges and opportunities. *IIE Transactions* **40** 800–819.

Izard, T. 2005. Managing the habitual no-show patient. *Family Practice Managment* **12** 65–66.

Keir, L., B. Wise, C. Krebs. 2002. *Medical Assisting: Administrative and Clinical Competencies*. 5th ed. Thomson Delmar Learning.

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

Lacy, N., A. Paulman, M. Reuter, B. Lovejoy. 2004. Why we dont come: Patient perceptions on no-shows. *Annals of Family Medicine* **2** 541–545.

Laganga, L., S. Lawrence. 2007. Clinic overbooking to improve patient access and provider productivity. *Decision Sciences* **38** 251–276.

Liu, N., S. Ziya, V. Kulkarni. 2009. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *to appear in MSOM* .

Lowes, R. 2005. Practice pointers: How to handle no-shows. *Medical Economics* **82** 62–65.

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

Rice, T., L. Nelson, DC. Colby. 1992. Will medicare beneficiaries switch physicians? a test of economic competition. *Journal of Health Politics, Policy and Law* **17** 3–24.

Robinson, L., R. Chen. 2009. The effects of patient no-shows on appointment scheduling policies. *to appear in MSOM* .

Rothstein, M. 1985. OR and the airline overbooking problem. *Operations Research* **33** 237–249.

Rust, C., N. Clark, W. Clark, D. Jones, W. Wilcox. 1995. Patient appointment failures in pediatric resident continuity clinics. *Archives of Pediatrics and Adolescent Medicine* **149** 693–695.

Sharp, D., W. Hamilton. 2001. Non-attendance at general practices and outpatient clinics. *British Medical Journal* **323** 1081–1082.

Smith, B., J. Leimkuhler, R. Darrow. 1992. Yield management at american airlines. *Interfaces* **22** 8–31.

US Census Bureau. 2002. Economic census industry series reports: Health care and social assistance.

van Baar, J., H. Joosten, J. Car, G. Freeman, M. Partridge, C. van Weel, A. Sheikh. 2006. Understanding reasons for asthma outpatient (non)-attendance and exploring the role of telephone and e-consulting in facilitating access to care: exploratory qualitative study. *Quality and Safety in Health Care* **15** 191–195.

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

Zeng, B., A. Turkcan, J. Lin, M. Lawley. 2008. Clinic scheduling models with overbooking for patients of heterogeneous no-shows probabilities. *to appear in Annals of Operations Research* .

Technical Appendices

A-1. Proof of Lemma 1

Proof. From (6), we observe that the continuous relaxation of $\hat{q}(i')$ for $i' \in R^+$ (non-negative real number) converges to $\hat{q}(i)$ for $i \in \mathbb{Z}^+$ when i' approaches i. So, we first allow i to take any real value to study its structure (e.g., take derivatives) and then investigate its values on those integer points. Taking the first derivative of $\hat{q}(i)$ with respect to (w.r.t.) i, we have

$$\hat{q}'(i) = -\frac{\alpha(C_u + \alpha w_0)}{2[C_l + C_u + \frac{\alpha}{2}(S + i - 1)]^2}. \tag{A-1}$$

Thus, it is easy to see that $\hat{q}(i)$ is strictly convex decreasing in i for $i \in Z^+$. Further, since $q(i) = \min\{\hat{q}(i), q_0\}$ where q_0 is a constant, we know $q(i)$ is non-increasing and quasi-convex for $i \in Z^+$.

The result on $i(q)$ follows directly from (4) and the definition of quasi-concave function. □

A-2. Proof of Proposition 2

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

$$i(q) = \begin{cases} 0 & \text{if} \quad q = 1 \\ k & \text{if} \quad \frac{S-k}{S} \le q < \frac{S-k+1}{S}. \end{cases} \tag{A-2}$$

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, $\hat{q}(0) = \frac{C_u+\alpha w_0}{C_l+C_u+\frac{\alpha}{2}(S-1)} \le \frac{C_u+C_l}{C_l+C_u+\frac{\alpha}{2}(S-1)} \le 1$. In addition, $\hat{q}(S) = \frac{C_u+\alpha w_0}{C_l+C_u+\frac{\alpha}{2}(2S-1)} > 0$. 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 $(i_c(q^c), q^c)$, and at least one intersection of $\hat{q}(i)$ and $i(q)$. Further, because $\hat{q}(0) \le 1$, $\hat{q}(i)$ is a decreasing function in i and the slope of $i_c(q)$ is $-\frac{1}{S}$, from Figure 2(a), we can see that $\hat{q}'(i_c(\hat{q}^c)) \ge -\frac{1}{S}$. If $i(\hat{q}^c) \in \mathbb{Z}_+$, we have that $i^* = i(\hat{q}^c)$ and $q^* = \hat{q}^c$ and (q^*, i^*) is a N.E. Otherwise, because $\hat{q}'(i) > \hat{q}'(i_c(\hat{q}^c)) \ge -\frac{1}{S}, \forall i > i_c(q^c)$ (due to convexity) and based on (A-2), we can see that an intersection of $\hat{q}(i)$ and $i(q)$ is $(q(i(\hat{q}^c)), i(\hat{q}^c))$, i.e. (q^*, i^*). It follows that (q^*, i^*) is a N.E. when $q(i) = \hat{q}(i)$. □

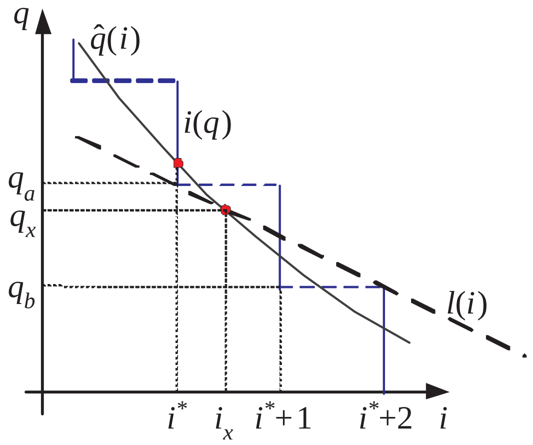


Figure A-1: Proof of 2 N.E.

A-3. Proof of Proposition 3

Proof. We use Figure A-1 to illustrate the proof. From the proof of Proposition 2, we know that in case of multiple N.E., (q^*, i^*) is the one with the least value on i. Further, because $\hat{q}(i)$ is convex and $\hat{q}'(i) > -\frac{1}{S}$, for $i > \hat{i}_c(q^c)$, we have $\hat{q}(i) > \frac{S-i}{S}$ 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, if there exist multiple N.E., 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 N.E. by contradiction. We first assume that there exist three N.E. 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 N.E., 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 $i > i_x \geq i^* \geq 1$.

However, from (A-1), we have

$$-\frac{1}{\hat{q}'(i)} = \frac{2(C_l + C_u)^2}{\alpha(C_u + \alpha w_0)} + \frac{\alpha(S+i-1)^2}{2(C_u + \alpha w_0)} + \frac{2(C_l + C_u)(S+i-1)}{C_u + \alpha w_0}. \tag{A-3}$$

Because of Assumption 1 ($C_l \geq \alpha w_0$), $\forall i \geq 1$, we have

$$\frac{2(C_l + C_u)(S+i-1)}{C_u + \alpha w_0} \geq 2(S+i-1) \geq 2S.$$

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-2) 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 $i > i_c(\hat{q}^c)$, 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^*). □

A-4. Proof of Lemma 6

Proof. From (20), it is easy to see that $q(k)$ is increasing in i. Since $\pi(k)$ are constants for a given K, from (22), we can see that $\overline{q}(i)$ is increasing in i. □

A-5. Proof of Lemma 8

Proof. Since $\overline{q}(i)$ increases in i, it is clear that $\overline{q}(i)' > 0$. From (24), to obtain $\overline{q}(i)'$, we need to study the first derivative of $e^{-\frac{kS}{C(S+i)}}$ with regard to i.

$$\begin{aligned}
(e^{-\frac{kS}{C(S+i)}})' &= \frac{1}{e^{\frac{kS}{C(S+i)}}} \frac{kS}{C(S+i)^2} \\
&= \frac{1}{1 + \frac{kS}{C(S+i)} + \frac{k^2S^2}{C^2(S+i)^2}/2 + \cdots + \frac{k^nS^n}{C^n(S+i)^n}/(n!) + \ldots} \frac{kS}{C(S+i)^2} \\
&< \frac{1}{1 + \frac{kS}{C(S+i)} + \frac{k^2S^2}{C^2(S+i)^2}/2} \frac{kS}{C(S+i)^2} \\
&= \frac{1}{\frac{C(S+i)^2}{kS} + (S+i) + \frac{kS}{2C}} \\
&\leq \frac{1}{(S+i) + \sqrt{2}(S+i)} \\
&< \frac{1}{2S}
\end{aligned}$$

Because $p_{max} - p_{min} < 1$, we can easily conclude that $\overline{q}(i)' < \frac{1}{2S}$. □

A-6. Proof of Proposition 10

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

$$\overline{q}_2 = \frac{C_u}{C_l + C_u} - (p_{max} - p_{min}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{C(S+i)}} \pi(k)) - \frac{0.5\alpha(S+i-1)\overline{q}_2 - \alpha w_0}{C_l + C_u}, \quad \text{(A-4)}$$

which leads to

$$\overline{q}_2(i) = \frac{C_u + \alpha w_0 - (C_l + C_u)(p_{max} - p_{min})\sum_{k=0}^{K}(1 - e^{-\frac{kS}{C(S+i)}}\pi(k))}{C_l + C_u + 0.5\alpha(S+i-1)} \quad \text{(A-5)}$$

$$= \frac{C_u + \alpha w_0}{C_l + C_u + 0.5\alpha(S+i-1)} - \frac{(C_l + C_u)(p_{max} - p_{min})\sum_{k=0}^{K}(1 - e^{-\frac{kS}{C(S+i)}}\pi(k))}{C_l + C_u + 0.5\alpha(S+i-1)} \quad \text{(A-6)}$$

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 $f(i) = (p_{max} - p_{min})\sum_{k=0}^{K}(1 - e^{-\frac{kS}{C(S+i)}}\pi(k))$. Then, the first derivative of $\overline{q}_2(i)$ is

$$\begin{aligned}
\overline{q}_2(i)' &= \frac{-\alpha(C_u + \alpha w_0) + \alpha(C_l + C_u)f(i)}{2[C_l + C_u + 0.5\alpha(S+i-1)]^2} - \frac{(C_l + C_u)f'(i)}{C_l + C_u + 0.5\alpha(S+i-1)} \\
&\geq \frac{-\alpha(C_l + C_u) + \alpha(C_l + C_u)f(i)}{2[C_l + C_u + 0.5\alpha(S+i-1)]^2} - \frac{(C_l + C_u)f'(i)}{C_l + C_u + 0.5\alpha(S+i-1)} \\
&= \frac{\alpha(C_l + C_u)(f(i) - 1)}{2[(C_l + C_u)^2 + \alpha(S+i-1)(C_l + C_u) + 0.25\alpha^2(S+i-1)^2]} \\
&\quad - \frac{(C_l + C_u)f'(i)}{C_l + C_u + 0.5\alpha(S+i-1)} \\
&= \frac{\alpha(f(i) - 1)}{2[(C_l + C_u) + \alpha(S+i-1) + 0.25\alpha^2\frac{(S+i-1)^2}{C_l+C_u}]} - \frac{(C_l + C_u)f'(i)}{C_l + C_u + 0.5\alpha(S+i-1)} \\
&\geq \frac{-\alpha}{2[(C_l + C_u) + \alpha(S+i-1) + 0.25\alpha^2\frac{(S+i-1)^2}{C_l+C_u}]} - \frac{(C_l + C_u)f'(i)}{C_l + C_u + 0.5\alpha(S+i-1)} \\
&\geq \frac{-\alpha}{2\alpha(S+i)} - \frac{(C_l + C_u)f'(i)}{C_l + C_u + 0.5\alpha(S+i-1)} \\
&= -\frac{1}{2(S+i)} - \frac{(C_l + C_u)f'(i)}{C_l + C_u + 0.5\alpha(S+i-1)}.
\end{aligned}$$

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 $\overline{q}_2(i)' \geq -\frac{1}{2S}$. □

A-7. 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 $\overline{q}$ is not available, set $\bar{q} = \frac{C_u}{C_l+C_u}$. Then set $\hat{S} = S + \lceil S(1-\overline{q})\rceil$;*

b. Compute expected waiting time $E[w]$ (equation (21)).

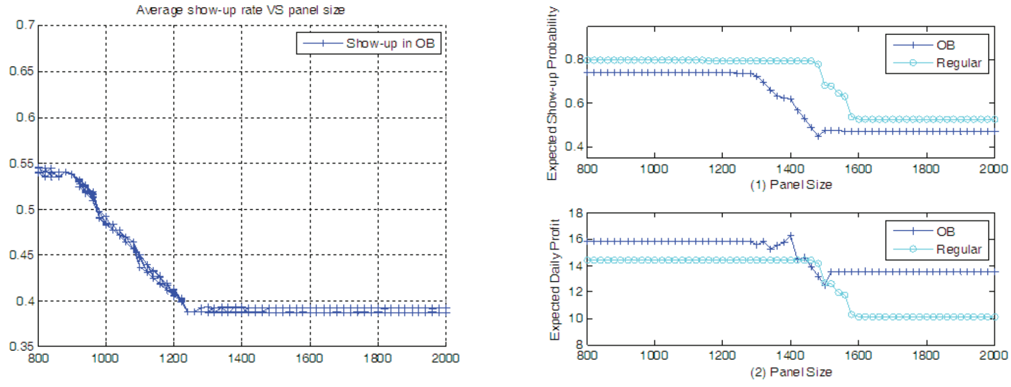
c. For $k = 0, \ldots, K$, compute the show-up probability $q(k)$ using (20), hence $p(k) = 1-q(k)$;

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

e. Calculate the expected show-up rate $\overline{q}$ 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-4(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 $\pi_1(k), k = 0, 1, .., K$ and apply Algorithm 1. Figure A-4(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.



(a) Convergence in I-SBS

(b) Performance of A 4-block Model

Figure A-2: Convergence in I-SBS and Performance of a 4-Block I-MBS

A-8. Algorithm 2 for I-MBS and Its Performance

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 $\max_{k=0,\dots,K}\{|\pi_1(k) - \pi_2(k)|\} > \epsilon$, do
    a. For each block $j = 1$ to $B$, do
        I. if $\overline{q}_j$ is not available, set $\overline{q}_j = \frac{C_u}{C_l+C_u}$ and $\overline{q}'_j = \overline{q}_j - 2\epsilon$ ($\overline{q}'_j$ is used to see whether $\overline{q}_j$ has converged). Also, set $\hat{S}_j = S_j + \lceil S_j(1-\overline{q})\rceil$.
        II. While $|\overline{q}_j - \overline{q}'_j| > \epsilon$, do
            i. Set $\overline{q}_j = \sigma\overline{q}_j + (1-\sigma)\overline{q}'_j$ (where $\sigma \in (0,1)$ is a randomly selected constant) and update $\hat{S}_j = S_j + \lceil S_j(1-\overline{q}_j)\rceil$;
            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 $k = 0, \dots, K$, compute the show-up probability $q_j(k)$ using (26);
            iv. Set $\overline{q}'_j = \overline{q}_j$ and obtain $\overline{q}_j$ 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 $\pi_2(k)$, $k = 0, 1, 2, ..., K$;
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 $j = 1, \dots, B$, 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-4(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.