The Impact of Overbooking on Primary Care Patient No-show

<table>
<thead>
<tr>
<th>Journal:</th>
<th>IIE Transactions on Healthcare Systems Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>UHSE-2012-0027.R2</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Health Care Operations Management</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>17-May-2013</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Zeng, Bo; University of South Florida, Department of Industrial &amp; Management Systems Engineering Zhao, Hui; The Pennsylvania State University, Smeal College of Business Lawley, Mark; Purdue University,</td>
</tr>
<tr>
<td>Keywords:</td>
<td>outpatient clinics, stochastic processes, patient no-shows</td>
</tr>
</tbody>
</table>

URL: http://mc.manuscriptcentral.com/UHSE  Email: john.fowler@asu.edu
The Impact of Overbooking on Primary Care Patient No-show

Overbooking has been widely adopted to deal with primary care’s prevalent patient no-show problem. However, there has been very limited research that analyzes the impact of overbooking on the major causes/factors of patient no-show and most importantly, its implications on patient no-show. In this paper, we take a novel approach and develop a game-theoretic framework (with queueing models) to explore the impact of overbooking on patient no-show through its effect on two important factors shown to affect no-show: appointment delay (time between a patient requesting an appointment and his actual appointment time) and office delay (the amount of time a patient waits in the office to see the doctor). While overbooking reduces appointment delay (which may positively affect patient no-show rate), it increases office delay (which may negatively affect patient no-show rate). Our results show that, considering both impacts of appointment delay and office delay, patient no-show rate always increases after overbooking. Further, there exists a critical range of patient panel size within which overbooking may also lead to lower expected profit for the clinic. Correspondingly, we propose two easy-to-implement strategies, which can increase clinic’s expected profit and reduce no-show at the same time.

Key words: Health care management, patient response, overbooking strategies.

1. Introduction and Literature Review

Patient no-show is one of the most challenging operational issues facing nearly all primary-care clinics (Cayirli and Veral, 2003; Gupta and Denton, 2008). Clinic no-show rates have often been significant in practice, ranging from 5% to 60% (Woodcock, 2003), reaching as high as 80% for some healthcare settings such as public pediatric (Rust et al., 1995). These statistics do not include late appointment cancelations, which have nearly the same damage as no-shows and can be dealt with similarly. Patient no-show has 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 delayed treatments and negative effects on patients’ health, in addition to lower staff productivity and reduced revenues for the healthcare providers.

Many factors contribute to patient no-show behavior, for example, long appointment delay (time between a patient requesting an appointment and his actual appointment time), patients’ dissatisfaction due to long office delay (waiting time in office) and the perception that the healthcare
system disrespects their time and beliefs, forgetfulness, time constraints, transportation, weather conditions, and patients’ emotions (Goldman et al., 1982; Bean and Talaga, 1992; Garuda et al., 1998; Lacy et al., 2004).

Customer no-show is not unique to 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 (Rothstein, 1985). As stated in Smith et al. (1992), more than $225 millions resulted from overbooking in 1990, which accounts for 40% of American Airline’s total benefit obtained through revenue management.

Following the successful stories in the airline industry, many clinics have implemented overbooking to stabilize revenue streams and improve healthcare access (Keir et al., 2002; Kim and Giachetti, 2006). The option of overbooking has also been provided in many commercial scheduling software, e.g., Encore2008 and Spectrasoft. Advanced overbooking methods have also been developed by healthcare engineering researchers, e.g., Kim and Giachetti (2006); Muthuraman and Lawley (2008); Zeng et al. (2010); Liu et al. (2010); Robinson and Chen (2010).

Despite its widespread usage to counter patient no-show, the impact of overbooking on the major causes/factors of no-show and its implications to patient no-show rate itself have not been analyzed. Notice that many of the aforementioned factors for no-show are random or uncontrollable, yet some (as described below) are affected by the administration’s overbooking strategies which may in turn affect patient no-show behavior. It is the purpose of this paper to examine these factors to better understand/evaluate the impact of overbooking and to propose, accordingly, easy-to-implement strategies that may reduce patient no-shows.

We will focus on two very important factors that have been shown to be particularly affected by overbooking strategies hence may in turn affect patient no-show, while capturing the many other random factors through a random variable. These two factors are appointment delay and office delay (patients waiting time in the office to see the doctor). Specifically, on the one hand, overbooking reduces appointment delay (because more patients can be scheduled) which may have positive impact on patient no-show rate. Many studies report the link between longer appointment delay with higher chances of no-show (e.g. Grunebaum et al. (1996); Gallucci et al. (2005); Dreither et al. (2008); Liu et al. (2010)). Bibi et al. (2007) also report that a clinic used managed overbooking (controlling each physician’s appointment volume based on his working speed) as an intervention tool to reduce appointment delay and observed reduced no-show rate. So overbooking has the potential to reduce no-show rate through its reduction of appointment delay. On the other hand, overbooking increases office delay which may have a negative impact on patient no-show rate. 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 (disutility) for this visit. For patients who are covered by third-party payments, the only costs are travel and waiting time (office delay) involved in keeping the appointment. As office delay increases due to overbooking, costs to the patient correspondingly increase. For some patients, this increase in disutility may be enough to tip the balance towards not attending (Sharp and Hamilton, 2001; Bibi et al., 2007). This can be especially true when the utility of attending is low (e.g., a follow-up check for a seemingly well-recovering patient) or has become low (symptoms significantly improved during the appointment delay). One may think of two alternative ways for patients to counter long office delay: arriving late or switching providers. Both ways, however, have not been viable, as shown from empirical studies, because of clinics’ widely used re-scheduling policies upon late arrivals (Cayirli et al. (2006)) and the many constraints (e.g., insurance network, provider specialty/availability) involved in switching providers (only 5.4% of patients voluntarily switch providers Rice et al. (1992)). Indeed, office delay is ranked patients’ top complaint (LaGanga and Lawrence, 2008) and many reported links between patients’ bad waiting time experiences/dissatisfaction with future no-show behavior (Lowes (2005); Dyer (2005); Lacy et al. (2004); Garuda et al. (1998); van Baar et al. (2006)). Capturing this link is important since as Sharp and Hamilton (2001) point out, overbooking may be counterproductive due to the increased no-show caused by the longer waiting time. However, there has been no analytic research looking at this link. To summarize, the impact of overbooking on patient no-show rate is two-fold: on the one hand, overbooking may reduce no-show due to the reduced appointment delay; on the other hand, overbooking may increase no-show due to the increased office delay. It is the goal of this paper to analyze the overall impact of overbooking through capturing these two conflicting effects.

Gupta and Denton (2008) define office delay and appointment delay as the direct and indirect waiting, respectively, and provide a comprehensive review of literature on appointment scheduling. As described in this review, the majority of the operations literature is on direct waiting (office delay), modeling the trade-off between patients’ waiting time on the day of their appointment and physician utilization. This includes the stream of work on patient no-show, e.g., Robinson and Chen (2003), Hassin and Mendel (2008), and those using overbooking to deal with no-shows, e.g., Kim and Giachetti (2006); Laganga and Lawrence (2007); Zeng et al. (2010); Muthuraman and Lawley (2008); Turkcan et al. (2011). As stated in Gupta and Denton (2008), “although evidence suggests that longer direct waiting times are a key reason why patients tend to miss appointments, none of the overbooking models have considered a linkage between suggested overbooking targets.
and the possible increase in no-show rates that could be caused by greater direct waiting times.”

Our paper exactly captures this link when modeling the impact of the increased office delay on patient no-show rate.

There has been much less work on indirect waiting (appointment delay), most of which develops effective dynamic scheduling policies considering the record of scheduled appointments, e.g., Patrick et al. (2008); Gupta and Wang (2008); Liu et al. (2010). The more closely related ones are Green and Savin (2008) and Liu and Ziya (2011). Green and Savin (2008) develop a queueing 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, it does not involve overbooking. Liu and Ziya (2011) look at tackling patient no-show through controlling panel sizes and using overbooking. While they consider reduced appointment delay due to overbooking using a queueing model, they do not model the link between overbooking and patient no-show rate. In contrast, we particularly model the patient response (in terms of no-show rate) to the impact of overbooking on both appointment delay and office delay, hence adopting a game-theoretic framework with queueing models. The focus of our results is to show the impact of overbooking on patient no-show rates and proposing more effective overbooking strategies that will increase clinics’ revenues and no-show rates at the same time. There has been even less work that considers both appointment delay and the office delay. The only other work that we are aware of is Kortbeek et al. (2011) which designs appointment schedule for outpatient clinics with scheduled and walk-in patients. This paper does not consider the impact of overbooking on these delays.

To summarize, our work has two important contributions: (1) There has been very limited research that considers both appointment delay and office delay, partly because of the potential complexity involved (Gupta and Denton, 2008; Liu and Ziya, 2011). We provide a model framework for doing so. (2) We develop a novel model, a game-theoretic framework, to consider not only the impact of overbooking on appointment delay and office delay, but further, how this impact affects patient no-show. No previous research has looked at this.

In presenting our model, since appointment delay and office delay correspond to two different 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 appointment delay, referred to as the integrated model. By including a patient office delay tolerance factor, we can control how much impact office delay has on patient no-show decision. 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 reflects only the impact of appointment delay (and other random factors) on patient no-show behavior. In the numerical study, we particularly study how this tolerance factor affects clinic performance.

Our results show that, surprisingly, no other measures taken, the reduction of appointment delay (due to overbooking) does not necessarily lead to improvement in patient show-up rate. This is because the additional patients scheduled at the end of the queue due to the extra scheduling capacity resulted from overbooking still experience long appointment delay. Therefore, considering both impacts of appointment delay and office delay, patient no-show rates always increase after overbooking. Further, although overbooking increases clinic's expected profit most of the time, there exists a critical range of the patient panel size within which overbooking may reduce the clinic's expected profit. Correspondingly, we propose two easy-to-implement strategies, overbooking with controlled appointment queue and selective dynamic overbooking, both of which can increase clinic's expected profit and improve no-show rates at the same time. We also provide some guidelines of how to choose the controlled appointment queue length.

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 other random factors) on no-show and derive results for both single block scheduling (SBS) and multiple block scheduling (MBS). In Section 3, we incorporate appointment delay to develop the integrated model for SBS and MBS. In section 4, we conduct a comprehensive numerical study to investigate the overall impact of overbooking under different parameters and obtain insights which lead to useful suggestions to the practitioners. In Section 5, we conclude with a summary of results, discussion of managerial insights, and some future research directions. Due to space constraint, we present all the analysis and results regarding MBS in the appendix.

2. The Basic Model

The basic model studies the interactions between patient no-show rate and office delay by overbooking (and other random factors). To capture the patient response to overbooking (in terms of no-show rate) through its impact on office delay, the basic model takes a game-theoretic framework. 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 SBS. 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 (we assume there is an \( \epsilon \) difference in their arrival time, which determines their sequence in the queue 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) for this visit from the factors such as the expected waiting time, weather conditions and so forth. Unless explicitly mentioned, we use capital letters to denote random variables and their lower cases to denote particular realization of the variables.

To model the fact that each clinic visit has a different utility for the patients 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 \( C \), defined to be a patient’s utility less the random disutility factors, and referred to in short, as the patient’s utility. While the patient himself knows his realization of \( C \), i.e., \( c \), we assume that \( C \) over the patient population follows a uniform distribution over \([-c_l, c_u]\) with \( c_l, c_u \geq 0 \), which is common knowledge to everyone, i.e., the probability density function is

\[
g(C = c) = \begin{cases} \frac{1}{c_l + c_u} & \text{if } c \in [-c_l, c_u] \\ 0 & \text{otherwise.} \end{cases}
\]

We allow \( C \) to take negative values to capture the fact that patients could fail to show up due to 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 \( q_0 \equiv 1 - p_0 \). Note that the maximum patient show-up rate is \( q_0 \).

Let \( w \) denote a patient’s waiting time in a clinic visit. Clinic experiences indicate that a patient’s disutility of waiting time, \( L(w) \), is a non-decreasing function in \( w \). For tractability, we assume patients are homogeneous in their disutility function (although heterogeneous disutility may be added with much more complexity), which can be approximated by a 2-piece linear function as follows:

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

where \( \alpha \) indicates patients’ sensitivity to waiting time, with a higher \( \alpha \) indicating a higher sensitivity.
(having a higher disutility) to office waiting time. To accommodate the findings from the literature (e.g., Hart (1995); Yeboah and Thomas (2010)) that patients are usually not sensitive to waiting time until it is greater than a threshold, we assume there exists a threshold of \( w_0 \) such that a patient incurs disutility from waiting only if his waiting time exceeds \( w_0 \). In the 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 particular patient’s clinic visit is his utility from the visit less the disutility from waiting, i.e., \( c - L(w) \). Recall that disutility from other random factors is already included in \( c \), as defined earlier. Obviously, at the time of a patient’s decision of whether to show up or not, he does not know (and no one knows) his waiting time, which depends on where in the office waiting queue he would be if he showed up. Therefore, the patient would make his decision based on his expected waiting time. Let \( E[W] \) denote the expected waiting time for his next clinic visit. We assume that a patient makes his decision in order to maximize his net utility. It is easy to see that a patient will show up if the net utility of his clinic visit is non-negative, i.e., \( c - L(E[W]) \geq 0 \).

Before the visit, a patient does not know how many other patients will be physically in front of him. 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, \( E[W] \), and patients’ utility drawn from the distribution, \( g(C) \), patient no-show rate in this block can be computed as

\[
p = \int_{-\infty}^{L(E[W])} g(C) dC = \frac{c_l + L(E[W])}{c_l + c_u} = p_0 + \frac{L(E[W])}{c_l + c_u}
\]

and the patient show-up rate will be

\[
q = 1 - p = 1 - \frac{c_l}{c_l + c_u} - \frac{L(E[W])}{c_l + c_u}
\]

We assume that \( c_u \) is greater than the largest possible \( L(E[W]) \), i.e., there always exist some patients whose utility of the visit warrant their visits to the clinic even with the longest waiting time. This guarantees \( p \) bounded up by 1. 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 \( p \) and \( q \) in the calculations, whichever is more convenient.

Now, we are ready to describe the interactions between office delay caused by overbooking and patient no-show. Suppose, initially \( S \) patients are scheduled in a block, the expected waiting time is
If the clinic overbooks to $S' \geq S + 1$ patients for this 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 $S'$ 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 $p(E[W']) \geq p(E[W])$. Such interactions will settle at a fixed point, i.e., a Nash Equilibrium (NE). In the next section, we use a game theoretic model to capture the above interactions between the clinic and the patient population and solve for the NE.

Since choosing which overbooking policy to use involves many factors, to focus on what is proposed for this research - capturing the impact of overbooking on no-show rate, we will build the model assuming the clinic adopts a specific overbooking policy - the *naive statistical overbooking* (NSOB). We choose NSOB for a few reasons: (1) it’s simple, hence provides analytical tractability for the derivation of NE; (2) it has been used in practice (Kim and Giachetti (2006)); and (3) it is a very mild policy (not too much overbooking), hence our results will be an underestimate of the impact of office delay compared to many other methods proposed in literature (heavier overbooking generally leads to more office delay). The model framework we establish can be applied to study other overbooking policies proposed in literature.

In the next section, we establish the specific relationship between overbooking and office delay it causes and characterize the NE of the patient no-show rate for the basic model.

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

In this section, we analyze the game between the clinic and the patient population under SBS. Specifically, given the patient no-show rate, the clinic’s decides how much to overbook, which is specified from the NSOB policy. On the other hand, given the overbooking level, patients determine whether they would show up, from which we obtain the patient population’s show-up rate as a response to the overbooking level. Next, we present the response functions of both sides.

Suppose the capacity of the block is $S$ patients (with no overbooking), where $S$ can only be integers. Given the current no-show rate, $p$, a clinic (she) determines the overbooking level, $i(p)$.

We assume, throughout the paper, there is plenty of demand to fill up all scheduling capacity.

Under NSOB, the clinic’s decision, $i$, can be written as a function of $p$ (or $q$):

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

(3)
where “⌈⌉” is the ceiling function to ensure the integer requirement. Hence, after overbooking, the clinic books \( S + i(p) \) patients in the block. In other words, given a no-show rate, the clinic’s response function is given in (3).

Then, given the amount of overbooking, \( i \), each patient makes his show-up-or-not decision (a binary decision) according to his utility and the expected waiting time, maximizing his achieved net utility. To put in math form, let \( z = 1 \) be a patient’s decision to show-up and \( z = 0 \) not to show up. To solve for the patient’s best decision, he faces the following optimization problem

\[
\begin{align*}
\max_{z \in \{0, 1\}} & \quad z \cdot (c - L(E[W])) \\
\text{subject to} & \quad z \cdot (E[W] - w_0) \\
\end{align*}
\]

where \( a^+ = \max\{0, a\} \) and the \((\cdot)^+\) term captures whether the expected waiting time is greater than the tolerance, \( w_0 \).

To obtain \( E[W] \), given the overbooking amount, \( i \), there are \( S + i - 1 \) patients (excluding the patient himself) scheduled to come. So, the expected waiting time for a patient can be computed from his expected position in the queue, i.e.,

\[
E[W|i] = \frac{1}{2} (S + i - 1)q(i).
\]

Plugging this into (4), we have

\[
q(i) = \frac{c_u - \alpha (q(i)(S + i - 1) - w_0)^+}{c_l + c_u}.
\]

Solving for \( q(i) \), we obtain

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

where

\[
\hat{q}(i) = \frac{c_u + \alpha w_0}{c_l + c_u + \frac{\alpha}{2}(S + i - 1)}
\]

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., \( w_0 = +\infty \)), \( q(i) = q_0 \). This indicates that patient show-up rate is overestimated when not considering the negative impact of office delay.
So far, we have captured the best response functions for the game between the clinic and the patient population, i.e. \( i(q) \) given in (3), and \( q(i) \) given in (5).

Solving these two responses to obtain the NE is non-trivial due to the ill behavior of \( q(i) \) and \( i(q) \) and the integer requirements for \( i(q) \). Details of the solution process is provided in Appendix A-1 including two very mild assumptions. Since \( i(q) \) is neither concave nor convex, we may have multiple NE.

Define

\[
\hat{q}^c = 1 + \frac{c_l + c_u}{\alpha S} - \frac{1}{2S} - \sqrt{\left(1 + \frac{c_l + c_u}{\alpha S} - \frac{1}{2S}\right)^2 - \frac{2(c_u + \alpha w_0)}{\alpha S}}.
\]  

(7)

Let \( i^* = i(\hat{q}^c) \) and \( q^* = \hat{q}(i^*) \), where \( i(\cdot) \) and \( \hat{q}(\cdot) \) are defined in (3) and (6), respectively. \( \hat{q}^c \) and all the NE (presented in the following theorem) can be demonstrated in Figure 1(a) and Figure 1(b) (all figures are shown at the end of the paper). 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 \(-\frac{1}{S}\).

**Theorem 1.** For the basic model with SBS, given \((q^*, i^*)\) defined above, 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.

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 this section, we have analyzed the basic model under single block schedule (SBS). Analysis of the basic model under multiple blocks (MBS) is presented Appendix A-4.1. In the next section, we discuss the integrated model which incorporates appointment delay into the basic model.
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 reduced appointment delay. To do this, we need to model the queue for obtaining an appointment. As we mentioned, there has been little operations literature on appointment delay. Assuming Poisson arrivals of the patient appointment requests and exponential service times, Green and Savin (2008) (referred to as GS in the rest of this paper) provide a nice $M/M/1/K$ queueing model to approximately capture appointment delay, but without considering overbooking. We use an approximate way to adapt GS model by adding the component of overbooking and then incorporate it into 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. It is assumed that when patients request an appointment, they like to have it as soon as they can. In the GS model, patient appointment requests come as Poisson arrivals waiting to be served. Suppose a single patient’s appointment request rate is $\lambda$ and the panel size of the clinic (the population served by the clinic) is $N$, then $\lambda \times N$ represents the arrival rate of the patients’ appointment requests. Patients are served in exponential service time with a rate of $T$ where $T^{-1}$ is defined as the capacity of a single day (in our model, $T^{-1} = B \cdot S$, where $B$ is the number of blocks in a day and $S$ is the service capacity in each block).

Defining the state as the queue length a patient sees when he requests an appointment (appointment delay $k$) and assuming a patient who sees an appointment delay ($k$) exceeding a certain limit ($K$) will leave for service elsewhere (for this visit), the authors model the appointment system as an $M/M/1/K$ queue in which patients entering the system have a state-dependent probability of not being served (no-show) and may rejoin the queue with a re-booking probability, $r$. Assuming the no-show rate as a function of $k$ obtained from empirical data, the steady-state distribution of the queue length $k$ can then be determined. Specifically, the empirical state-dependent no-show rate as a function of $k$ (equation (1) in GS), using our notation, can be rewritten as

$$p(k) = p_{max} - (p_{max} - p_{min})e^{-k/D}, \quad (8)$$

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 $D$ is the no-show
appliance delay sensitivity parameter.

Adapted into our notation, 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 ($p_k$) can be calculated using the following equations (which correspond to (17)-(18) in GS paper) where $p(k)$ is given in (8).

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

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

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

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

We then can use the steady state distribution $\pi(k)$, obtained from solving (9)-(12), to compute the expected no-show rate $\sum_{k=0}^{K} p(k)\pi(k)$, as discussed in next section.

Since the appointment request rate to the clinic, $\lambda N$, has an impact on the above calculations regarding appointment delay, by managing its panel size, the clinic can adjust appointment delay. As in the GS model, we also evaluate different panel sizes and their impacts when considering overbooking in our numerical study.

The above (GS) model does not take into consideration of overbooking, which clearly changes appointment delay and therefore no-show rates. In the following sections, we add overbooking into the above model and integrate the impact of appointment delay with that of office delay. As can be expected from the above description of the model, when considering overbooking and office delay, the state-dependent no-show rate (as an input into the queueing model), $p_k$, will change to the no-show rate considering both appointment delay and office delay (and other random factors). In addition, the service capacity will change from $S$ to $\hat{S}$ (or $B \cdot S$ to $\sum_{j=1}^{B} \hat{S}_j$ for multi-blocks) since although service rate has not changed, the clinic can schedule more patients because of overbooking. Details about integrating office delay and appointment delay model with overbooking is described in the following sections.

### 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 (random factors). These three sources do not overlap with each other by definition (i.e., office delay captures the impact of long
office waiting on the appointment day while appointment delay captures the deterioration of show-up rate due to long period time before the appointment day). In other words, we can decompose them in the calculation and write \( p = p^a + p^o + p_0 \), 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 \( p_0 = \frac{c_t}{c_t + c_u} \) is the unassignable no-show introduced in Section 2.1.

Recall that \( p_{\min} \), introduced in (8), is the minimum observed no-show rate, i.e., the no-show rate when there is no appointment delay (\( k = 0 \)). Hence, \( p_{\min} \) is not related to appointment delay and can be simplified as \( p_{\min} = p^o + p_0 \). Therefore, (8) can be equivalently represented as

\[
p(k) = p_{\max} - (p_{\max} - p_{\min})e^{-k/D} = (1 - e^{-k/D})(p_{\max} - p_{\min}) + p_{\min} = (1 - e^{-k/D})(p_{\max} - p_{\min}) + p^o + p_0, \tag{13}
\]

which allows us to separate the impacts of appointment delay and office delay. 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 include office delay, and also overbooking, as detailed below.

To consider overbooking, from (13) we write the patient no-show rate as

\[
p(k|i) = (1 - e^{-k\hat{S}/D})(p_{\max} - p_{\min}) + p^o(i) + p_0. \tag{14}
\]

Comparing (14) and (8), we see how (14) captures the benefits of overbooking in reducing appointment delay. A patient who originally sees an appointment delay of \( k \) will see an appointment delay reduced to \( k\hat{S}/S \) after overbooking. As a result, combining office delay captured in Section 2.2 and considering impact of overbooking on both delays, patient show-up rate for an appointment delay \( k \), \( q(k) \), given the amount of overbooking \( i \), can be written as

\[
q(k|i) = 1 - p(k|i) = 1 - \frac{c_t}{c_t + c_u} - \frac{\alpha(E[W|i] - w_0)^+}{c_t + c_u} - (1 - e^{-k\hat{S}/D})(p_{\max} - p_{\min}) \tag{15}
\]

where

\[
E[W|i] = \frac{1}{2}(S + i - 1)\bar{q}(i). \tag{16}
\]

and

\[
\bar{q}(i) = \sum_{k=0}^{K} q(k|i)\pi(k|i), \tag{17}
\]

with \( \pi(k|i) \) solved from equations (9)-(12) (in which \( B \cdot S \) is replaced by \( \hat{S} \) hence \( \pi(k) \) depends on \( i \)).
In (15), the last three terms are the no-show caused by random factors \( p_0 \), office delay, and appointment delay, respectively. Note that the first three terms are the same as in the basic model (equation (4)). \( \tilde{q}(i) \) in (17) is the expected patient show-up rate considering all possible appointment delays, as the response to the amount of overbooking, \( i \). Since an arrival seeing \( K \) patients ahead of him will not join the queue, the exact expected patient show-up rate should be derived by taking expected value over the show-up rate of \( K - 1 \) patients. However, equation (17) provides a close and trackable approximation. Numerical study shows a 1-2% difference in expected show-up rate on average.

The other response function is the clinic’s overbooking level, \( i(\tilde{q}) \) (or equivalently, the scheduling level, \( \hat{S}(\tilde{q}) \equiv S + i(\tilde{q}) \)), given the expected patient show-up rate \( \tilde{q} \). This follows the NSOB policy, i.e.,

\[
\hat{S}(\tilde{q}) = S + \lceil S(1 - \tilde{q}) \rceil.
\]

Therefore, by solving equations (15)-(18) (and (9)-(12) to obtain \( \pi(\cdot) \)), we can obtain the equilibrium solutions to the integrated model with single block scheduling. In summary, the major difference between the integrated model and the basic model is the inclusion of the impact of appointment delay in the patient show-up equation (15) and the expected patient show-up rate over all appointment delay (equation (17)).

Before further analysis, we would like to discuss two different operating policies depending on whether the clinic fills the extra scheduling capacity resulted from overbooking (since the service rate is assumed not to change, when scheduling more patients and patients show up, overtime is used to serve patients. So, the scheduling capacity is increased because more patients are scheduled/served in a day when 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 this appointment. 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 \( 20S/\hat{S} \) days after overbooking. Therefore, some of the patients originally seek medical service elsewhere may stay for appointments if 20-day is still the threshold of seeking service elsewhere, i.e., \( K \) 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 scheduling capacity resulted from overbooking is filled. Alternatively, clinics may take advantage of $K$ and control it such that not all extra scheduling capacity is used up for more patients. A possible way is to keep $K$ the same as that before overbooking by not allowing appointments to be made beyond certain days ($20S/\hat{S}$ days), referred to as the constant-$K$ overbooking scenario. Although adjusted-$K$ 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. This is also a good example to show that we can turn a parameter that reflects patients’ behavior ($K$ in this case) into an administrative tool to help improve clinic’s performance. More details about implementation is also discussed in Section 4.

The complicated inter-relationship among equations (15)-(18) 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, 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. We first investigate the situation when only appointment delay (i.e., patients are insensitive to office delay) is considered. We analyze the properties of $\overline{q}(i)$ and characterize the NE solutions. Then, we extend our study to the situation where both delays are considered, respectively.

3.2.1 Constant-$K$ Overbooking with Appointment Delay Only

This case corresponds to the situation where $e_0$ is very large so that patient responses towards office delay can be ignored. In this case, as shown in the following, 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 (the intersection of $\overline{q}(i)$ and $i_c(q)$) for the game when not considering the integer requirements. 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, we establish bounds on its first derivative although concavity/convexity of $\overline{q}(i)$ cannot be obtained.

Lemma 2. $0 < \overline{q}(i)' < +\infty$.

Based on Lemma 2, we characterize the NE when considering the integer requirement.

Proposition 3. Considering only appointment delay, Let $\overline{q}_e^*$ be the equilibrium show-up rate without the integer requirements and $i^* = i(\overline{q}_e^*)$ be the corresponding integer overbooking level. For the
constant-K overbooking, there is either no NE or a unique NE at \((\tilde{q}(i^*), i^*)\), where \(\tilde{q}(i^*)\), calculated using (15)-(17), is the equilibrium patient show-up rate given the amount of overbooking, \(i^*\).

Although Proposition 3 states that it is possible to have no NE, this is of a small probability since \(S\) is sufficiently large in single block scheduling. In this case, the model can be treated as a continuous game (game without considering the integer requirements), which always has a solution.

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, \(q(i)\) becomes the minimum of two differentiable functions, i.e.,

\[
\tilde{q}(i) = \frac{c_u}{c_l + c_u} - (p_{max} - p_{min}) + (p_{max} - p_{min}) \sum_{k=0}^{K} e^{-\frac{kS}{D(S+\alpha)} \pi(k)} - \frac{\alpha(E[W|i] - w_0)}{c_l + c_u},
\]

\[
= \min\{\tilde{q}_1(i), \tilde{q}_2(i)\},
\]

(19)

where \(\tilde{q}_1(i) = \frac{c_u}{c_l + c_u} - (p_{max} - p_{min}) + (p_{max} - p_{min}) \sum_{k=0}^{K} e^{-\frac{kS}{D(S+\alpha)} \pi(k)}\), and

\[
\tilde{q}_2(i) = \frac{c_u}{c_l + c_u} - (p_{max} - p_{min}) + (p_{max} - p_{min}) \sum_{k=0}^{K} e^{-\frac{kS}{D(S+\alpha)} \pi(k)} - \frac{\alpha(E[W|i] - w_0)}{c_l + c_u}.
\]

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

Proposition 4. \(-\frac{1}{2S} < \tilde{q}_2(i)' < +\infty\) for \(i \in [0, S]\).

Based on Proposition 4, we can easily see that there exists at most one intersection of \(\tilde{q}_2(i) < 1\) and \(i_c(q)\). We use \(\tilde{q}_{2c}\) to denote the patient show-up rate at the intersections of \(\tilde{q}_2(i)\) and \(i_c(q)\) and set \(\tilde{q}_{2c}^* = \infty\) if there is no such intersection. Also, we use \(\tilde{q}_{1c}^*\) to denote the intersection of \(\tilde{q}_1(i)\) and \(i_c(q)\), whose existence is guaranteed. Then, we have the following result.

Define \(\tilde{q}_c^* = \min\{\tilde{q}_{1c}^*, \tilde{q}_{2c}^*\}\) and \(i^* = i(\tilde{q}_c^*)\), based on Propositions 3 and 4, we can characterize the NE when considering both delays and also integer requirement.

Proposition 5. For the game with integer restriction considering both delays, there exist at most two NE. If there exists a unique NE, it is \((\tilde{q}(i^*), i^*)\). If two NE exist, they are \((\tilde{q}(i^*), i^*)\) and \((\tilde{q}(i^* + 1), i^* + 1)\).

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 in Appendix A-5) can be
developed which offers quick solutions to equations (15)-(18), providing a numerical way to obtain
the equilibrium.

In this section, we have analyzed the integrated model for single block scheduling. Analysis of
the integrated model for multiple blocks is presented in Appendix A-4.2. The analysis and results of
the single block model provides understanding and insights into the more complicated and realistic
multiple block model.

4. Numerical Study

In this section, we present results of a numerical study which investigates the impact of overbooking
under different parameters, considering its opposing effects on appointment delay and office delay.
Based on these results, we obtain insights that lead to two 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 clinic’s strategies.

Results are reported for the single block model because they are computationally more tractable
and their figures are much cleaner to demonstrate the insights. We do, however, perform a set of
numerical study on overbooking in the multi-block scheduling model. Results are qualitatively
similar to those in the single block model. Numerical results also show that the calculation of
single block model provides a good approximation to the multiple block model. Detailed discussion
and additional insights on the multi-block model are presented in Appendix A-4.3.

Whenever appropriate, we adopt parameters obtained from real data or from existing literature.
Specifically, we use parameters based on data collected from a public mental health clinic at the
Johns Hopkins Bayview Medical Center in Baltimore that was reported in Gallucci et al. (2005).
This data set was also adopted for GS model. Based on this data set, the no-show appointment delay
sensitivity parameter D is set to 9 days, the maximal no-show percentage due to appointment delay
p_{max} is set to 36%. The average appointment request arrival rate from a single patient is \lambda = 0.008
(i.e., 0.008 per day), service rate is S = 20 per day, and rebooking rate is r = 1. In addition, taking
20 days as the threshold of appointment delay above which patients will seek service elsewhere for
this visit, the upper bound on the queue length (in regular booking) is K = 20 \cdot 20 = 400. We set
waiting time tolerance for single block scheduling to be w_0 = \{5, 8, 11\}, the disutility rate \alpha = 3,
unassignable no-show rate \( p_0 = \{\frac{1}{4}, \frac{1}{7}\} \), and \( c_l = 15 \) for the uniform distribution parameters for patient utility \( (c_u) \) can be obtained from \( p_0 \) since \( p_0 = \frac{c_l}{c_l + c_u} \). We normalize the revenue for one patient visit to 1 and set the overtime cost to 2 per time unit. In following, we discuss numerical results and the corresponding insights.

### 4.1 The Impact of Overbooking

Figure 2 compares 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), there exists a critical range of the panel size, 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 clinic’s expected profit decrease in a less drastic fashion in the range. This indicates that overbooking mitigates the impact of panel size on clinic’s 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 \textit{always} increases patient no-show rate. Hence, no other measures taken, overbooking does not solve or improve the root problem of patient no-show, in fact, it makes the no-show rate worse. It only improves clinic’s profit (although we show above that profit could also suffer for panel sizes in the critical range). Therefore, instead of imposing higher and higher degrees of overbooking, clinics should think of approaches to more effectively conduct overbooking if they want to improve no-show.

The above two observations, especially the second one, seem somewhat puzzling since although overbooking increases office delay which leads to lower show-up rate, overbooking reduces appointment delay which could increase show-up rate as we discussed. So, why show-up rate is always lower under overbooking? A closer look in the next subsection provides reasons to the above observations as well as suggestions that can improve the performance after overbooking.

### 4.2 Clinic Intervention: Controlled Appointment Queue and Overbooking Level

As we mentioned, with adjusted-\( K \), clinics fill all the extra scheduling 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 3 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 office delay is always higher with overbooking. As a result, with adjusted-$K$ overbooking, we can expect higher no-show rates because of the not-improved appointment delay and the increased office delay. On the other hand, if we control $K$ such that not all extra scheduling capacity resulted from overbooking is used up for scheduling more patients, it may take advantage of the positive effect of overbooking on reducing 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. We will discuss the value of $K$ in next subsection.

Figure 4 shows the same cases as in Figures 2, adding constant-$K$ overbooking. While the observations of the critical ranges 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 negative impact of overbooking on office delay is so small that it is dominated by the positive impact of overbooking on appointment delay regardless of the panel size, leading to higher show-up rates with overbooking under all panel sizes. In contrast, when patients are very sensitive to 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 the panel size, 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 size, because the positive effect of overbooking on reducing appointment delay is higher with a larger panel size. Therefore, when panel size is large, the positive effect of overbooking on appointment delay exceeds its negative effect on office delay, but when panel size is small, it cannot counter its negative effect on office delay. As a result, we see overbooking reduces patient show-up rate for small panel sizes but increases the patient show-up rate for large panel sizes.
As for the clinic’s expected profit, unlike the adjusted-$K$ case, constant-$K$ overbooking always improves clinic’s profitability, for all panel sizes. Further, overbooking increases clinic’s expected profit more for larger panel sizes because, as we mentioned, the positive impact of overbooking on appointment delay is higher with large panel sizes. Finally, compared to the adjusted-$K$ case, the critical range for the constant-$K$ case arrives later (curves shifted to the right), indicating that constant-$K$ overbooking can handle bigger panel sizes.

The above observations indicate that when adopting overbooking, instead of filling all the extra scheduling capacity resulted from overbooking, clinics which impose some control over their appointment queue will see higher expected profits and lower 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 (see e.g., Kopach et al. (2007); Qu et al. (2007); Robinson and Chen (2010) for more on open access). However, instead of only allowing appointments for 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-$K$ allows appointments scheduled up to a certain day from the time an appointment is made. Constant-$K$ overbooking is an improvement of the traditional scheduling method through overbooking and controlled queue length because it remains control over demand fluctuation and enjoys higher expected profits and potentially improved no-show rate.

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), considering again both appointment delay and office delay. To reflect the degree of overbooking, we revise (3) to become
\[ i(p) = \lceil aSp \rceil, \] (20)
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 5 compares expected patient show-up rate and clinic’s expected daily profit when $a = 0.0, 0.5,$ 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.

### 4.3 Re-booking Rate and Discussion of Controlled Appointment Queue Length

In this and the following subsections, we will explore two other important characteristics for clinics, both related to patients. The first one is the re-booking rate, $r$, which will be shown to be an important factor for the performance of different overbooking policies. Re-booking rate is 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 6 and 7 explore the impact of $r$. Specifically, Figure 6 shows the impact of $r$ 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 range, 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 $r$ decreases from 1.0 to 0.5, the panel size may increase by 600 – 800).

As we proposed in the previous subsection, instead of filling up all capacity made available from overbooking, overbooking with controlled appointment queue is one simple way to achieve both a higher expected profit and a lower no-show rate. We also mentioned that although constant-$K$ is desired, any $K$ value between the constant-$K$ and the adjusted-$K$ would be more beneficial than adjusted-$K$. In the following, we briefly analyze the choosing of $K$ and its interactions with two important parameters we discussed, rebooking rate and overbooking level.

We first compare the performance of regular booking and overbooking with different controlled queue length. Specifically, let $CTK$ and $ADK$ represent the $K$ values for constant-$K$ and adjusted-$K$, respectively, and $\beta ADK$ represent a level-$\beta$ value between $CTK$ and $ADK$, where $\beta ADK = CTK + \beta \times (ADK − CTK)$. Figure 7 compares the performance of regular booking with overbooking with three levels of controlled queue length ($CTK$, $0.5 ADK$, and $ADK$) under different re-booking
rates. We have a couple of interesting observations. First, the reduction in patient show-up rates due to overbooking increases as \( r \) increases. This is because an increase in \( r \) means more no-shows will re-schedule their missed appointments, taking additional scheduling capacity but may lead to no-show again. Second, the three \( K \) curves (\( CTK \), \( 0.5ADK \), and \( ADK \)) differ significantly over the critical range when rebooking rate is high, and the difference reduces as the rebooking rate reduces. In other words, the advantages of constant-\( K \) overbooking over larger \( K \) (i.e., increasing patient show-up rate and clinic’s expected profit and delaying the arrival of critical range to allow a bigger panel size) are larger for clinics with a larger re-booking rate. An intuitive explanation is that, if a shorter queue length is adopted, a no-show patient’s request to re-join the appointment queue through rebooking will be less likely to be accepted. So, its detrimental impact will be restricted. When rebooking rate becomes smaller, such advantage will become less significant and there will be less difference among \( K \)’s. Therefore, using a controlled queue length is particularly important when re-booking is higher. Since higher rebooking rate worsens the performance (both the show-up rate and the expected profit), when rebooking rate is high, to achieve the same performance, a lower \( K \) is needed.

Figure 8 furthers our analysis on controlled queue length by comparing more levels of controlled queue length under two different overbooking levels, i.e., \( a \) is set to 1 and 0.5, respectively (see (20) for definition of \( a \) ). Five levels of controlled queue length are obtained by setting \( \beta = 0, 0.25, 0.5, 0.75, 1.0 \), where \( \beta = 0 \) and \( \beta = 1.0 \) correspond to the two boundary cases, \( CTK \) and \( ADK \), respectively. As we can see, the five different \( K \) levels are very similar in shapes, with constant-\( K \) being the best and adjusted-\( K \) the worst (in terms of both the show-up rate and the expected profit). The other three levels spread almost evenly between these two levels with \( \beta = 0.25 \) closer to constant-\( K \) and \( \beta = 0.75 \) closer to adjusted-\( K \). This is true with other parameter sets as well. This indicates that an analysis of the two boundary cases tells much about the performances of other levels of \( K \) values which approximately spread over the range between the constant-\( K \) and the adjusted-\( K \). Further, comparison between the two overbooking levels indicate that as overbooking rate decreases (less overbooking), curves of all \( K \) values move closer to the regular booking curve, with a sharper slope in the critical range (hence bigger sensitivity to panel sizes in the critical range).

To summarize, in choosing the appropriate queue length, constant-\( K \) always delivers the best performance in terms of both the expected profit and the show-up rate, while adjusted-\( K \) delivers the worst. Roughly speaking, the performance of any \( K \) values in between of these two values evenly spread out between those of the two boundary cases and the closer it is to \( CTK \), the better
the performance is. The clinic can choose the value according to their needs. Further, overbooking with controlled queue length is particularly important for clinics facing moderate or high rebooking rate. For higher rebooking rate, to achieve the same performance (expected show-up rate and/or expected profit), a lower $K$ value is needed. In addition, for higher overbooking level, to achieve the same expected show-up rate, a lower $K$ value is needed.

### 4.4 Unassignable No-shows and Selective Dynamic Overbooking

In this subsection, we consider another parameter related to patients, the unassignable no-show rate $p_0$, a parameter indicating patients’ no-show tendency. Figures 9(a) and 9(b) show its impact on the expected patient show-up rate and clinic’s expected profit with adjusted $K$. Both figures have the same parameters except the values of $p_0$. Two important points stand out: (1) With higher $p_0$, the critical panel size range shifts left. This is because, with a certain $r$, 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 no-show tendency ($p_0$): 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.

The second observation indicates 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 office delay 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.

Discussions with different practitioners confirm that clinics do observe “habitual” no-show patients and that selective dynamic overbooking is insightful. A similar approach was 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 a 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.

In addition to the above, we also performed numerical analysis on multi-block scheduling, looking at the impact of the multiple blocks. Results show that the expected show-up rate decreases from the first to the last block, due to the increasing office delay caused by overflow. Refer to Appendix A-4.3 for details of the study and additional insights.

5. Conclusions

Although overbooking has been proposed and implemented in practice to reduce the impact of patient no-show on the clinic’s side (i.e., stabilizing revenue), little work has been done to study the impact of overbooking on patients’ side, which is directly related to the factors causing patients to no show and hence affects patients’ no-show decisions. In this paper, we analyze the impact of overbooking from the patients’ perspective, with the goal of finding important factors which clinic’s administration may potentially influence to improve the patients’ experience and in turn their no-show behavior.

While overbooking has little impact on many random factors that affect patient no-show, it does impact two very important factors - appointment delay and office delay, which directly affect the patients’ experience. We develop a game-theoretic framework with queueing models to analyze the impact of overbooking on patient no-show through these two important factors. Specifically, while overbooking seems to positively affect patient show-up rate due to the reduced appointment delay, it may also negatively affect patients’ show-up rate due to the increased office delay. The overall impact of overbooking depends on the relative magnitude of these two effects. Being the first work to consider both appointment delay and office delay, our analysis has provided a few interesting and important insights which can help clinics better understand overbooking and improve their performance.

First, unlike what one expects, no other measures taken, the reduction of appointment delay due to overbooking does not necessarily translate to an increase in patient show-up rate. As a result, although overbooking will likely increase the clinic’s expected profit, it always increases patient no-show rate. 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. Further, due to the increased no-show rate, overbooking may reduce clinic’s expected profits for panel sizes within a critical range, even when patients are quite tolerant of office delay. 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 to better bring out the benefits of reduced appointment delay is overbooking with controlled appointment queue, with which clinics only make appointments within 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. It also increase the panel size that the clinics can handle. Numerical study of different controlled appointment queue length also provides some insights into how to choose the approximate controlled queue length.

Second, 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, clinic’s expected profit decreases in a less drastic fashion in the critical range. Hence, overbooking mitigates the impact of panel size on the 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.

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 over all scheduling sessions, we encourage what we call a selective dynamic overbooking strategy, under which overbooking is applied to patients according to their no-show history. This strategy again helps clinics achieve higher expected profits and better patient show-up rates at the same time.

Finally, no-shows’ probability to re-schedule their missed appointments \( (r) \) is an important factor to consider when ministering overbooking because they take additional scheduling capacity and may lead to no-show again. An increase in \( r \) may imply significant reduction in the panel size a clinic can handle. With no control of appointment queue, the reduction of patient show-up rate after overbooking increases as \( r \) increases and the benefit of using overbooking with controlled appointment queue is larger for clinics with a larger \( r \).

As the first work to consider the impact of overbooking on both appointment delay and office delay, in the current study, we adopted a simple overbooking policy, NSOB, to demonstrate the model framework and obtain insights. We may apply this framework to other simple or sophis-
ticated overbooking strategies in the literature. We expect more involved analytical results for more complex overbooking policies, with similar insights. Finally, this work has opened doors for some interesting future study. For example, we have proposed two useful strategies for clinics. There remains further investigation regarding tactics in applying these strategies. In particular, for overbooking with controlled appointment queue, we provide some analysis and insights about how to choose the controlled queue length. However, like GS model and most of the other works, we assume patients who are turned away from appointments due to the controlled queue length do not leave the panel (switch to other clinics). Although patients turned away for appointments may not necessarily switch clinics (hence this is not an unreasonable assumption), when this is a concern, we can expect the choice of controlled queue length to be affected. It requires a different model to fully address this question in that case. 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


URL: http://mc.manuscriptcentral.com/UHSE Email: john.fowler@asu.edu


Figure 1: NE of the Patient Response Game for SBS

(a) Unique NE

(b) Two NE

Figure 2: Expected Patient Show-up Rate and Clinic Profit for Adjusted $K$ when $p_0 = \frac{1}{7}$

(a) $w_0 = 5$

(b) $w_0 = 8$

(c) $w_0 = 11$
Figure 3: Expected Appointment Delay and Office Delay when \(p_0 = \frac{1}{7}\)

(a) \(w_0 = 5\)

(b) \(w_0 = 8\)

(c) \(w_0 = 11\)

Figure 4: Expected Patient Show-up Rate and Clinic Profit when \(p_0 = \frac{1}{7}\)

(a) \(w_0 = 5\)

(b) \(w_0 = 8\)

(c) \(w_0 = 11\)
Figure 5: Expected Patient Show-up Rate and Clinic Profit for Different Overbooking Levels when $p_0 = \frac{1}{7}$

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

Figure 6: Expected Patient Show-up Rate and Clinic 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 7: Expected Show-up Rate and Profit for Different Re-booking Rates when \( p_0 = \frac{1}{7}, u_0 = 8, a = 1 \)

(a) re-booking \( r = 1 \)

(b) re-booking \( r = 0.5 \)

(c) re-booking \( r = 0 \)

Figure 8: Expected Show-up Rate and Profit for Different Re-booking Rates when \( p_0 = \frac{1}{7}, u_0 = 8, r = 1 \)

(a) OB \( a = 1 \)

(b) OB \( a = 0.5 \)
Figure 9: Expected Show-up Rate and Clinic Profit with $w_0 = 5$
Technical Appendices

A-1. Solution Process and Proof to the NE of Basic Model with SBS

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 (3), i.e.,

\[ i_c(q) = S(1 - q). \]  

(A-1)

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

\[ \hat{q}^c = 1 + \frac{c_l + c_u}{\alpha S} - \frac{1}{2S} - \sqrt{\left(1 + \frac{c_l + c_u}{\alpha S} - \frac{1}{2S}\right)^2 - \frac{2(c_u + \alpha w_0)}{\alpha S}}. \]  

(A-2)

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 6.** Let \( i^* = i(\hat{q}^c) \) and \( q^* = \hat{q}(i^*) \), where \( i(\cdot) \) and \( \hat{q}(\cdot) \) are defined in (3) 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 \( (q(i^* + 1), i^* + 1) \).
Proof. First, a more explicit expression of \( i(q) \) given in (3) is

\[
i(q) = \begin{cases} 
0 & \text{if } q = 1 \\
\frac{S-k}{S} & \text{if } \frac{S-k-1}{S} < q < \frac{S-k+1}{S}.
\end{cases}
\] (A-3)

In Figure 1(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, \( \hat{q}(0) = \frac{c_a+c_u+\alpha u_0}{c_l+c_u+\frac{q}{2}(S-1)} \leq \frac{c_a+c_l}{c_l+c_u+\frac{q}{2}(S-1)} \leq 1 \). In addition, \( \hat{q}(S) = \frac{c_a+c_u+\alpha u_0}{c_l+c_u+\frac{q}{2}(2S-1)}>0 \). 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 \((i_c(q^*), q^*)\), and at least one intersection of \( \hat{q}(i) \) and \( i(q) \). Further, because \( \hat{q}(0) \leq 1, \hat{q}(i) \) is a decreasing function in \( i \) and the slope of \( i_c(q) \) is \(-\frac{1}{S}\), from Figure 1(a), we can see that \( q^*(i_c(q^*)) \geq -\frac{1}{S} \).

If \( i(q^*) \in \mathbb{Z}_+ \), we have that \( i^* = i(q^*) \) and \( q^* = q_c^* \) and \((q^*, i^*)\) is a NE. Otherwise, because \( q^*(i) > q^*(i_c(q^*)) \geq -\frac{1}{S}, \forall i > i_c(q^*) \) (due to convexity) and based on (A-3), we can see that an intersection of \( \hat{q}(i) \) and \( i(q) \) is \((q(i(q^*)), i(q^*))\), i.e., \((q^*, i^*)\). It follows that \((q^*, i^*)\) is a NE when \( q(i) = q^* \).

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 \( \hat{q}(i) > -\frac{1}{S} \), for \( i > i_c(q^*) \), we have \( \hat{q}(i) > \frac{S-1}{S} \) for \( i > i_c(q^*) \), i.e., the curve \( \hat{q}(i) \) is above the line \( i_c(q), \forall i > i_c(q^*) \). Therefore, we can conclude that 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 \( q^*(i^* + 2) < q_b \). 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 \( q^*(i) < -\frac{1}{2S} \) 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

\[
\hat{q}'(i) = -\frac{\alpha(c_u + \alpha u_0)}{2(c_l + c_u + \frac{q}{2}(S + i - 1))^2},
\] (A-4)
i.e.,
\[ -\frac{1}{\hat{q}(i)} = \frac{2(c_l + c_u)^2}{\alpha(c_u + \alpha w_0)} + \frac{\alpha(S + i - 1)^2}{2(c_u + \alpha w_0)} + \frac{2(c_l + c_u)(S + i - 1)}{c_u + \alpha w_0}. \]  
(A-5)

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-5), 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-3) that when \(\hat{q}(i^* + 1) > \frac{S - i^*}{2}\) 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 \((q^*, i^*)\).

The NE characterized above can be demonstrated in Figure 1(a) and Figure 1(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.

**A.2. Proof of Lemma 2**

**Proof.** It is sufficient to show that \(\pi(\hat{i}) > \pi(i)\) if \(\hat{i} > i\) and show that \(\pi(i)\) is bounded for all \(i\). To simplify our exposition, we use \(\pi(k)\) to replace \(\pi(k|i)\) and apply \(\hat{\pi}\) to introduce parameters specific to \(\hat{i}\).

Let \(\rho = \frac{\lambda N}{BS}\). As derived in GS,
\[ \pi(0) = \frac{1}{\sum_{k=0}^{K} \rho^k \prod_{l=0}^{k-1} (1 - rp(l|i))} \]
and \(\pi(k) = \pi(0) \frac{\rho^k \prod_{l=0}^{k-1} (1 - rp(l|i))}{\prod_{l=0}^{k-1} (1 - rp(l|i))}\), where \(\prod_{l=0}^{k-1} \rho^k\) is set to 1 when \(k = 0\).

We observe that with \(i\) becomes larger, both \(\rho\) and \(p(k|i)\) decreases and \(\prod_{l=0}^{k-1} (1 - rp(l|i))\) increases for all \(k \geq 1\). So, given \(\hat{i} > i\), it can been seen that \(\hat{\pi}(0) > \pi(0)\). Indeed, there always
exist an index \( t \leq K \) such that \( \tilde{\pi}(t) < \pi(t) \). Otherwise equations \( \sum_k \tilde{\pi}(k) = \sum_k \pi(k) = 1 \) do not hold. Without loss of generality, let \( t \) be the smallest index of this property. Given that 

\[
\pi(k + 1) = \frac{\rho}{1 - p(k+1)} \pi(k),
\]

it is clear that \( \tilde{\pi}(k) < \pi(k) \) for all \( l \geq t \). Furthermore, it can be seen that

\[
\sum_{k \leq t} (\tilde{\pi}(k) - \pi(k)) = \sum_{t+1 \leq k \leq K} (\pi(k) - \tilde{\pi}(k)). \tag{A-6}
\]

Let \( E = e^{-\pi_{\tilde{\pi}t+1}} \). We next show that \( \sum_{k \leq K} \tilde{E}^{-k} \tilde{\pi}(k) - \sum_{k \leq K} E^{-k} \pi(k) > 0 \). For the case where \( t = K \), the conclusion follows easily given that \( \tilde{E}^{-k} > E^{-k} \) for all \( k \). We therefore consider the general case where \( t \leq K - 1 \). Equivalently, it can be proven by showing that

\[
\sum_{k \leq t} (\tilde{E}^{-k} \tilde{\pi}(k) - E^{-k} \pi(k)) > \sum_{t+1 \leq k \leq K} (E^{-k} \pi(k) - \tilde{E}^{-k} \tilde{\pi}(k)).
\]

For the left-hand-side, we have

\[
\sum_{k \leq t} (\tilde{E}^{-k} \tilde{\pi}(k) - E^{-k} \pi(k)) > \sum_{k \leq t} (E^{-k} \pi(k) - E^{-k} \pi(k)) = \sum_{k \leq t} E^{-k} (\tilde{\pi}(k) - \pi(k)) \geq E^{-t} \sum_{k \leq t} (\tilde{\pi}(k) - \pi(k)). \tag{A-7}
\]

For the right-hand-side, we have

\[
\sum_{t+1 \leq k \leq K} (E^{-k} \pi(k) - \tilde{E}^{-k} \tilde{\pi}(k)) < \sum_{t+1 \leq k \leq K} (E^{-k} \pi(k) - E^{-k} \tilde{\pi}(k)) \leq \sum_{t+1 \leq k \leq K} E^{-k} (\pi(k) - \tilde{\pi}(k)) \leq E^{-t-1} \sum_{t+1 \leq k \leq K} (\pi(k) - \tilde{\pi}(k)). \tag{A-8}
\]

Because of (A-6), we have \( E^{-t-1} \sum_{t+1 \leq k \leq K} (\pi(k) - \tilde{\pi}(k)) < E^{-t} \sum_{k \leq t} (\tilde{\pi}(k) - \pi(k)) \). It follows that \( \bar{q}(\tilde{i}) \geq \bar{q}(i) \). In addition, for any \( i \),

\[
\bar{q}(i) = \frac{c_u}{c_l + c_u} - (p_{max} - p_{min}) \sum_{k=0}^{K} (1 - e^{-kS/\bar{\pi}(k)}) \pi(k) \leq \frac{c_u}{c_l + c_u}.
\]
Result in Lemma 2 suggests that $\overline{q}(i)$ is an increasing function with respect to $i$. So, it has a unique intersection with $i_c(q)$. Furthermore, this lemma indicates that if the office delay impact is ignored, overbooking under constant $K$ always increases patient show-up rates.

A-3. Proof of Proposition 4

Proof. From the definition of $E[W|i]$, we have

$$\overline{q}_2 = \frac{c_u}{c_l + c_u} - (p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{c_l + c_u} \pi(k|i)}) - \frac{0.5\alpha(S + i - 1)\overline{q}_2 - \alpha w_0}{c_l + c_u}, \quad (A-9)$$

which leads to

$$\overline{q}_2(i) = \frac{c_u + \alpha w_0 - (c_l + c_u)(p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{c_l + c_u} \pi(k|i)})}{c_l + c_u + 0.5\alpha(S + i - 1)}$$

$$= \frac{c_u + \alpha w_0}{c_l + c_u + 0.5\alpha(S + i - 1)} - \frac{(c_l + c_u)(p_{\text{max}} - p_{\text{min}}) \sum_{k=0}^{K} (1 - e^{-\frac{kS}{c_l + c_u} \pi(k|i)})}{c_l + c_u + 0.5\alpha(S + i - 1)}. \quad (A-10)$$

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

$$\overline{q}_2(i)' = \frac{-\alpha(c_u + \alpha w_0) + \alpha(c_l + c_u) f(i)}{2(c_l + c_u + 0.5\alpha(S + i - 1))^2} - \frac{(c_l + c_u) f'(i)}{c_l + c_u + 0.5\alpha(S + i - 1)}$$

$$\geq \frac{-\alpha(c_l + c_u) + \alpha(c_l + c_u) f(i)}{2(c_l + c_u + 0.5\alpha(S + i - 1))^2} - \frac{(c_l + c_u) f'(i)}{c_l + c_u + 0.5\alpha(S + i - 1)}$$

$$= \frac{2[(c_l + c_u)^2 + \alpha(S + i - 1)(c_l + c_u) + 0.25\alpha^2(S + i - 1)^2]}{(c_l + c_u) f'(i)} - \frac{c_l + c_u + 0.5\alpha(S + i - 1)}{c_l + c_u + 0.5\alpha(S + i - 1)}$$

$$\geq \frac{2[(c_l + c_u) + \alpha(S + i - 1) + 0.25\alpha^2(S + i - 1)^2]}{(c_l + c_u) f'(i)} - \frac{(c_l + c_u) f'(i)}{c_l + c_u + 0.5\alpha(S + i - 1)}$$

$$\geq \frac{-\alpha}{2\alpha(S + i)} - \frac{(c_l + c_u) f'(i)}{c_l + c_u + 0.5\alpha(S + i - 1)}$$

$$= \frac{1}{2(S + i)} - \frac{(c_l + c_u) f'(i)}{c_l + c_u + 0.5\alpha(S + i - 1)}. \quad (A-11)$$
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 $c_l + c_u > \alpha$. From the proof of Lemma 2, it is clear that $f(i)$ decreases in $i$, i.e. $f'(i) < 0$. So, we conclude that $q_2(i) \geq -\frac{1}{2\gamma}$.

A-4. Results for Multiple Block Scheduling

A-4.1 The Basic Model for Multiple Block Scheduling

Different from the SBS model which treats the whole session (e.g., shift, day) as a single block, the multiple block scheduling, i.e. 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. While the single block model is a special case of multi-block model where all patients are scheduled in one block, the time slot based model (where each patient is scheduled a specific slot) is another special case where each patient slot is considered an individual block.

We focus on equal-length blocks, the most typical scheduling fashion where each block originally has the same number of patients, and assume that the NSOB policy is applied to each individual 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 (see Muthuraman and Lawley (2008); Zeng et al. (2010)), i.e., patients that were not able to be seen in an earlier block will flow to the next block. Specifically, since a physician is treated as a single server and her patients are examined in a sequential fashion, patients’ expected waiting time is determined by the number of patient arrivals in this block and the overflowing patients from the 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.

Let $X_j, Y_j$ and $L_j$ denote the number of patients arriving in block $j$, the number of patients remaining in clinic at the end of block $j$, and the number of patients served in block $j$, respectively.
It is easy to see that
\[ Y_j = (X_j + Y_{j-1} - L_j)^+ . \]

Next, we present the computation method for the expected waiting time. Readers are directed to Muthuraman and Lawley (2008) for more detailed derivation.

First, define \([Q_{j,l}]\) as the patient arrival matrix, where \(Q_{j,l}\) is the probability that \(l\) patients arrive at the beginning of block \(j\), and \([R_{j,k}]\) as the overflow matrix, where \(R_{j,k}\) is the probability that \(k\) patients overflow from block \(j\) to block \(j+1\). Note that the number of patients follows a Poisson distribution if the queue is not empty and the number of patient who show up follows a binomial distribution. So, for block \(j\) with \(\hat{S}_j (\equiv S + i_j^*)\) patients scheduled, these matrices are computed as
\[
Q_{j,l} = \binom{\hat{q}_j^*(l)}{\binom{p_j^*(l) + (S + i_j^*) - l}} \binom{(S + i_j^*)}{l}
\]
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 \(f_{L_j}(m) = e^{-\mu t}(\mu t)^m/m!\) 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 \(E[Y_{j-1}] = \sum_k kR_{j-1,k}\) and therefore the expected waiting time for patients in block \(j\), given the amount of overbooking in block \(j\), \(i_j\), is
\[
E[W_j|i_j] = \frac{1}{2}(S + i_j - 1)q_j(i_j) + E[Y_{j-1}] . \tag{A-12}
\]

Using the same approach as used for SBS, we have the patient show-up rate in block \(j\), \(q_j\), in response to the amount of overbooking for this block, \(i_j\), as
\[
q_j(i_j) = \min\{\hat{q}_j(i_j), q_0\} , \tag{A-13}
\]
where
\[
\hat{q}_j(i_j) = \frac{c_u + \alpha(w_0 - E[Y_{j-1}]}){c_q + c_u + \frac{3}{2}(S + i_j - 1)} . \tag{A-14}
\]
To ensure \(\hat{q}_j(i_j) > 0\), 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 $c_u$, as shown in the numerical study.

Define

$$\hat{q}_j^c = 1 + \frac{c + c_u}{\alpha S} - \frac{1}{2S} - \sqrt{(1 + \frac{c + c_u}{\alpha S} - \frac{1}{2S})^2 - \frac{2(c_u + \alpha(w_0 - E[Y_{j-1}])))}{\alpha S}.$$ (A-15)

Comparing (A-14) with (6) and (A-15) with (7), we find that $\hat{q}_j$ 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 with a vertical translation resulted from $E[Y_{j-1}]$. So, given $E[Y_{j-1}]$, we can obtain the following results of NE in block $j$ of the MBS model.

**Theorem 7.** 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 (A-14). 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 $j$. Therefore, we can solve the NE sequentially, starting from block 1. Since there is no patient overflow for block 1, we can solve for the NE of this block following Theorem 1 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 7. 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.

**A-4.2 The 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 calcu-
lated 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 and how we address them.

First, for multiple blocks, we again have the patient overflow problem and we can follow the same equations in Section A-4.1 to calculate the expected overflow from block \( j - 1 \) to block \( j \), \( E[Y_{j-1}] \). Then, we can calculate the block-specific parameters for block \( j \) using (A-16)-(A-19), corresponding to equations (15)-(18) for the single block model. In equations (A-16)-(A-19), \( i \) represents the overbooking levels in all blocks, i.e., \( i = \{i_1, i_2, ..., i_B\} \). The dependence of the block parameters (e.g., \( \bar{q}_j \)) on overbooking levels of all blocks, \( i \), reflects the inter-relationship among all the blocks.

For all \( j = 1, ..., B \),

\[
q_j(k|i) = \frac{c_u}{c_l + c_u} - (1 - e^{-k/A'}) (p_{\text{max}} - p_{\text{min}}) - \frac{\alpha (E[W_j|i] - w_0)}{c_l + c_u}, \tag{A-16}
\]

\[
E[W_j|i] = \frac{1}{2} (S_j + i_j - 1) \bar{q}_j(i) + E[Y_{j-1}], \tag{A-17}
\]

\[
\bar{q}_j(i) = \sum_{k=0}^{K} q_j(k|i) \pi(k|i), \tag{A-18}
\]

\[
\hat{S}_j(\bar{q}_j) = S + [S(1 - \bar{q}_j)], \tag{A-19}
\]

where \( A' = \frac{\sum_{j=1}^{B} \hat{S}_j}{BS} D \). Note that \( \bar{q}_j(i) \) 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 overbooking level (due to different office delay and patient show-up rate), appointment delay seen by patients requesting appointments 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 average no-show rate over all blocks when appointment delay is \( k \), i.e.,

\[
p(k|i) = \frac{\sum_{j=1}^{B} p_j(k|i) * \hat{S}_j}{\sum_{j=1}^{B} \hat{S}_j}, \tag{A-20}
\]

together with equations (9)-(12) where \( B \cdot S \) is replaced by \( \sum_{j=1}^{B} \hat{S}_j \) to calculate the steady-state probabilities of appointment delay.

Due to the complexity mentioned above, analytical characterization of the NE is intractable. We again develop an algorithm (Algorithm 2 in Appendix A-6) to solve for the block-specific as
well as the all-block parameters iteratively to obtain the equilibrium solutions.

A-4.3 Numerical Study for Multiple Block Scheduling

We performed a set of numerical study that investigates the performance of overbooking in the multi-block model. Results from the multi-block model are qualitatively similar to those in the single block model (e.g. similar shape of the no-show rate figures and the expected profit figures, the existence of the critical panel size range, etc.). Hence, the insights from the single block model are also applicable to the multi-block model. Next, we present results that are particular to the multi-block models.

In the study, we particularly look at how the number of blocks affects the patient show-up rates. We look at scheduling the same length of a session (e.g., a shift or a day) where the session is divided into (1) a single block with the whole session length, (2) two blocks each with a half of the session length, and (3) 4 blocks each with a quarter of the session length. Overbooking is individually applied to each block based on its respective no-show rate. To keep apple-to-apple comparison, block size $S_j = \frac{S}{2}$ for all blocks and office waiting time tolerance $w_0$ is adjusted to $\frac{w_0}{2}$.

We make the following observations from Figure A-1 and Figure A-2: (a) The results from the multi-block models are in general quite similar to those from single block model. In fact, single-block models, which are computationally much easier, provide good approximations to multi-block models for both the expected show-up rate and clinic’s expected daily profit. The larger $w_0$ is, the better the approximation. When patients’ tolerance of office delay ($w_0$) is sufficiently large, the expected show-up rates for all blocks under multi-block model converge to the same value under the single block model (we increased $w_0$ to a larger value to show this trend in Figure A-1 and Figure A-2). This is because compared to the single block model, the biggest difference is the overflow of patients between the adjacent blocks, which brings in additional office delay. With larger $w_0$, the negative impact of office delay is smaller. (b) The expected show-up rate decreases from the first block to the last block, due to the increasing office delay caused by overflow. (c) The impact of the number of blocks is less significant when the panel size is beyond the critical range.
Figure A-1: Expected Patient Show-up Rate and Clinic Profit for Multi-block Models when $p_0 = \frac{1}{7}$

Figure A-2: Expected Patient Show-up Rate over Blocks
A-5. Algorithm 1 for I-SBS

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 $\bar{q} = \frac{c_u}{c_u + c_a}$. Then set $\hat{S} = S + [S(1 - \bar{q})]$;
   
   b. Compute expected waiting time $E[W]$ (equation (16)).
   
   c. For $k = 0, \ldots, K$, compute the show-up probability $q(k)$ using (15), hence $p(k) = 1 - q(k)$;
   
   d. Set $\pi_1(k) = \pi_2(k)$ and apply (9)-(12) to compute $\pi_2(k)$, $k = 0, 1, 2, \ldots, K$;
   
   e. Calculate the expected show-up rate $\bar{q}$ according to equation (17).

3. Compute the expected daily profit.

A-6. Algorithm 2 for I-MBS

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 $\bar{q}_j = \frac{c_u}{c_l + c_u}$ and $\bar{q}_j^\prime = \bar{q}_j - 2\epsilon$ ($\bar{q}_j^\prime$ is used to see whether $\bar{q}_j$ has converged). Also, set $\hat{S}_j = S_j + [S_j(1 - \bar{q})]$.
      
      II. While $|\bar{q}_j - \bar{q}_j^\prime| > \epsilon$, do
           
           i. Set $\bar{q}_j = \sigma\bar{q}_j + (1 - \sigma)\bar{q}_j^\prime$ (where $\sigma \in (0, 1)$ is a randomly selected constant) and update $\hat{S}_j = S_j + [S_j(1 - \bar{q}_j)]$;
           
           ii. Compute the expected waiting time $E[W_j]$ (equation (A-17)) 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 A-4.1;
           
           iii. For $k = 0, \ldots, K$, compute the show-up probability $q_j(k)$ using (A-16);
iv. Set $\bar{q}_j' = \bar{q}_j$ and obtain $\bar{q}_j$ using $\pi_2(k)$ from (A-18);

III Compute overflow $E[Y_j]$;

b. Use (A-20) to compute the weighted no-show rate $p(k)$ using $q_j(k)$ ($p_j(k) = 1 - q_j(k)$) and $\bar{S}_j$;

c. Set $\pi_1(k) = \pi_2(k)$ and apply (9)-(12) to compute $\pi_2(k)$, $k = 0, 1, 2, ..., K$;

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