

# APPOINTMENT SCHEDULING WITH OVERBOOKING TO MITIGATE PRODUCTIVITY LOSS FROM NO-SHOWS

Linda R. LaGanga & Stephen R. Lawrence

Mental Health Center of Denver  
4141 East Dickenson Place  
Denver, CO 80222  
(303) 504-6665  
[Linda.Laganga@mhcd.org](mailto:Linda.Laganga@mhcd.org)

Leeds School of Business, UCB 419  
University of Colorado at Boulder  
Boulder, CO 80309-0419  
(303) 492-4351  
[Stephen.Lawrence@colorado.edu](mailto:Stephen.Lawrence@colorado.edu)

## ABSTRACT

The challenge of balancing the interests of patients with those of healthcare providers is increased when patients fail to show up for scheduled appointments. Overbooking appointments mitigates the lost productivity caused by no-shows but increases patient wait time and provider overtime. In this paper, simulation analysis is used to develop and test the performance of scheduling rules that are designed specifically to accommodate excess overbooked appointments. Our analysis provides new insights into rules that perform well to increase provider productivity while balancing the increased waiting time and overtime costs of overbooked schedules.

**Keywords:** Appointment Scheduling, No-shows, Overbooking, Service Operations, Simulation

## INTRODUCTION

In this paper, we build upon and extend the double-booking, block scheduling, and wave-scheduling policies devised by practicing clinicians to develop and measure the performance of a number of scheduling rules based on these policies. We adjust traditional appointment scheduling performance measures to capture the operating dynamics of overbooked appointment scheduling systems, determine their effectiveness when overbooking is used to compensate for the lost productivity of no-shows, and provide recommendations for improving performance in overbooked appointment scheduling systems. Our analysis is useful for schedulers and health care providers to identify and evaluate operational and policy changes that will boost clinic productivity and improve patient service.

The scheduling rules that we develop and test in our simulation experiments are designed specifically to analyze the effects of the placement of the extra appointments in an overbooked appointment schedule. We compare traditional double-booking and other multiple-booking scheduling patterns suggested by providers with alternative scheduling patterns that use compressed inter-appointment arrival times instead of, or in combination with, multiple booking.

## RESEARCH METHODOLOGY

Modeling was done in two stages. First, we developed spreadsheet-based planning models of all experimental schedules to calculate patient wait time at every appointment time and for the entire

schedule for the case in which every scheduled patient shows up. This analysis was useful in identifying where large wait times are likely to accumulate and to support experimentation and development of alternative scheduling rules. Then, to test the performance of various schedules with stochastic no-shows, we simulated the operation of the planning models for various levels of show rate  $S$ .

### The Clinic Model and General Scheduling Rules

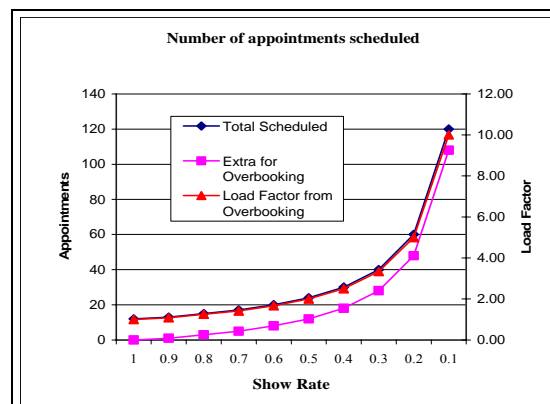
We model a realistic clinic session using the parameters of an actual outpatient psychiatric clinic that we studied. In this clinic, a normal morning clinic session runs for four hours from 8:00 a.m. – 12 p.m. The service time for each patient is 20 minutes without variation so that the clinic size is  $N = 12$ .

If every patient showed up with certainty, then there would be no need to overbook, there would be no patient wait time, the provider would be utilized 100% of the time, and there would be no overtime required to serve all patients. However, the clinic experiences a significant no-show rate ( $S < 1$ ), so if the target number of patients to be served remains at  $N = 12$ , then additional overbooked appointments must be added to the clinic schedule. One way to schedule extra appointments into the clinic session is to compress the inter-arrival time between appointments  $T$  proportionally to the show-rate  $S$ , so that  $T = DS$ , or for the clinic under consideration,  $T = 20S$ .

### Analysis of Overbooking Dynamics

Overbooking a scheduling system introduces new challenges in constructing schedules and comparing performance. The first consideration is the process of scheduling the extra  $K-N$  appointments. Overbooking is achieved by adjusting the time intervals between appointments, using block scheduling of multiple patients at one or more schedule times, or combinations of these approaches to fit the extra appointments into the schedule. For high show rates, developing potential schedules can be handled by fitting a small number of extra appointments into the schedule through adding individual appointments to the baseline schedule of  $N$  appointments spaced  $D$  time units apart. This becomes more challenging as show rates decrease because the number of appointments that must be added to the schedule becomes large, as shown in Figure 1.

**FIGURE 1: Total and extra appointments scheduled due to overbooking**



We construct and evaluate potential wave schedules, identify schedule block sizes that contribute to congestion and overtime, and modify the schedule to attempt to alleviate these conditions. Although we can calculate the probability of every possible number of patient shows or no-shows occurring, the analytical calculation of expected wait time and overtime is intractable because overtime occurs not only when capacity is exceeded, but also as the result of the time-based appointment positions in which patients show up. Thus we use simulation to model the performance of our schedules at varying patient show rates.

### SIMULATION RESULTS AND ANALYSIS

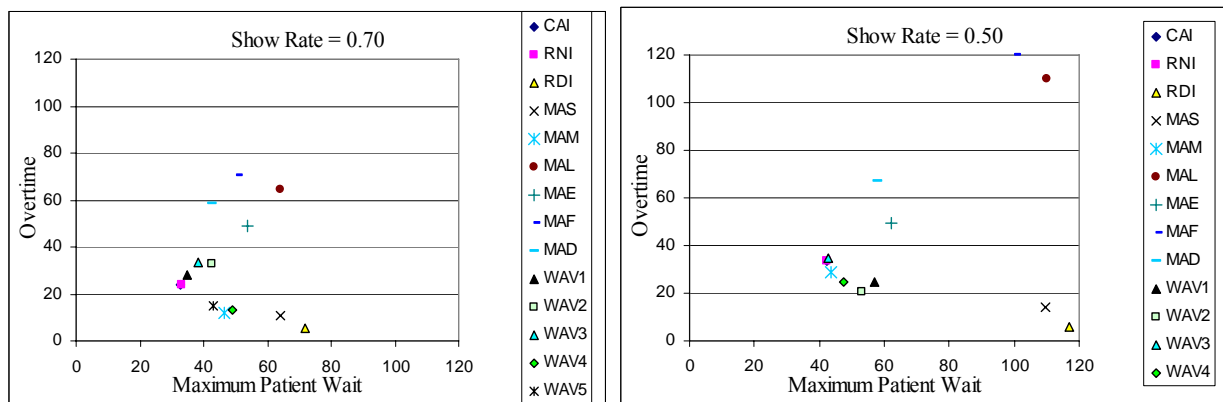
We developed a simulation model with deterministic patient no-shows for each of the 67 schedules that we developed and analyzed in our planning models. For each schedule simulated, we completed 10,000 replications for a total of 670,000 observations of the schedules' performance. The number of replications was determined by conducting pilot studies preceding the main experiment that indicated that the half-widths of the 95% confidence intervals were less than 1% of the point estimates for the performance measures of interest.

We graphically present our results to display and analyze the performance trade-offs between wait time and overtime. For each show rate  $S$ , we consider the objective of choosing the schedule with the best performance, measured as the minimum of the weighted sum of maximum patient wait time,  $W$ , and provider overtime,  $\mathcal{O}$ . Then, for  $\pi =$  the cost per minute of maximum patient wait time and  $\omega =$  cost per minute of provider overtime, the expected cost of using a particular schedule is

$$C = \pi W + \omega \mathcal{O}. \tag{1}$$

The results are useful in separating the schedules that should be considered for implementation from those that should not because they are dominated by one or more others that have smaller values for both patient wait time and provider overtime [3]. We can identify the best schedule that minimizes cost among those tested by finding the point where  $\pi/\omega$  has a value between the slopes of two adjacent line segments on the efficient frontier [4].

**FIGURE 2: Simulation Results for  $S = \{0.7, 0.5\}$**



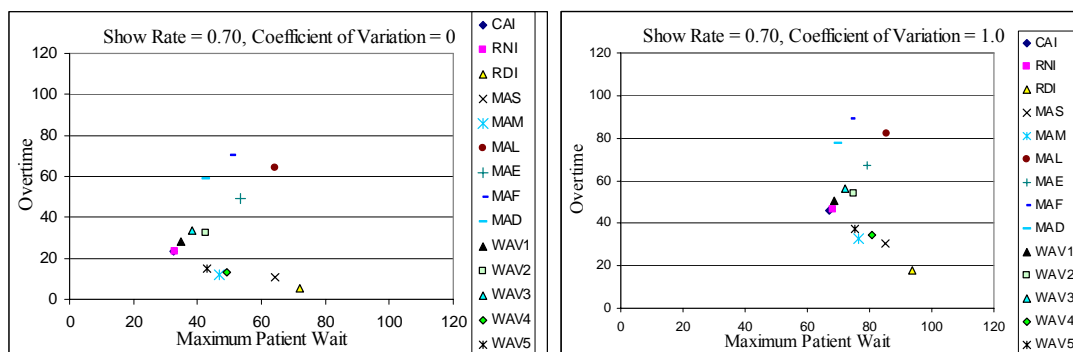
As shown in Figure 2, the set of schedules on the efficient frontier varies with show rate. The points for  $S = 0.70$  differ from those of  $S = 0.90$ . At  $S = 0.70$ , MAS (analogous to the Bailey-Welch rule [1][2] with 5, instead of 2, appointments scheduled in the initial block) is far off the efficient frontier, but another block-scheduling variation is MAM, which schedules several blocks of size 2 and is on the efficient frontier. WAV5 was developed to improve performance of schedule WAV1 by reducing overtime, which did, indeed, occur in the simulation results. The trade-off is that wait time increased with the schedule change, but the improved rule made it onto the efficient frontier, unlike WAV1, which was dominated by the first two rules.

At show rates  $S = \{0.70, 0.50\}$ , WAV4 was designed to represent the wave scheduling pattern used by some providers to schedule repeated patterns of appointments in block sizes of  $\{3, 2, 1\}$  [2]. The schedule performs better for  $S = 0.70$ , with 13.23 minutes of overtime and 49.078 minutes of maximum wait time, than for  $S = 0.50$ , with 24.44 minutes of overtime and 47.638 maximum wait time (unless the cost of patient wait time is much higher than provider overtime). But the schedule would not be considered for implementation for  $S = 0.70$  because it is dominated by other schedules. In contrast, for  $S = 0.50$ , the schedule results are positioned on the efficient frontier and would be considered, which illustrates that the best schedule for a given show rate depends not only on the scheduling rule and the show rate itself but also on the relative performance of the alternative rules. At each of the levels of  $S$  analyzed, several rules emerged for possible selection according to weighting of cost factors and practical considerations.

### SENSITIVITY ANALYSIS

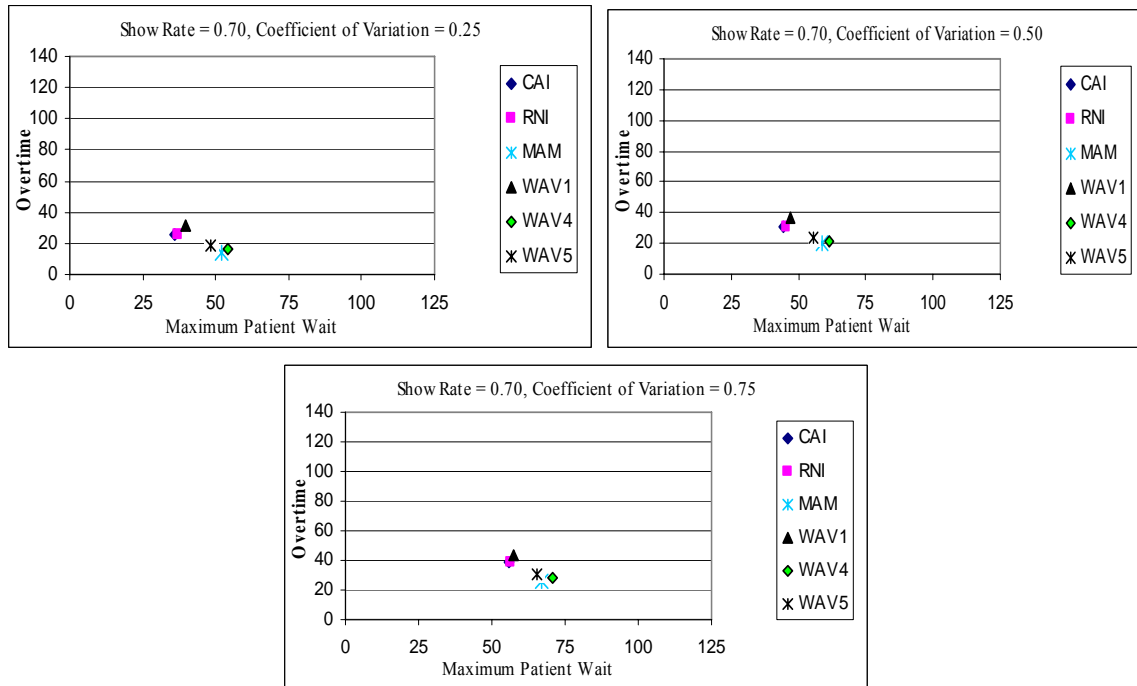
An initial assumption in our analysis was that service times were constant, which is consistent with psychiatric clinics that we studied. To determine the effects of variable service times, we conducted additional simulation experiments for  $S = 0.70$  and the five rules on or near the efficient frontier. We modeled service time as a Gamma distribution with four parameter pairs  $(\alpha, \beta)$  of (16, 1.25), (4, 5), (1.78, 11.25), (1, 20) for four levels of *provider service time variability*,  $c_s = \sigma/\mu$ , where  $\sigma$  is the standard deviation of service time and  $\mu$  is its mean, for five levels of service time variability,  $c_s = \{0.0, 0.25, 0.5, 0.75, 1.0\}$ , including the original experiments where  $c_s = 0$ . We used the Gamma distribution because it is bounded from below by 0 (no negative service times) and is skewed to the right with a long right tail allowing the possibility of extended service times, which may occur in handling emergency service.

**FIGURE 3: Comparison of plots for constant service time ( $c_s = 0$ ) and high variation  $c_s = 1.0$**



As shown in Figure 3, when service time changes from constant ( $c_s = 0$ ) to high variation ( $c_s = 1$ ), the pattern of plotted points for each scheduling rule in relationship to the other scheduling rules remains very similar, and the same scheduling rules {CAI, RNI, RDI, MAM, WAV1, WAV4, WAV5} remain on or near the efficient frontier. For every scheduling rule, maximum patient wait time and provider overtime increase with increased service time variation.

**FIGURE 4: Schedules on the efficient frontier for  $c_s = \{0.25, 0.5, 0.75\}$**



Focusing on the scheduling rules on or near the efficient frontier in Figure 4 for the three additional levels of service time variation tested, for  $c_s = \{0.25, 0.50, 0.75\}$ , reveals that the pattern of points on the efficient frontier remains unchanged with changes in service time variation, and the magnitude of maximum patient wait time and provider overtime increase with increased service time variability. Therefore, we conclude that service time variability has little impact on the efficient frontier and is, therefore, immaterial in comparing the performance of various scheduling rules; however, it does impact the magnitude of the performance components — maximum patient wait time and provider overtime — of the scheduling rules.

## SUMMARY AND RECOMMENDATIONS

The simplest practical overbooked schedule, the RNI rule, compresses all inter-appointment times by the same factor  $S = \text{show rate}$ , and then places each resulting appointment time on the nearest practical clock time. Our simulation results show that for the full range of show rates tested, this schedule is always among the schedules that perform well and should be considered for implementation because it is not dominated by other policies.

In contrast, we would not recommend scheduling policies with very tight uniform compression at any show rate because they allow no catch-up time, and large accumulations of patient wait time occur, even with stochastic no-shows. In particular, the RDI scheduling rule results in high patient wait times for all show rates  $S$ ; this would never be acceptable to patients or their payers and advocates, especially with the existence of alternative schedules that result in much less wait time. Similarly, schedules that use block scheduling of multiple patients at the same time, especially in large block sizes, are not recommended unless  $S \leq 0.50$ . Otherwise, they result in large patient wait time. Breaking large blocks into smaller ones improves performance.

From our experiments with  $S = 0.90$ , we found that patient wait time can be avoided entirely by scheduling one extra appointment at the end of the clinic session, resulting in an average of only 18 minutes of overtime. If less overtime is desired, this can be accomplished with the wave schedule that compresses selected inter-appointment times from  $D = 20$  minutes to 15 minutes to avoid a large accumulation of patient wait time anywhere in the schedule and results in average maximum patient wait time of about 12 minutes. Hence, we recommend that these schedules be considered if patients and providers can accept these relatively small impacts for the benefit of adding capacity to offer service access to one extra patient per provider per clinic session (two per day if sessions run both morning and afternoon) and to allow providers to be more productive.

For schedules under consideration, formal analysis of the benefits and costs can be evaluated using a net utility  $U$  measure [5]:

$$U = MS(K - N) - \pi W - \omega \mathcal{O} \quad (2)$$

where  $M$  is the marginal benefit of servicing an additional patient.

## REFERENCES

[1] Bailey, N. T. (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society, Series B, 14(2)*, 185-199.

[2] Baum, N .H. (2001). Control your scheduling to ensure patient satisfaction. *Urology Times, 29(3)*, 38-43.

[3] Fries, B. E., & Marathe, V. P. (1981). Determination of optimal variable-sized multiple-block appointment systems. *Operations Research, 29(2)*, 324-345.

[4] Ho, C., & Lau, H. (1992). Minimizing total cost in scheduling outpatient appointments. *Management Science, 38(12)*, 1750-1763.

[5] LaGanga, L. R. & Lawrence, S. R. (2007). Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences, 38(2)*, 251-276.

[6] Welch, J. D., & Bailey, N. T. (1952). Appointment systems in hospital outpatient departments. *The Lancet*, May 31, 1105-1108.