Clinical Appointment Scheduling:
The Cost of No-Shows and the Value of Overbooking

Linda LaGanga [†]
Stephen Lawrence [‡]

October 19, 2005

Leeds School of Business
University of Colorado
419 UCB
Boulder, CO 80309-0419

ABSTRACT

We analyze appointment scheduling and daily patient flow through an outpatient mental health clinic. Using empirical data, we develop several new scheduling heuristics that work to increase patient throughput and clinic productivity, while decreasing patient wait-times. Simulation experiments demonstrate the effectiveness of our methods compared to various other heuristics reported in the practitioner literature.

[†] Linda.LaGanga@colorado.edu
[‡] Stephen.Lawrence@colorado.edu

Introduction

Cost-effective delivery of health care services requires service providers to attain high levels of productivity (Tonges, 1985) and utilization (Managed Care Weekly Digest, 2003). The failure of some patients to arrive for their scheduled appointments increases provider idle time and reduces the expected number of patients that are actually seen in a day (Shonick and Klein, 1977), which reduces the clinic's revenues and denies some patients timely access to needed services. As the literature included in the next section demonstrates, some researchers in outpatient appointment scheduling recommend overbooking to solve the problem of no-shows, but have given little or no consideration to the costs.

The contribution of this paper is to examine the performance impacts of patient no-shows and overbooking, including a new utility model that includes both the costs and benefits of overbooking and shows clinical providers and administrators how beneficial or costly overbooking would be for the specific operating conditions of their clinics.

The rate of patient failures to arrive ("no-show rates," Barron (1980)) can be significant. The problem may be particularly severe for community mental health centers, pediatric clinics, hospitals, and neighborhood medical and dental clinics (Bean and Talaga, 1995). For example, in an outpatient community mental health center that we observed, almost 30% of adult patients failed to show up for their appointments with psychiatrists. Thus for every 100 scheduled appointments, 30 of the time slots were unused, which reduced provider utilization and productivity. This also means that 30 patients who needed appointments were denied access until a later date, which may reduce customer satisfaction and quality of health care (Chesanow, 1996; Larkin, 1999; Murray and Berwick, 2003).

As Barron (1980) recognized, the problems caused by no-shows may be solved by reducing the occurrence of no-shows or by scheduling to reduce their impact. On one hand, reducing the no-show rate reduces the source of the uncertainty in provider productivity, but may be uncontrollable by the clinic or costly to accomplish. On the other hand, overbooking compensates for uncertainty by boosting productivity but has the potential to harm customer service because of the day-to-day variability in the number of patients that actually show up. When the number of patients who show up exceeds normal system capacity, there are costly increases in patient wait time and the length of the clinical work day.

The purpose of this paper is to support decision-making in responding to the problems caused by no-shows. We compare system performance at varying levels of no-show rate with and without overbooking. Using analytical and simulation models, we determine the value of overbooking in terms of expected utility obtained from serving patients, including the costs of patient wait time and overtime operation. We consider non-financial utility (Metters and Vargas, 1999) in our model because providers in not-for-profit health care systems often value serving patients in need more than they value revenue benefits. Therefore this study is relevant both to for-profit and not-for-profit health care providers.

We extend previous outpatient scheduling literature by focusing specifically on no-shows and explicitly examining the additional costs incurred by overbooking. We add overtime operation to the cost function used in previous research that focused only on patient wait time and provider idle time. In addition, we analyze scheduling performance for the wide range of realistic no-show rates that have been cited in medical and health care literature to identify the conditions under which overbooking is helpful or harmful to scheduling performance, thus providing guidance in the important decision of whether or not to overbook.

Overview of No-shows and Appointment Scheduling

Many authors have contributed to the literature on no-shows from a variety of disciplines and perspectives including medical practice, health care administration, operations management, transportation planning (particularly airline revenue management), and marketing. Health care researchers and some practitioners have focused on finding causes of no-shows and eliminating or reducing them. They consider costs such as analysis of patients and their behavior and the implementation costs of programs or practices to boost patient attendance rates (Bean and Talaga, 1995; Campbell, et al., 2000; Garuda et al., 1998; Shonick and Klein, 1977).

Reported reasons for no-shows include lack of transportation, scheduling problems, overslept or forgot, and lack of child care (Campbell, et al., 2000). The probability of patient no-shows may relate to factors such as patient age, gender, number of previous appointments (Shonick and Klein, 1977), appointment lead time (Bean and Talaga, 1995) and Medicaid status (Rust et. al, 1995). McCarthy et al. (2000) and Sharp and Hamilton (2001) suggest that no-show rates might increase if wait times grow too long at the clinic. Approaches that have been successfully applied to reduce no-shows include sending patients reminder cards (Rust et al., 1995), calling patients to remind them of appointments, and providing information about public transportation (Bean and Talaga, 1995).

Operations management and statistical perspectives are evident in studies of clinical appointment scheduling systems that measure performance as the weighted sum of patient wait time and provider idle time costs (Bailey, 1952; Bailey and Welch, 1953; Ho and Lau, 1992; Welch and Bailey, 1952). These studies identify the no-show rate as a significant factor in schedule performance and measure some of the effects, but do not focus on how to handle no-shows or reduce their negative impacts in the scheduling system. Out of 36 articles categorized

in a recent review of outpatient scheduling literature by Cayirli and Veral (2003), only 11 include the possibility of no-shows. Recommendations for handling the problem are even more limited. Only four of the articles reviewed (Blanco White and Pike, 1964; Fetter and Thompson, 1966; Vissers and Wijngaard 1979; Vissers, 1979) include scheduling adjustments or operational considerations such as overbooking to mitigate the effects of no-show behavior.

Shonick and Klein (1977) show how to use the probabilities of patient no-shows to overbook enough patients so that the expected number of arrivals is equal to the target number to be seen, but do not consider overtime as a potential risk that could increase costs. Rohleder and Klassen (2002) consider overtime and overbooking as possible methods to deal with temporary or chronic high demand for appointments. They hold the no-show rate constant at 5% and do not include no-shows as an experimental factor in their simulation model of appointment scheduling rules. They also include ending time of the clinical day and server utilization as server-oriented measures of schedule performance, but do not integrate them into an overall performance measure.

The medical practitioner literature recommends "wave scheduling," using variable appointment intervals and patient batch-sizes to build small queues of patients while allowing the provider time to catch up at the end of each period in order to balance provider productivity with patient wait time (Barron, 1980; Baum, 2001; Chesanow, 1996; Chung, 2002; Cole, 2003; McCarthy, 2002; McCord, 1996; Schroer and Smith, 1977; Silver, 1975; Zeff, 1995). But this literature does not differentiate no-shows from varying service times in their impacts on schedule performance.

There are large variations in no-show rates among medical specialties and geographic regions (Sharp and Hamilton, 2001) and patient populations and their reasons for no-show

behavior (Garuda et al., 1998). Some studies, including a case study by Brahimi and Worthington (1991) and an official study of hospitals in England and Wales (Warden, 1995), have reported patient no-show rates of 10%. Sharp and Hamilton (2001) reported a 12% no-show rate at outpatient clinics in the UK. According to Barron (1980), eight studies at inner city, community health centers, and university medical centers indicate no-show rates of 10-30% while the estimated no-show rates for private practice are 2-15%. An even wider range of no-show rates, 3-80%, is reported in a study by Rust et al. (1995) of 200 public pediatric clinics. Our overbooking models are designed to handle a wide range of no-show rates, which, as these studies demonstrate, is necessary for the models to be useful in a variety of clinical practices.

In contrast to the health care industry, in the airline industry the practice of overbooking to compensate for no-shows has been extensively studied as revenue management to predict and balance the costs and benefits of overbooking (Hillier and Lieberman, 2001; Rothstein, 1971; Smith et al., 1992; Van Ryzin and Talluri, 2003). Similar to airline seats, daily clinical appointments can be viewed as perishable assets that cannot be held in inventory, thus driving the need to overbook to minimize the number of assets that perish unused because of the occurrence of no-shows.

Overbooking Model

We model the impacts of no-shows by first considering the expected daily cost of no-shows and the value of no-show reduction without overbooking, and then comparing the expected utility from overbooking at varying levels of no-show rates and clinic sizes. The model is based on the following definitions and assumptions:

1) $\boldsymbol{M}$ = Marginal benefit obtained from servicing each patient.

This may be the revenue received for each patient on a fee-for-service basis, or it could be the perceived gain in utility obtained in a not-for-profit system.

2) ***D*** =Service time duration.

We assume *D* is constant to allow us to focus on the uncertainty caused by no-shows rather than on any uncertainty introduced by the variation of service times. Varying service times could be easily incorporated into our models by specifying a probability distribution for service time.

3) ***N*** = provider capacity or "Clinic size."

This is the target number of patients to receive service from each provider during one "clinical session," which is defined as the entire clinical day or a segment of the day such as the morning or afternoon that has defined times at which the clinic starts and is intended to end. Our assumption that *N* is finite precludes the use of steady-state queuing models.

4) ***C*** =Allocated duration of the clinical session. This is the total duration of time the clinic session would take for a provider to service *N* consecutive patients all of whom arrive at the time of their scheduled appointments which are scheduled *D* units of time apart. Thus, *C=ND.*

5) ***S*** = the show rate. If all patients show up with certainty, then $S = 1$.

6) ***R*** = the no-show rate. $R = 1 - S$.

7) ***K*** = the total number of appointments booked.

The clinic allows overbooking by scheduling $K \geq N$ patients.

8) ***x*** = the number of patients who show up. We assume that all patients who show up are served, even if they must wait or if the provider must work overtime. Therefore, *x* is the number of patients served.

9) We assume that for each appointment, the outcome that the scheduled patient shows up for the appointment is an independent Bernoulli event that occurs with probability *S*, so that the number

of scheduled patients who do show up when *K* patients are scheduled is a random variable that is binomially distributed (van Ryzin and Talluri, 2003), so that $E[x] = KS$.

9) ***T*** = the interval of time between scheduled appointments. We assume that appointments, including those that are overbooked, are spread at even intervals throughout the clinical session (Vissers, 1979). Therefore, we divide the duration *C* of the clinic session by *K*, the number of appointments booked, to obtain $T = C / K = ND / K$.

10) Demand for appointments is greater than or equal to the supply of appointments slots.

11) Appointments are scheduled with a specific provider. Therefore, if a clinic employs multiple providers, the system is modeled to represent a collection of individual providers, each operating as a single-server appointment system.

12) If the number of patients who show up exceeds clinic size, then overtime costs are incurred.

13) If patients show up in excess of capacity during any interval of time within the clinical work day, then costs are incurred due to patient wait time.

Appointment Scheduling without Overbooking

We start by considering the expected utility of an appointment scheduling system in which there are no-shows but no overbooking. This provides a baseline against which to determine whether the cost of no-shows justifies the expense of remedial actions such as attempting to reduce the no-show rate or implementing overbooking policies to compensate for lost productivity.

Let *M* be the marginal benefit obtained from servicing each patient. This may be the revenue received for each patient on a fee-for-service basis, or it could be the perceived gain in utility obtained in a not-for-profit system. Let *N* be the target number of patients to be seen during the entire clinical day or clinic session (such as a morning or an afternoon.) Let *S* be the show rate,

so if all patients show up with certainty, then $S = 1.0$. Then $R = 1\text{-}S$ is the no-show rate. The gross utility U_G of a clinic session can be expressed as $U_G = MN$. The net utility U_N with no-shows is $U_N = MSN$, so the impact of no-shows on utility can be expressed as the utility cost (reduction of utility):

$$U_C = U_G - U_N = MN - MSN = MN(1 - S) = MNR \quad (1)$$

Thus, the marginal cost of no-shows with respect to no-show rate *R* is simply *MN*. For example, if the capacity of a provider is 20 patients a day, then at a marginal utility of $100 per patient, the provider's maximum daily utility is $20 \times \$100 = \$2{,}000$ and the cost of each percent increase in the no-show rate is $0.01 \times 20 \times \$100 = \$20$ per day. The daily cost of a no-show rate of 30% is $0.30 \times \$2{,}000 = \$600$. If the clinic can reduce the no-show rate to 20%, then the daily cost of the no-show rate decreases to $400. When there is no overbooking, predicting costs is easy because there is a direct linear relationship between *S* and total daily utility, and between *R* and the daily cost of no-shows. Costs and benefits multiply per provider.

Appointment Scheduling with Overbooking

If it were possible to reduce the no-show rate, then the cost of lost utility would be reduced. But no-shows occur for a variety of reasons. However, there are costs associated with finding out the specific reasons the patients of a particular clinic fail to show up and with attempting to reduce or eliminate the obstacles to appointment attendance (Bean and Talaga, 1995; Campbell et al., 2000; Garuda et al., 1998). These costs might exceed that of the lost utility, and clinical providers and administrators may consider overbooking to compensate for the utility loss caused by no-shows when they are not confident that they can change the no-show behavior of their patients (Barron, 1980; Shonick and Klein, 1977). Before proceeding to implement a policy of overbooking, clinical decision-makers need to consider total utility as the sum of the benefits and

the costs, so we determine the expected utility if overbooking is used to compensate for no-shows.

By assumption, the number of patients x who show up on any day varies according to a Bernoulli process, impacting the total capacity utilization of the clinic. The problem facing the clinic is to determine the number of appointments, $K \geq N$, to book. For a given show rate S, the expected number of patients served in a day, E[x], should be equal to clinic size N. Our assumption that x follows a binomial distribution, with each of K scheduled patients having the independent probability S of showing up, allows us to conclude that E[x] = KS. Then to achieve E[x] = N, we set $K = N/S$. Earlier, we defined C as total clinic time (allocated at target capacity N), so that for constant service duration D, $C = ND$. We also assumed that the K scheduled appointments are allocated evenly over the total clinic time C, so that the inter-appointment time interval is $T = ND/K$. Notice that with overbooking and this equal allocation of appointments, T is compressed by factor $N/K = S$, so that $T = ND / K = \frac{ND}{N / S} = DS$.

With overbooking, patient wait time occurs when more patients arrive than can be seen in any interval of time. When the patients scheduled at the end of the day show up, provider overtime occurs because D, the actual time needed to service each patient, is greater than the time allocated, T. As demonstrated in Table 1, patient wait time and provider overtime occur not only as a function of the total number of scheduled patients that show up during a clinic session but also as a function of these patients' arrival times. Therefore we must consider the varying possible sequences of patient shows and no-shows and their impact on patient wait time and overtime operation. In Table 1, we provide an example of each of the four possible combinations of wait time and overtime. There is a baseline case in which there are no added costs because there is neither patient wait time nor provider overtime; a case in which there is

patient wait time but no provider overtime; a case in which there is no patient wait time but there is provider overtime; and a case that has both patient wait time and provider overtime. These examples demonstrate the dynamics of arrival uncertainty and how it contributes to costs when overbooking is used.

In these examples, the clinic size $N = 5$, show rate $S = 0.5$, and service duration $D = 1$. Thus the number of appointments scheduled is $K = 5/0.5 = 10$, the regular time for the clinic session to end is $C = 5$, and the time between appointments, T, is compressed from 1 to 0.5 time units. To demonstrate how costs depend not only on x, the number of patients who arrive for each session, but also on the arrival times of these patients, we hold $x = 5$ fixed for each case. For the parameters specified above and from our earlier assumption that x is a binomially distributed random variable, the probability $p(x = 5) = \binom{10}{5} 0.5^5 (1 - 0.5)^5 = 0.2461$.

Table 1. Service times for patient show patterns.
O_i indicates the i^{th} patient who shows, (scheduled and arriving) in the timeslot shown.
The placement of D_i shows the timeslots in which the duration D of service time occurs for the i^{th} patient.

	Regular Time										**Overtime**				
Time Slot	1	2	3	4	5	6	7	8	9	10					
Start Time	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7
Case 0 Arrivals	O_1		O_2		O_3		O_4		O_5						
Service Durations	D_1		D_2		D_3		D_4		D_5						
Case 1 Arrivals	O_1	O_2	O_3		O_4		O_5								
Service Durations	D_1		D_2		D_3		D_4		D_5						
Case 2 Arrivals	O_1		O_2		O_3		O_4			O_5					
Service Durations	D_1		D_2		D_3		D_4			D_5 *Begin*	D_5 *End*				
Case 3 Arrivals	O_1		O_2				O_3	O_4	O_5						
Service Durations	D_1		D_2				D_3		D_4		D_5				

Case 0: Base case with no patient wait time and no provider overtime.

All patients who show up arrive when service is completed for the previous patient. Therefore there is no patient wait time and there is no provider overtime because the last arrival receives and completes service within the regular clinic session time $C = 5$. Notice that this arrival pattern is exactly what it would be if $S = 1$ and there were no overbooking because there would be no need for it. In this case, however, overbooking is used and this pattern is just one of $2^{10} = 1024$ possible arrival patterns that could occur by chance with probability 0.0010. This pattern results in no patient wait time and no provider overtime because each arrived appointment is followed, by chance, by a non-arrived appointment.

Case 1: Patient wait time, no provider overtime.

This case demonstrates what happens when more patients arrive than can be seen in any interval of time. There is patient wait time because the second arrival occurs before service is completed for the first arrival. The second patient waits T units of time. The third patient arrives as scheduled in the appointment immediately following the second, whose appointment started late, and waits $2T$ units of time. Even though non-arrived appointments separate the remaining patient arrivals, these patients also wait $2T$ units of time because the earlier arrivals delayed the availability of the provider to begin to service each of these patients. In this case, the delayed start of each patient's service does not result in late finish time for the clinic session. The non-arrived appointments after the fifth arrival allow the provider to catch up by end of the clinic session so there is no provider overtime.

Case 2: No patient wait time but there is provider overtime.

The non-arrived appointments following the first, second, third, and fourth arrived appointments result in no wait time for any of the five patients who arrive. But the fifth arrival occurs in the last appointment slot. Due to overbooking, the time allocated for each appointment is compressed from 1 to 0.5 time units, and therefore only half of the fifth arrival's service occurs before regular session end time C. To complete service for the last arrival, $T = 0.5$ units of overtime are used.

Case 3: There is both patient wait time and provider overtime.

The first three arrivals in this case follow a pattern similar to that of Case 0 and Case 2. Each of the first two arrivals is followed by at least one non-arrived appointment so that none of the first three arrived patients has wait time. But starting with the third arrival, three patients arrive sequentially. The lack of non-arrived appointments between arrived appointments results in wait

time for the fourth and fifth arrived patients. And, similar to the first three arrivals in Case1, the last three arrived appointment in this case require a total of $3D$ units of time to complete their service, but there are only $4T = 2D$ units remaining until the regular session end time C. Thus in addition to the wait time incurred by patients because the arrivals occur in an uninterrupted sequence, there is overtime because the sequence of arrivals occurs near the end of the clinic session.

Expected Net Utility from Overbooking

We used the cases above to show what causes overtime operation and patient wait time when overbooking is used. As these cases demonstrate, it is not only the number of patient arrivals but the order and times of these arrivals that impact costs. We held the number of arrivals fixed as $x = 5$ across the four cases. From Case 3, however, we make an extended observation. If the first two arrived appointments had instead been non-arrived, then the number of arrivals would decrease to $x = 3$ but patient wait time and overtime would be the same. Hence, we conclude that patient wait time and provider overtime are possible even when the number of arrivals is less than N, the clinic size or target number of patients to be seen. This means that overbooking to compensate for no-shows can result in instances in which costs increase because of wait time and overtime but there is no increase in productivity or revenue because, due to the uncertainty in patient behavior in showing up for their appointments, fewer patients show up than expected.

To calculate the expected *net* utility from overbooking, we must consider both the additional expected utility obtained by servicing additional patients and the expected costs of patient wait time and overtime created by overbooking. Let π = cost per hour of patient wait time, δ = cost per hour of clinic overtime operation, $W(p)$ = wait time of patient p,

F = the actual finish time of a clinic session, and C = the total clinic time allocated for a full capacity clinic of N patients. Considering both the benefit and the costs of overbooking, the expected net utility from overbooking is:

$$E[U_{Net}] = MS(K-N) - \pi E\left(\frac{\sum_p W(p)}{x}\right) - \delta E[(F-C)] \qquad (2)$$

The first term, $MS(K-N)$, is obtained as follows. As noted earlier for the baseline case of no overbooking, if we do not overbook and schedule only N appointments, the expected utility is MSN. If instead we overbook with a total of $K > N$ appointments, then the expected utility is MSK. Therefore, the additional expected utility obtained by servicing additional patients is $MS(K\text{-}N)$. The next term, $\pi E\left(\frac{\sum_p W(p)}{x}\right)$, is the expected cost of average patient wait time, which is the product of the unit cost of patient wait time and the expected value of the average patient wait time. Due to uncertainty in patient behavior in showing up for appointments, the derivation of patient wait time is complicated because it involves a large number of possible sequences of shows and no-shows. For details of the derivation, see Appendix 1.

The last term, $\delta E[(F-C)]$, is the expected cost of overtime, which is the product of the unit cost of overtime and the expected difference between actual finish time and the total clinic time allocated for a full capacity clinic of N patients.

Examples of Cost Calculations

We present the two small examples below, for the cases $K = 2$ and $K = 3$, to demonstrate how to obtain expected costs analytically and to show that the relationships among the parameters

change as *K*, the number of scheduled appointments, increases. We obtain $K = N/S$ by rounding up or down to the nearest integer.

Example 1: *K* = 2

Using the rule $K = N/S$ to determine how many appointments to schedule, we schedule $K = 2$ appointments when N =1 and $0.41 \le S \le 0.66$, or when N = 2 and $0.81 \le S \le 1$. The scheduling system for two appointments has the following four possible sequences of shows and no-shows:

[(1,1), (1,0), (0,1), (0,0)] with corresponding *x*=[2,1,1,0]. The accumulated wait time and finish time of the session, derived from the definitions and relationships in Appendix 1, are shown for each possible sequence in Table 2.

Table 2. All possible show and no-show sequences for *K*=2 → $\{(N=1)\ AND\ (0.41 \le S \le 0.66)\}OR\ \{(N=2)\ AND (0.81 \le S \le 1)\}$

Show and No-Show sequence	Probability	Accumulated wait time	Average Wait time per patient	Finish time of session
1,1	S^2	D-T	(D-T)/2	2D
1,0	S(1-S)	0	0	D
0,1	(1-S)S	0	0	T+D
0,0	(1-S)^2	0	0	0

Note that for the special case S =1, the only possible show and no-show sequence is the first row, and then, because $T = SD = D$, there is no patient wait time. For any value of *D* and for values of *S* in the ranges above and resulting values of *T*, we can calculate expected values for total patient wait time, average patient wait time, finish time of the clinic session, and costs of patient wait time and overtime as follow.

$$E\left[\sum_p W(p)\right] = S^2(D-T) + 0\left[S(1-S) + (1-S)S + (1-S)^2\right] = S^2(D-T) \quad (9)$$

$$E\left[\overline{W}\right] = S^2(D-T)/2 + 0\left[S(1-S) + (1-S)S + (1-S)^2\right] \\ = S^2(D-T)/2 \tag{10}$$

$$E[F] = S^2 2D + S(1-S)D + S(1-S)(T+D) + \left[(1-S^2)\times 0\right] \\ = S(T - ST + 2D) = S[T(1-S) + 2D] \tag{11}$$

When we overbook by scheduling K patients ($K > N$), we are scheduling K-N more patients than we would have scheduled if we did not overbook. The expected number that shows up is $S(K\text{-}N)$, and each that shows up contributes M units of utility. Thus the expected added utility from overbooking is expressed as $MS(K\text{-}N)$. Unlike previous research that considered only the benefits of overbooking, we also consider the costs of patient wait time and provider overtime to determine U_{Net}, the expected net utility. Using the expected costs we calculated above in (10) and (11) and the relationship $T = SD$, we express the net utility for $K = 2$ as:

$$E[U_{Net}] = MS(K-N) - \pi S^2(D-T)/2 - \delta\left[S\left(T - ST + 2D\right) - C\right] \\ = MS(K-N) - \pi S^2 D(1-S)/2 - \delta\left[SD\left(2-S\right)(S+1) - C\right] \tag{12}$$

Then it is advantageous to overbook by scheduling (K-N) extra patients only if the expected net utility is greater than zero. This means that the expected benefit from seeing additional patients must exceed the sum of the expected costs. Thus from (12)

$$MS(K - N) > \pi S^2(D - T)/2 + \delta\left[S\left(T - ST + 2D\right) - C\right] \\ MS(K - N) > \pi S^2 D(1-S)/2 + \delta\left[SD\left(2-S\right)(S+1) - C\right] \tag{13}$$

Example 2: $K = 3$

The expected costs, expressed above as functions of S, D, and T, change when K is increased to schedule more patients because new sequences are included that result in more complicated expressions of wait time and finish time. We see this in Table 3 with the calculations for $K = 3$,

which is the appropriate number of appointments to schedule if any one of the following three scenarios applies:

1) $\{(N=1)\ AND(0.29 \le S \le 0.4)\}$, or

2) $\{(N=2)\ AND(0.58 \le S \le 0.8)\}$, or

3) $\{(N=3)\ AND(0.84 \le S \le 1)\}$.

Table 3. All possible show and no-show sequences for K=3 → $\{(N=1)\ AND(0.29 \le S \le 0.4)\}$ or $\{(N=2)\ AND(0.58 \le S \le 0.8)\}$, or $\{(N=3)\ AND(0.84 \le S \le 1)\}$.

Sequence #	Show and No-Show sequence	Probability	Accumulated wait time	Average Wait time per patient	Finish time of session
1	1,1,1	S^3	3(D-T)	(D-T)	3D
2	1,1,0	S^2(1-S)	D-T	(D-T)/2	2D
3	1,0,1	S^2(1-S)	Max{0,D-2T}	Max{0,D-2T}/2	Max{2T,D}+D
4	1,0,0	S(1-S)^2	0	0	D
5	0,1,1	S^2(1-S)	D-T	(D-T)/2	T+2D
6	0,1,0	S(1-S)^2	0	0	T+D
7	0,0,1	S(1-S)^2	0	0	2T+D
8	0,0,0	(1-S)^3	0	0	0

For non-restricted values of S, expected values are calculated and expressed as:

$$E\left[\sum_p W(p)\right] = 3S^3(D-T) + S^2(1-S)[2(D-T) + Max\{0, D-2T\}]$$
$$= S^2[(S+2)(D-T) + (1-S)Max\{0, D-2T\}] \quad (14)$$

$$E[\overline{W}] = S^3(D-T) + S^2(1-S)[(D-T)/2 + \frac{Max\{0, D-2T\}}{2} + (D-T)/2]$$
$$= S^2\left[(D-T) + (1-S)\frac{Max\{0, D-2T\}}{2}\right] \quad (15)$$

$$E[F] = 3S^3D + S^2(1-S)(5D + Max\{2T, D\} + T) + 3S(1-S)^2(T+D) \quad (16)$$

Then for K=3, it is advantageous to overbook (K-N) patients only if

$$MS(K-N) > \pi S^2\left[(D-T) + (1-S)\frac{Max\{0, D-2T\}}{2}\right] + \delta[3S^3D + S^2(1-S)(5D + Max\{2T, D\} + T) + 3S(1-S)^2(T+D) - C] \quad (17)$$

It is apparent that the first, second, and fifth sequences incur wait time because *T*, the scheduled time between appointments, is compressed and these sequences contain more than one patient arrival in a row. In these sequences, we can calculate the unique value of the wait times for each patient because we know that each patient that directly follows another patient cannot start until the end time of service for that previous patient. However, in the third sequence, (1,0,1), the third patient, whose scheduled arrival time is $(3\text{-}1)T = 2T$, incurs wait time only if the service time for the first patient ends later than the arrival time of the third patient, i.e., only if $D > 2T$. Therefore for the third sequence, the wait times and finish time of the session contain "Max" functions because patients who show up begin to receive service at the later of their arrival time and the end time of service for the previous patient. We know that $T = SD$ for all values of *S* and *D*, but if we haven't specified the value of *S*, we cannot conclude whether $D\text{-}2T > 0$. This situation did not arise for the earlier case $K = 2$, and it occurs for only one sequence for the general case $K = 3$, but for cases $K > 3$, there are more sequences containing maximum functions in the calculation of wait time and finish time. For example, for $K = 4$, there are arrival sequences in which there are no-shows between shows, such as (1,1,0,1), (0,1,0,1), (1,0,1,0), (1,0,1,1), (1,0,0,1).

It is not possible to specify a closed-form analytical model in terms of the parameters *S* and *D* that is valid for all possible values of the show rate *S*. However, when we specify a range of values for *S*, we can simplify the equations above by re-expressing patient wait time and session finish time as precise equalities.

Notice that for $S < 1/2$, $Max\{0, D - 2T\} = Max\{0, D - 2SD\} = D - 2SD = D(1 - 2S)$, and

$Max\{2T, D\} + D = Max\{2SD + D, 2D\} = 2D$,

but for $S \geq 1/2$, $Max\{0, D - 2T\} = Max\{0, D - 2SD\} = 0$ and

$Max\{2T, D\} + D = Max\{2SD + D, 2D\} = 2SD + D = D(2S + 1)$.

Thus we can express wait time and finish time for Sequence #3, with arrivals (1,0,1), as shown in Table 4.

Table 4: Wait time and finish time for arrival sequence (1,0,1)

S	Probability	Accumulated wait time	Average Wait time per patient	Finish time of session
$S < 1/2$	$S^2(1\text{-}S)$	$D\text{-}2T = D(1\text{-}2S)$ or	$(D\text{-}2T)/2 = D(1\text{-}2S)/2$	$2D$
$S \geq 1/2$	$S^2(1\text{-}S)$	0	0	$2T + D = D(2S + 1)$

For $S < 1/2$, equations (14) – (17) become:

$$E\left[\sum_p W(p)\right] = S^2[(S + 2)(D - T) + (1 - S)(D - 2T)] = S^2 D(S - 1)(S - 3) \quad (14\text{ a})$$

$$E[\overline{W}] = S^2\left[(D - T) + (1 - S)\frac{Max\{0, D - 2T\}}{2}\right] = S^2 D(1 - S)\left(\tfrac{3}{2} - S\right) \quad (15\text{ a})$$

$$E[F] = 3S^3 D + S^2(1 - S)(5D + Max\{2T, D\} + T) + 3S(1 - S)^2(T + D) = 6D^3 - 7S^3 D + S^4 D - 3SD \quad (16\text{ a})$$

Then for K=3 and $S < 1/2$, it is advantageous to overbook (K-N) patients only if

$$MS(K - N) > \pi S^2 D(1 - S)\left(\tfrac{3}{2} - S\right) + \delta[(6D^3 - 7S^3 D + S^4 D - 3SD) - C] \quad (17\text{ a})$$

For $S \geq 1/2$, expected values are calculated and expressed as:

$$E\left[\sum_p W(p)\right] = 3S^3(D - T) + S^2(1 - S)[2(D - T) + 0] = S^2 D(2 - S - S^2) \quad (14b)$$

$$E\left[\overline{W}\right] = S^3(D-T) + S^2(1-S)[(D-T)/2 + \frac{0}{2} + (D-T)/2] \\ = S^2D - S^3D = S^2D(1-S) \tag{15b}$$

$$E[F] = 3S^3D + S^2(1-S)(5D+2T+T) + 3S(1-S)^2(T+D) \\ = SD(3+5S-2S^2) \tag{16b}$$

Then for K=3 and $S \geq 1/2$, it is advantageous to overbook (K-N) patients only if

$$MS(K-N) > \pi(S^2D(1-S)) + \\ + \delta[SD(3+5S-2S^2) - C] \tag{17b}$$

Comparison of net benefit to expected costs for varying *K* and *S*

As demonstrated in our examples of two values of K, to determine when the benefits of overbooking exceed the costs, we must re-express the cost functions for each different value of K, and for $K > 2$, for specific ranges of S. Overbooking is recommended when $E[U_{Net}] > 0$, which happens when the expected marginal benefit, $MS(K-N)$, exceeds the expected total cost of average patient wait time and provider overtime. Table 5 shows how the expression of the expected total cost changes for our examples. This emphasizes why a general recommendation to overbook in any clinic that has no-shows can be dangerous. Determining whether overbooking is an effective policy requires accurate information about the no-show rate and the clinic size.

Table 5: Expected total costs for a varying number of scheduled appointments K and no-show rate S.

K	Valid ranges for S	Expected total cost of average patient wait time and provider overtime
2	All values $S \le 1$	$\pi S^2 D(1-S)/2 \quad + \delta SD(2-S)(S+1) \quad - \delta C$
3	$S < 1/2$	$\pi S^2 D(1-S)\left(3/2 - S\right) + \delta SD(S^3 - 7S^2 - 3) + 6D^3 - \delta C$
3	$1/2 \le S \le 1$	$\pi S^2 D(1-S) \quad + \delta SD(2S+1)(S-3) \quad - \delta C$

Simulation Results for Overbooking Model

We developed analytical expressions for the expected costs associated with scheduling up to K = 3 patients and demonstrated how these costs must be recalculated as K increases. The number of possible sequences of shows and no-shows grows exponentially so that for a schedule of K patients this number is 2^K. For example, for K=20, the number of possible sequences is 2^{20}=1,048,576 and for K=30, the number of possible sequences is 2^{30}= 1,073,741,824. Thus for realistic problems and large clinics, the number of calculations required to solve the problem analytically makes real-time solution impractical. For this reason, for cases in which K > 3, we use simulation to analyze the results for varying levels of K and S.

To determine the expected net utility obtained from overbooking to compensate for no-shows, we conducted simulation experiments with five realistic sizes of N = {10, 20, 30, 40, 50}, and ten levels of S = {100%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%}. We estimated a realistic range for N as 10-50 because the number of patients seen per day by a clinician who sees patients as frequently as every ten minutes in an eight-hour work day is usually less than or equal to 48, but in a clinic we observed with longer service times, part-time providers saw as few as 10. This range is consistent with other research such as Ho and Lau's (1992) model with N at 10, 20, and 30, and Vissers' (1979) model with six levels of N in the range of 10-60 and fixed

no-show rate of 10%. We included a wide range of no-show rates because a variety of studies report widely varying rates. For example, Rust et al. (1995) reported no-show rates ranging from 3-80% in a study of 200 public pediatric clinics. [To do: Check article to see how many appointments were scheduled.] K was calculated (N/S) rounded to the nearest integer. For each experiment, 10,000 replications were made. Pilot simulations indicate that the half-width of the 95% confidence intervals were less than or equal to 2% of the point estimate of simulated values of session finish time, F, and patient wait time (total and average $W(p)$).

Results of simulation experiments are shown in Figures 1-5 and 8-10. First we consider separately each of the individual performance measures of average patient wait time and overtime operation. For all five levels of clinic size, both average patient wait time and overtime operation increase as the no-show rate increases. This is as expected because as the no-show rate increases, then the number of patients scheduled increases so that the expected number of patients that show up equals the clinic size. There is a probability that the number of patients that show up is greater than the clinic size and when this happens, there is *always* positive patient wait time and overtime operation because more patients must be seen than the number of patients that fits into the clinic size.

Wait Time and Overtime

As shown in Figures 1 and 2, as no-show rates increase in an overbooked system, patient wait time and overtime operation increase. As clinic size, N, increases, both measures are larger and increase at a higher rate as the no-show rate increases. In other words, with overbooking, as the clinic size, N, increases, the no-show rate and increases in it have larger effects. This is because the larger the no-show rate, the more compressed the inter-appointment time intervals must be in order to fit the proportionately higher number of total appointments scheduled. For

larger clinics, the possible number of consecutive patient shows is larger, which leads to higher probabilities that wait time occurs and higher values when it does. Patients who arrive later in a series of consecutive arrivals incur more wait time. As start times become later, appointment finish times are later and the entire clinic is more likely to end with a large amount of overtime.

Figure 1. Simulation Results for Average Patient Wait Time.

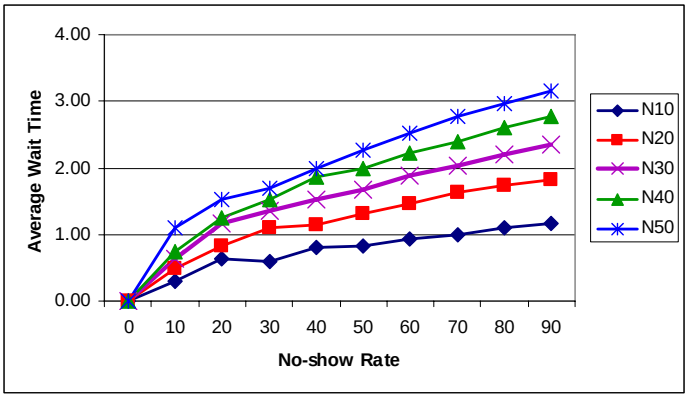
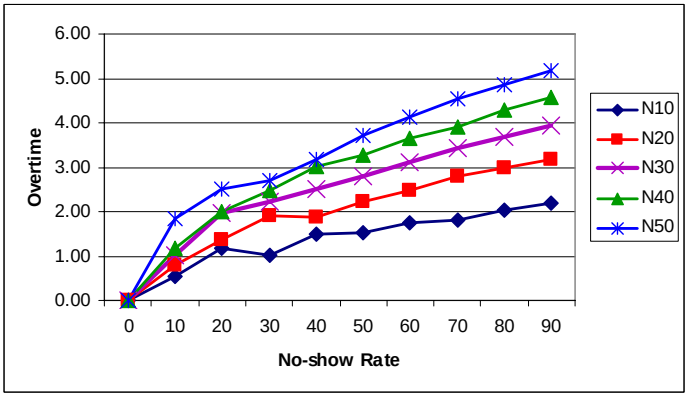


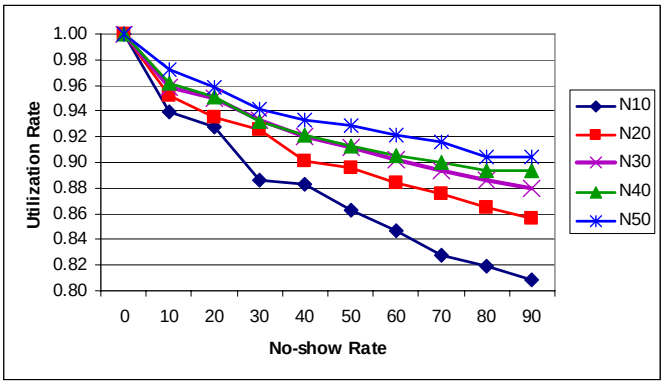
Figure 2. Simulation Results for Overtime.



Provider Productivity

Although productivity is not directly included in the model we defined of benefits and costs, many health care providers consider it to be an important measure of provider performance (Baum, 2001; Chesanow, 1996; Cole, 2003; Chung, 2002; McCarthy, 2002; Tonges, M.C., 1985). In Figure 3 we see that the increasing overtime resulting from overbooking is detrimental to the provider's utilization per unit of time, a measure of productivity that divides the total time a provider is busy delivering service by the total length of the provider's work day.

Figure 3. Utilization (Productivity)

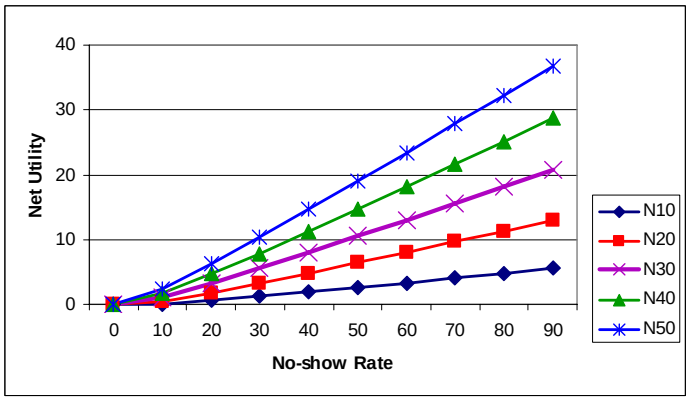


This is because the provider's workday is lengthened due to overtime, but because of the positive no-show rate, he/she does not serve, on average, additional patients per time unit worked. With overbooking, however, as the clinic size increases, productivity is higher for all no-show rates and decreases less as the no-show rate increases. This suggests that if the performance focus is on provider productivity, overbooking is more effective in larger clinics than in smaller ones.

Net Utility

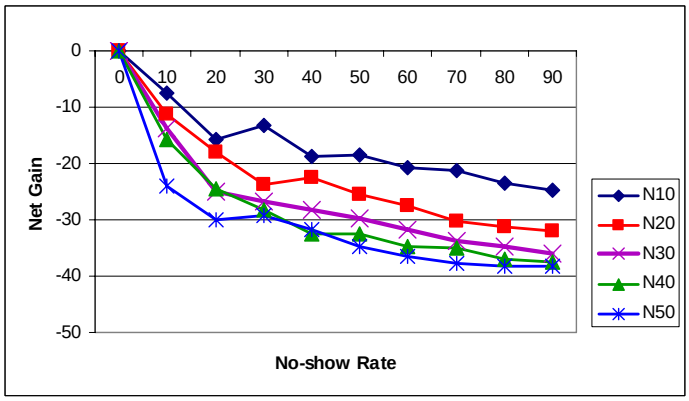
Next we use our simulation results in our net utility model to show the net effects of overbooking, calculated as a weighted sum of utility benefits and costs. If each component had equal value or weight, we see in Figure 4 that net utility in an overbooked system increases as the no-show rate increases and the rate of increase is steeper for increasing clinic sizes.

Figure 4. $U_{Net} = MS(K-N) - \pi\overline{W} - \delta(F-C)$
$M = \pi = \delta = 1$



But if the costs of patient wait time and overtime are both 10 times greater than the marginal benefit of each additional patient, then we reach a different conclusion. As shown in Figure 5, net utility from overbooking is increasingly negative as the no-show rate increases and as the clinic size increases.

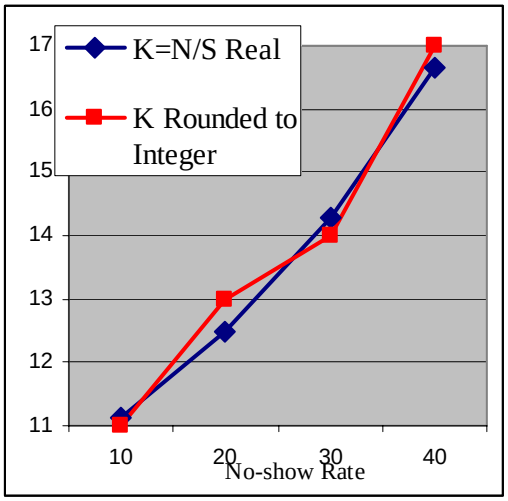
Figure 5. $U_{Net} = MS(K-N) - \pi\overline{W} - \delta(F-C)$
$M = 1;\ \pi = \delta = 10$



In Figure 5, for no-show rate R=0.30, there is a noticeable spike in Net Gain for N=10 and a dip for N=20. This happens because in setting the number of scheduled appointments as $K=N/S$, K is rounded to the integer value nearest to the real value of N/S. As Figure 6 shows for N=10, the rounded integers values of K for the no-show rates surrounding R=0.30 are larger than the real

values N/S, but for R=0.30 the rounded integer value is smaller. Hence, the spike in the resulting Net Gain is attributable to a downward shift in the input value of K.

Figure 6. Rounded values K for N=10



A similar phenomenon is apparent in Figure 7. For N=20, the rounded integers values of K for the no-show rates surrounding the rate R=0.30 are smaller than the real values N/S, but for R=0.30 the rounded integer value is larger. Hence, the dip in the resulting Net Gain is attributable to an upward shift in the input value of K.

Figure 7. Rounded values K for N=20

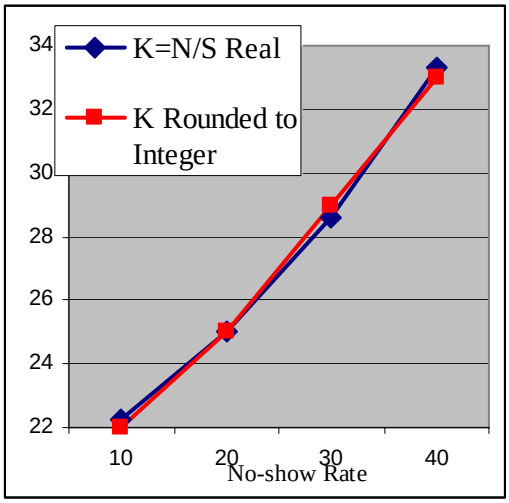


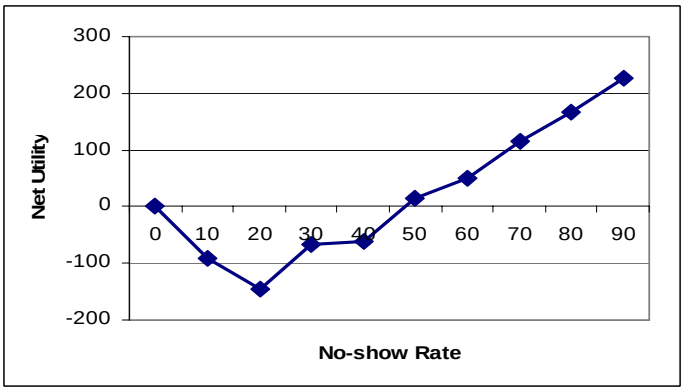
Figure 8 shows mixed weights for a hypothetical clinic that receives payment M = \$110 per patient, is penalized by its contracted payers at the rate π = \$500 per hour of average patient

wait time, and has overtime costs δ = \$80 per hour. For a clinic size of 10, net utility is increasingly negative as the no-show rate increases from 0 to 10% to 20%, then increases with increased no-show rates and becomes positive at a no-show rate of 50%. The sharp dip in Net Utility at no-show rate R=20 is influenced by the steep rise in rounded K that is shown in Figure 6 as R increases from 10% to 20%. The negative net utility that prevails until R = 50% indicates that until this point, the costs outweigh the benefits of overbooking. This is because until no-show rates become sufficiently high, there is a high probability of patients showing up in a series that is uninterrupted by no-shows. This causes patient wait time, which has the heaviest weight of the components of net utility, to become large. But at higher no-show rates, wait time becomes less likely and has smaller values, and unit cost of overtime is less than the marginal revenue from additional patients that are seen.

Figure 8. Mixed weights in a small clinic.

$M = 110;\ \pi = 500; \delta = 80.\ N = 10$

$$U_{Net} = \left[\$110 \times S(K - N)\right] - 500\overline{W} - 80\left(F - C\right).$$



For the larger clinic size of 50, shown in Figure 9, net utility is negative only for no-show rates between 0 and approximately 15%. Then net utility becomes positive and steeply increases for increasing no-show rates.

Figure 9.
Mixed weights in a large clinic.
$M = 110;\ \pi = 500; \delta = 80.\ N = 50.$
$U_{Net} = [\$110 \times S(K - N)] - 500\overline{W} - 80(F - C).$

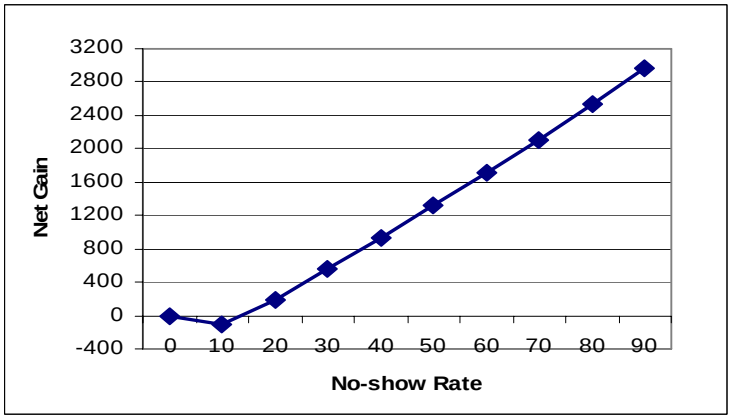
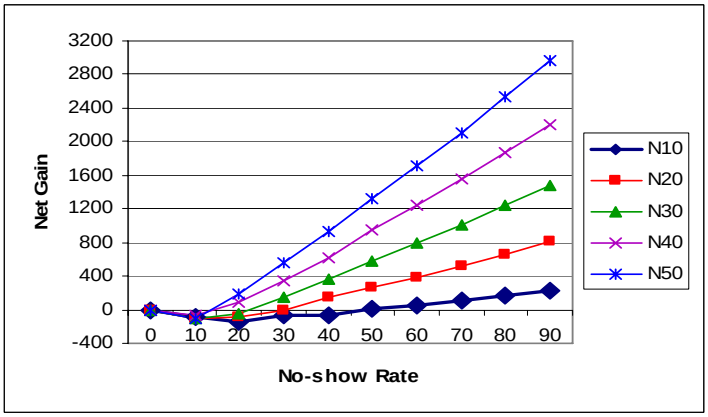


Figure 10 shows how net utility varies across no-show rates for varying clinic sizes. As the clinic size increases, net utility increases more rapidly across increasing no-show rates.

Figure 10. Mixed weights in varying clinic sizes, $N = \{10, 20, 30, 40, 50\}$.
$U_{Net} = [\$110 \times S(K - N)] - 500\overline{W} - 80(F - C).$



Discussion and Conclusions

Patient no-shows lead to loss of revenue or utility. When clinical providers and administrators can estimate the marginal revenue or utility per patient seen and know the average daily no-show rate, then the cost of no-shows can be determined. This provides a useful starting point for determining whether to invest in reducing the no-show rate and/or to consider

overbooking. Although earlier research in clinical appointment scheduling has suggested overbooking as the solution to no-shows, we have shown that overbooking may actually be more costly than beneficial when the costs of patient wait time and clinic overtime operation are included in the utility measure.

Our research demonstrates that for a realistic range of clinic sizes, net utility across no-show rates depends upon the relative weightings of the marginal value of each additional patient and the relative costs of patient wait time and provider overtime. For a fixed clinic size, net utility across no-show rates may be strictly positive, strictly negative, or mixed. In addition to the relative weightings of benefit and costs, net utility depends also upon clinic size. Therefore it would be incorrect to conclude that all clinics with positive no-show rates should overbook.

Decreasing no-show rates decrease uncertainty in patient behavior, but a surprising result of this research is that increased no-show rates may correspond to *higher* net utility. We interpret this counter-intuitive result by considering system behavior as a function of no-shows and the policy chosen for dealing with them. At higher show rates (lower no-show rates), the probability of patients showing up is higher. Thus if overbooking is employed, the risk of overtime and patient wait time is higher, incurring higher costs that may exceed the expected benefits from seeing more patients. Therefore, for sufficiently high costs of patient wait time and overtime, there is more benefit in overbooking when no-show rates are higher and more harm in overbooking when no-show rates are lower.

We continue to explore the operational implications and performance factors important to clinical providers in their appointment scheduling. Further application of probability models and revenue management techniques from the airline industry appear promising for helping clinics to choose appropriate levels of overbooking and to analyze the impacts on the level of service to

patients. Further application to wave schedules of the dynamics we revealed here of patient wait time and overtime may lead to improved schedule performance for practitioners.

References

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.

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

Barron, W.M., 1980. Failed appointments: Who misses them, why they are missed, and what can be done. Primary Care 7(4), 563-574.

Bean, A.B. and Talaga, J., 1995. Predicting appointment breaking. Journal of Health Care Marketing 15(1), 29-34.

Blanco White, M.J. and Pike, M.C., 1964. Appointment systems in out-patients' clinics and effect of patients' unpunctuality. Medical Care 2, 133-144.

Brahimi, M. and Worthington, D.J., 1991. Queueing models for out-patient appointment systems – a case study. Journal of the Operational Research Society 42(9), 733-746.

Campbell, J.D., Chez, R.A., Queen, T., Barcelo, A., Patron, E., 2000. The no-show rate in a high-risk obstetric clinic. Journal of Women's Health & Gender-based Medicine 9(8), 891-895.

Cayirli, T. and Veral, E., 2003. Outpatient scheduling in health care: A review of literature. Production and Operations Management 12(4), 519-549.

Chesanow, N., 1996. Can't stay on schedule? Here's a solution. Medical Economics 73(21), 174-180.

Cole, R.M., 2003. Four steps to a streamlined schedule: Analyzing your schedule in minute detail can help you gain control and keep your practice at peak performance. Review of Optometry 140(6), 33-35.

Chung, M.K., 2002. Tuning up your patient schedule. Family Practice Management 9(1), 41-48.

Fetter, R.B. and Thompson, J.D., 1966. Patients' waiting time and doctors' idle time in the outpatient setting. Health Services Research 1, 66-90.

Garuda, S.R., Javalgi, R.G., Talluri, V.S., 1998. Tackling no-show behavior: A market-driven approach. Health Marketing Quarterly 5(4), 25-45.

Hillier, F.S. and Lierbermain, G.J. (2001). Introduction to Operations Research, seventh edition. New York: McGraw-Hill.

Ho, C. and Lau, H., 1992. Minimizing total cost in scheduling outpatient appointments. Management Science 38(12), 1750-1763.

Klassen, K.J. and Rohleder, T.R., 1996. Scheduling outpatient appointments in a dynamic environment. Journal of Operations Management 14, 83-101.

Larkin, H., 1999. Satisfaction pays. American Medical News 42(30), 17.

McCarthy, E.L, 2002. Physician office productivity improvement through operations analysis and process redesign. The Journal of Ambulatory Care Management 25(4) 37-52.

McCarthy, K., McGee, H.M., and O'Boyle, C.A., 2000. Outpatient clinic waiting times and non-attendance as indicators of quality. Psychology, Health & Medicine 5(3), 287-293.

McCord, R., 1996. "Waitless" waiting room: The ultimate practice builder. Physician's Management 36(10, 71-76.

Metter, R. and Vargas, V., 1999. Yield management for the nonprofit sector. Journal of Service Research 1(3), 215-226.

Murray, M. and Berwick, D.M., 2003. Advanced access: reducing waiting and delays in primary care. Journal of the American Medical Association 289(8), 1035-1040

Rohleder, T.R. and Klassen, K.J., 2002. Rolling horizon appointment scheduling: A simulation study. Health Care Management Science 5, 201-209.

Rothstein, M., 1971. An airline overbooking model. Transportation Science 5(2), 180-192.

Rust, C.T., Gallups, N.H., Clark, W.S., Jones, D.S., Wilcox, W.D; 1995. Patient appointment failures in pediatric resident continuity clinics. Archives of Pediatrics & Adolescent Medicine 149(6), 693-695.

Schroer, B.J. and Smith, H.T., 1977. Effective patient scheduling. The Journal of Family Practice 5(3), 407-411.

Sharp, D.J. and Hamilton, W., 2001. Non-attendance at general practices and outpatient clinics: local systems are needed to address local problems. British Medical Journal 323(7321), 1081-1082.

Shonick, W. and Klein, B.W., 1977. An approach to reducing the adverse effects of broken appointments in primary care systems. Medical Care 15(5), 419-429.

Silver, M., 1975. Scheduling: Least developed art. Family Practice News 5(23), 34.

Smith, B.C., Leimkuhler, J.T., and Darrow, R.M., 1992. Yield management at American Airlines. Interfaces 22(1), 8-31.

Sweeney, D.R., 1996. Your office: A lot of things will have to change. Medical Economics 73(7), 97-102.

Thomas, S., Glynne-Jones, R., and Chait, I., 1997. Is it worth the wait? A survey of patients' satisfaction with an oncology outpatient clinic. European Journal of Cancer Care 6(1), 50-58.

Tonges, M.C., 1985. Quality with economy: Doing the right things for less. Nursing Eonomics 3, 205-211.

Study: Hospitals can increase revenue by proactive revenue-cycle management. Managed Care Weekly Digest, February 3, 2003, 33-34.

Van Ryzin, G.J. and Talluri, K.T., 2003. Revenue Management. In R.W. Hall (ed.), Handbook of Transportation Science (pp. 599-659). Boston: Kluwer Academic Publishers.

Vissers, J., 1979. Selecting a suitable appointment system in an outpatient setting. Medical Care 17(12), 1207-1220.

Vissers, J. and Wijngaard, J., 1979. The outpatient appointment system: Design of a simulation study. European Journal of Operational Research, 3(6), 459-463.

Warden, J., 1995. 4.5 Million outpatients miss appointments. British Medical Journal 310, 6 May 1995, 1158.

Welch, J.D. and Bailey, N.T., 1952. Appointment systems in hospital outpatient departments. The Lancet, May 31, 1952, 1105-1108.

Zeff, P., 2002. Delayed reaction. American Medical News 38(8), 14-16.

Appendix 1: Derivation of Patient Wait Time and Session Finish Time

To analytically derive expected patient wait time, let y_p represent the behavior of patient *p* such that

$$y_p = \begin{cases} 1 \; if \; the \; p^{th} \; scheduled \; patient \; shows \; up \\ 0 \; if \; the \; p^{th} \; scheduled \; patient \; does \; not \; show \; up \end{cases} \quad (4)$$

Then a sequence of the shows and no-shows of K patients can be expressed as a vector $(y_p, y_{p+1}, \ldots y_K)$. The total number of scheduled patients who show up is

$$x = \sum_{p=1}^{K} y_p \quad . \tag{5}$$

As shown in (4), there are two possible outcomes for each scheduled patient. Thus for K scheduled patients, the number of different possible sequences of patient shows and no-shows is 2^K. Each different possible sequence is numbered with index $q = \{1,2,\ldots 2^K\}$. For example, for the case $K = 3$, there are $2^3 = 8$ different sequences, expressed as the eight vectors of y_p as $\{(0,0,0),(0,0,1),(0,1,0),(1,0,0),(0,1,1),(1,0,1),(1,1,0),(1,1,1)\}$. As we observed for the three cases shown in Table 1, patient wait time depends on the entire vector of y_p, not just on the total $x = \sum_{p=1}^{K} y_p$. Referring to the eight vectors above, there are three vectors, i.e., for $q = \{5,6,7\}$, in which $x_q = 2$, but because there is one no-show separating the two shows in the 6th sequence (1,0,1), the average patient wait time is different from the average patient wait time in the 5th and 7th sequences, (0,1,1) and (1,1,0). Thus we must consider the average patient wait time specifically for each of the q sequences. To obtain the expected average wait time, we multiply the average wait time of each sequence by the probability of each sequence's occurrence, which is the probability that K appointments result in x_q patients who show up and $K - x_q$ patients who do not show up. Then the expected value of average patient wait time is

$$E\left(\frac{\sum_{p} W(p)}{x}\right) = \sum_{q=1}^{2^K}\left(S^{x_q}(1-S)^{K-x_q} \times \left(\frac{\sum_{p_q} W(p_q)}{x_q}\right)\right) \tag{6}$$

The left-hand side is the expected value of the sum, over all *p* patients, of patient wait time, divided by the total number of patients who show up. The right-hand side is the sum, over all 2^K different possible sequences, of the probability of each sequence occurring, $S^{x_q}(1-S)^{K-x_q}$, multiplied by the average patient wait time that occurs for that sequence, $\left(\frac{\sum_{p_q} W(p_q)}{x_q} \right)$, obtained by summing wait time over all scheduled patients in each sequence and dividing by the total number of patients who show in that sequence.

Next, we express the wait time for any patient within a sequence. We start with the definitions of Ho and Lau (1992) for patient *p* and let *A(p)* = scheduled arrival time, *b(p)* = start time of service, and *f(p)* = end time of service. Let $A(1) = b(1) = 0$. We hold the service time *D* constant, and we assume that when patients do show up, they are punctual, which means they are neither late nor early (Blanco White and Pike, 1964). Recall that $y_p = 1$ only if patient *p* shows up; otherwise, $y_p = 0$. We model actual patient arrival time as $y_p A(p)$ and actual service duration as $y_p D$ so that the actual arrival time and actual service duration of patients who do not show up is set to 0 and therefore there is no addition to wait time for subsequent patients or to the finish time of the clinic session. We define $b(p) = \max\{y_p A(p), f(p-1)\}$ to represent the actual begin time of an appointment as occurring no earlier than the actual finish time of the previous appointment. The relationship $f(p) = b(p) + y_p D$ indicates equality between the beginning and end time of service for patients who don't show up – i.e., no actual operating time is accrued in the clinic schedule for patients who do not show up. Then the wait time of any patient within a

sequence is:

$$W(p) = Max\{y_p A(p) - f(p-1), 0\} = Max\{y_p (p-1)T - (b(p-1) + y_{p-1} D), 0\} \quad (7)$$

The finish time for a session is the end time of the last arrival,

$$F = f[Max\{p \ni y_p = 1\}]. \quad (8)$$
