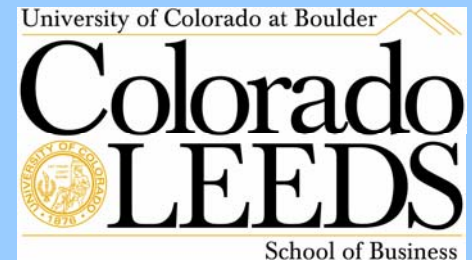


Overbooking of Clinical Appointment Schedules with Wave Heuristics



Linda LaGanga, Ph.D., LPC
Director of Quality Systems
Mental Health Center of Denver
Linda.Laganga@mhcd.org

Stephen Lawrence, Ph.D.
University of Colorado-Boulder
Stephen.Lawrence@colorado.edu



37th Annual Meeting of the Decision Sciences Institute
November 18-21, 2006
San Antonio, Texas

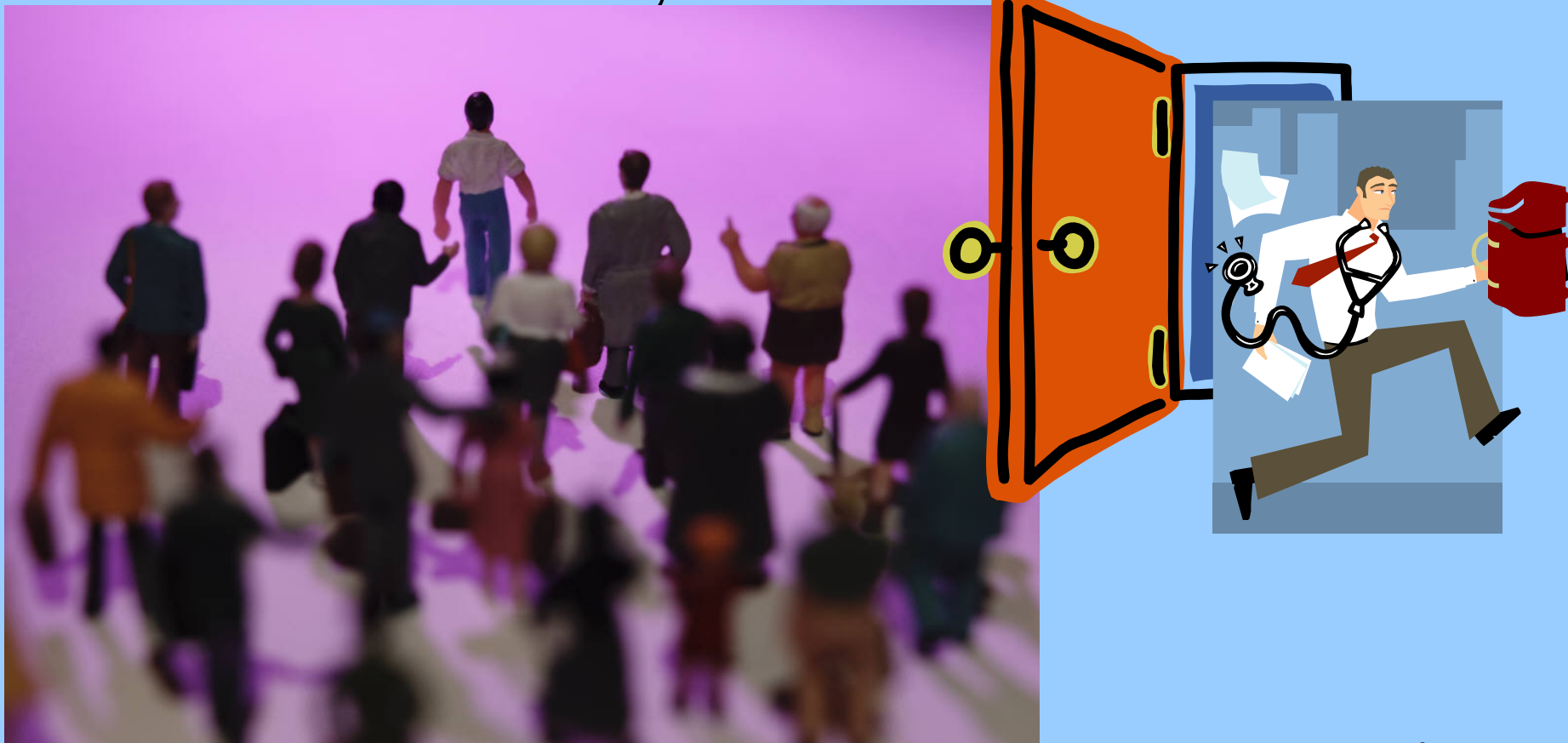
Abstract

We concentrate on no-shows and analyze the performance of scheduling rules to determine their effectiveness when overbooking is used to compensate for the lost productivity of no-shows.

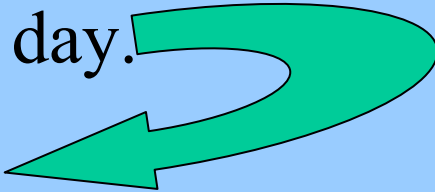
We adjust traditional appointment scheduling performance measures to capture the operating dynamics of overbooked appointment scheduling systems, and develop heuristics to improve performance and decision-making under uncertainty.

Overbooking is used in many clinics...

- often ineffectively!



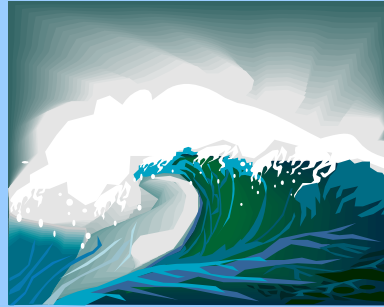
Why Overbook?

- To increase provider capacity and patient access to services
- $\frac{100\%}{70\%} = 1.43 = 43\%$ increase in capacity.
- In Denver we bring in 1 homeless person off the streets every day.
- 157 additional people served per year with overbooking.

Where to put overbooked Appointments: What is a Wave?

- A clinical appointment schedule has been described as, “a WAVE – people come in and keep coming and continue to come in! The idea is to modify that wave so as to have peaks and valleys in the schedule, allowing you to finish the day approximately on time” (Silver, 1975).

Why Wave Scheduling?



- “Wave scheduling” recommended by practitioners
- Anecdotal reports of success but little systematic study
- Provides new scheduling options
- Leaves room to catch up after a backlog
- For a fixed number of appointments, where to place them in the schedule
- Varies spacing and number of appointments booked

Model Assumptions

- Each patient is assigned to a specific provider, and providers do not service each other's patients. Thus, each provider is modeled as a single-server system with fixed appointment times and constant service time.
- Patients who show up are punctual.

Model Assumptions

- Constant service duration D
- Clinic capacity N constant as the total number of patients that can be serviced within the normal operating time of a clinic session without overtime.
- For show rate S , we set $K = N/S$ rounded to the nearest integer so that the expected number of patients serviced in a clinic session is $E(X) = SK = N$ (or is very close to N .)

Model Parameters

- A morning clinic session runs at regular capacity (without overtime) for four hours from 8:00 a.m. – 12 p.m.
- The service time for each patient is 20 minutes, constant time, and therefore, the clinic size is 12.

Why Wave Heuristics?

Show Rate	Appointments	Possible Schedules
1	12	21,090,682,613
0.9	13	73,006,209,045
0.8	15	751,616,304,549
0.7	17	6,499,270,398,159
0.6	20	125,994,627,894,135
0.5	24	4,355,031,703,297,280
0.4	30	450,883,717,216,035,000
0.3	40	285,219,402,396,401,000,000
0.2	60	5,680,916,595,331,740,000,000,000
0.1	120	744,986,412,434,507,000,000,000,000,000,000,000

- To identify a focused set of schedules for which we can evaluate and compare performance and choose among several alternatives that perform well.

Two-stage Modeling

- *Planning* models are based on the assumption that all scheduled patients show up.
 - Spreadsheet calculations of Wait Time and Overtime
 - Identify and alleviate scheduling congestion
 - Heuristic predictor of overtime

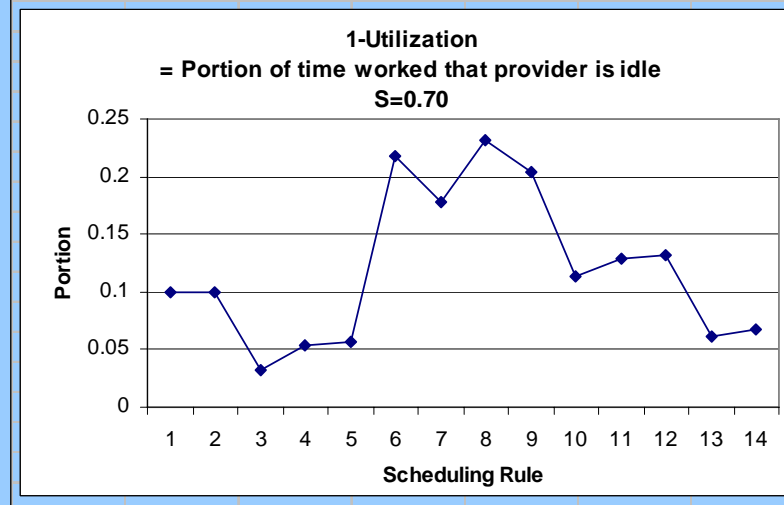
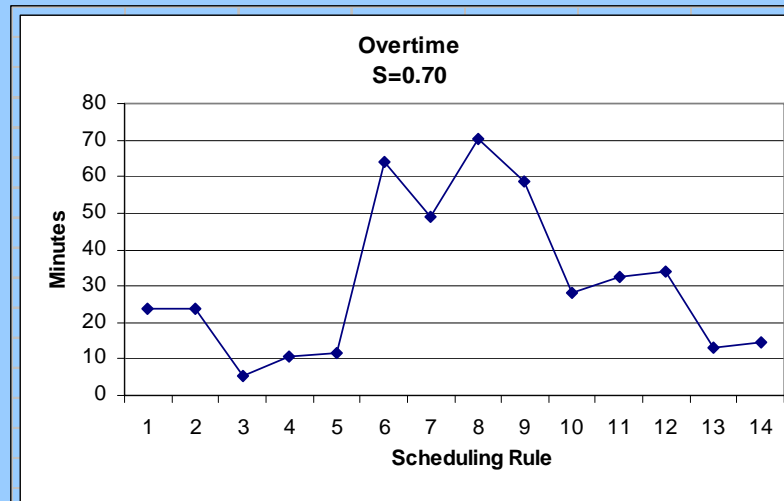
Second Stage Models

- *Simulation* models are used to model patient arrivals as a binomial probability distribution with two possible outcomes of each appointment: the patient either shows up or “no-shows.”
 - Tests schedule performance
 - Provider overtime
 - Patient wait time
 - Performance of overtime factor

Performance Measures

- Provider Side: Overtime
 - Overtime reflects the increased amount of time required to perform a fixed amount of work.
 - Idle time represents an efficiency loss, which is already captured in the measurement of overtime.

Simulation Results: Overtime and Idle Time of Varying Scheduling Rules for fixed N, S



Performance Measures

- Patient Side: Maximum Wait Time
 - Average wait time can increase when *fewer* patients show up.
 - Total wait time does not accurately capture the way waiting time is experienced by patients, who are concerned only about their own wait times.
 - Maximum wait time reveals congestion points in a schedule.

Planning Model of a schedule with heavy congestion at 11:20 a.m., concealed by Average Wait Time.

	Arrivals	Start	Finish	<i>Pt Wait</i>	<i>Max Wait</i>	Total=
Appt Time	0 = N	0	0	<i>in block</i>	<i>in Block</i>	13
8:00 AM	1	0	20	0	0	OvTime=
8:10 AM	0	20	20	0	0	20
8:15 AM	0	20	20	0	0	AvgWait=
8:20 AM	1	20	40	0	0	4.62
8:30 AM	0	40	40	0	0	MaxWait=
8:40 AM	1	40	60	0	0	40
8:45 AM	0	60	60	0	0	
8:50 AM	0	60	60	0	0	
9:00 AM	1	60	80	0	0	
9:10 AM	0	80	80	0	0	
9:15 AM	0	80	80	0	0	
9:20 AM	1	80	100	0	0	
9:30 AM	0	100	100	0	0	
9:40 AM	1	100	120	0	0	
9:45 AM	0	120	120	0	0	
9:50 AM	0	120	120	0	0	
10:00 AM	1	120	140	0	0	
10:10 AM	0	140	140	0	0	
10:15 AM	0	140	140	0	0	
10:20 AM	1	140	160	0	0	
10:30 AM	0	160	160	0	0	
10:40 AM	1	160	180	0	0	
10:45 AM	0	180	180	0	0	
10:50 AM	0	180	180	0	0	
11:00 AM	1	180	200	0	0	
11:10 AM	0	200	200	0	0	
11:15 AM	0	200	200	0	0	
11:20 AM	3	200	260	60	40	
11:30 AM	0	260	260	0	0	
11:40 AM	0	260	260	0	0	
11:45 AM	0	260	260	0	0	
11:50 AM	0	260	260	0	0	
12:00 PM	0	260	260	0	0	
Total=	13	Overtime=	20	60	=Total Wait	

Observations and Guidelines for constructing balanced schedules

- Patient wait time is minimized when all appointments are scheduled as late as possible and are spread as far apart as possible.
- Consider only reasonable schedules – if all appointments were scheduled at the latest time, then overtime would be maximized, but it would not be reasonable to keep all of the earlier appointment times unscheduled.
- Overtime is minimized when all appointments occur as early as possible in the schedule.
- Scheduling all appointments at the same time at the start of the clinic maximizes patient wait time.

How low show rates must be to justify multiple booking for $P \leq 0.25$

		Probability that all K scheduled patients show up	Probability that more than one patient shows up when block size is:									
S	K		2	3	4	5	6	7	8	9	10	
1	12	1	1	1	1	1	1	1	1	1	1	
0.9	13	0.2542	0.81	0.972	0.9963	0.9995	0.9999	1	1	1	1	
0.8	15	0.0352	0.64	0.896	0.9728	0.9933	0.998	0.9996	0.9999	1	1	
0.7	17	0.0023	0.49	0.784	0.9163	0.9692	0.989	0.996	0.9987	0.9996	0.99986	
0.6	20	3.6562E-05	0.36	0.648	0.8208	0.913	0.959	0.981	0.9915	0.996	0.99832	
0.5	24	5.9605E-08	0.25	0.500	0.6875	0.8125	0.891	0.938	0.9648	0.980	0.98926	
0.4	30	1.1529E-12	0.16	0.352	0.5248	0.663	0.767	0.841	0.8936	0.929	0.95364	
0.3	40	1.2158E-21	0.09	0.216	0.3483	0.4718	0.580	0.671	0.7447	0.804	0.85069	
0.2	60	1.1529E-42	0.04	0.104	0.1808	0.2627	0.345	0.423	0.4967	0.564	0.62419	
0.1	120	1.0000E-120	0.01	0.028	0.0523	0.0815	0.114	0.150	0.1869	0.225	0.2639	

Planning Model Calculations

- Patient wait time

$$\begin{aligned}
 W_j &= \sum_{k=1}^{y_j} W_{j,k} = \text{Max}\{0, F_{j-1} - A_j\} + (\text{Max}\{0, F_{j-1} - A_j\} + D) \\
 &\quad + (\text{Max}\{0, F_{j-1} - A_j\} + D) + D + \dots \\
 &= y_j \text{Max}\{0, F_{j-1} - A_j\} + D \sum_{i=1}^{y_j-1} (y_j - i)
 \end{aligned}$$

- Overtime Factor to predict Overtime

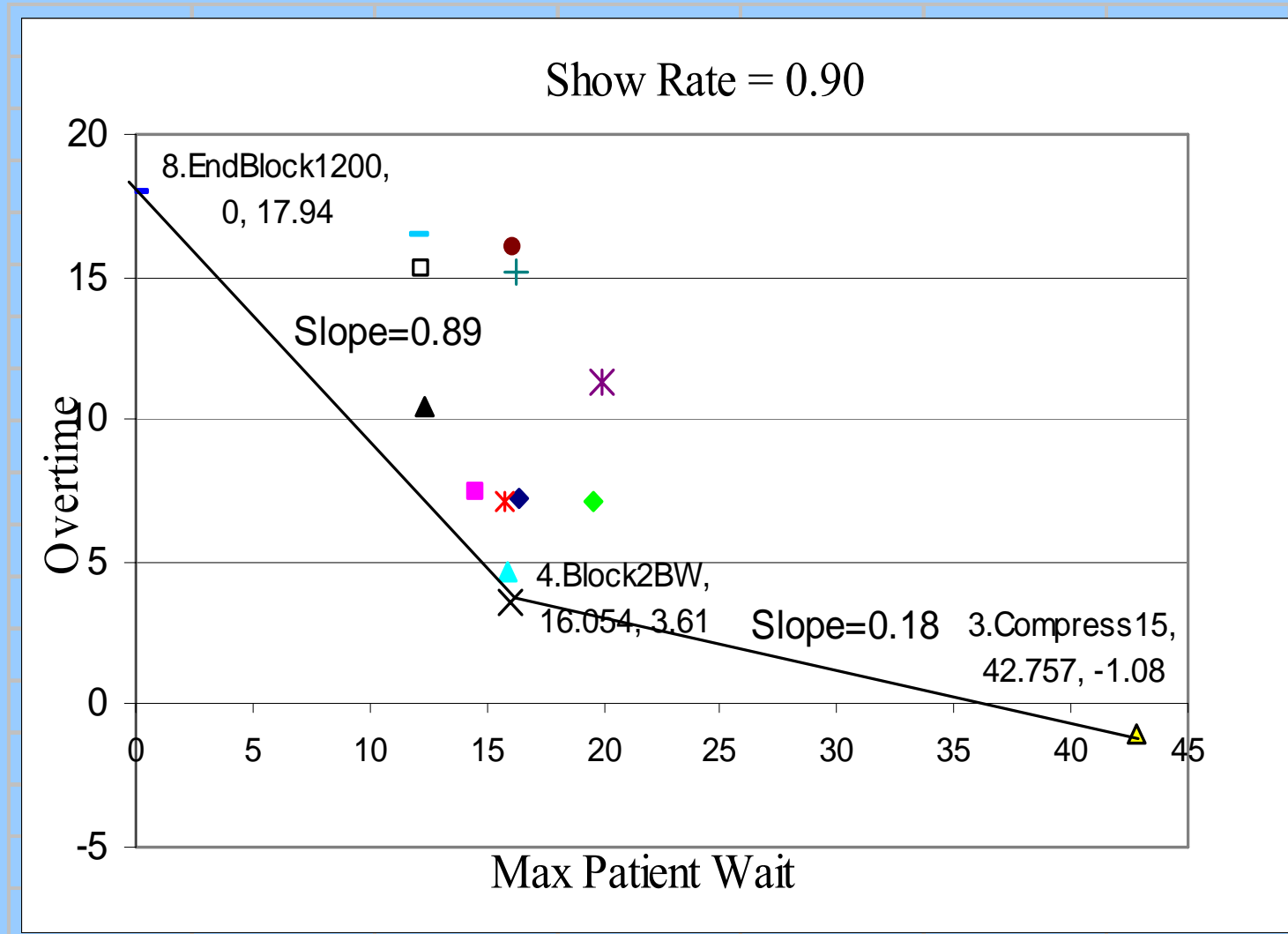
$$V_{S,r} = \frac{F_{L,S} \sum_{j=1}^L (y_j A_j^2)_{S,r}}{F_{L,1} \sum_{j=1}^L (y_j A_j^2)_{1,1}} \quad \forall (S, r \in S)$$

Appointment Time	Schedule Number											
	4	5	6	7	8	9	10	11	12	13	14	
8:00 AM	2	2	1	1	1	1	1	1	1	1	1	1
8:10 AM	0	0	0	0	0	0	0	0	0	0	0	0
8:15 AM	0	0	0	0	0	0	1	0	1	0	0	0
8:20 AM	1	1	1	1	1	1	0	1	0	1	1	1
8:30 AM	0	0	0	0	0	0	1	0	1	0	0	0
8:40 AM	1	1	1	1	1	1	0	1	0	1	1	1
8:45 AM	0	0	0	0	0	0	1	0	1	0	0	0
8:50 AM	0	0	0	0	0	0	0	0	0	0	0	0
9:00 AM	1	0	1	1	1	1	0	1	1	1	1	1
9:10 AM	0	0	0	0	0	0	1	0	0	0	0	0
9:15 AM	0	0	0	0	0	0	0	0	0	1	1	1
9:20 AM	1	2	1	1	1	1	0	1	1	0	0	0
9:30 AM	0	0	0	0	0	0	1	0	0	1	1	1
9:40 AM	1	1	1	1	1	1	0	1	1	0	0	0
9:45 AM	0	0	0	0	0	0	1	0	0	1	1	1
9:50 AM	0	0	0	0	0	0	0	0	0	0	0	0
10:00 AM	1	1	1	1	1	1	0	1	1	1	1	1
10:10 AM	0	0	0	0	0	0	1	0	0	0	0	0
10:15 AM	0	0	0	0	0	0	0	0	0	1	0	0
10:20 AM	1	0	1	1	1	1	0	1	1	0	1	1
10:30 AM	0	0	0	0	0	0	1	0	0	0	0	0
10:40 AM	1	2	1	1	1	1	0	1	1	1	1	1
10:45 AM	0	0	0	0	0	0	1	0	0	0	0	0
10:50 AM	0	0	0	0	0	0	0	0	0	0	0	0
11:00 AM	1	1	1	1	1	1	0	1	1	1	1	1
11:10 AM	0	0	0	0	0	0	1	0	0	0	0	0
11:15 AM	0	0	0	0	0	0	0	1	0	0	0	0
11:20 AM	1	1	1	1	1	1	0	0	1	1	1	1
11:30 AM	0	0	0	1	0	0	1	1	0	0	0	0
11:40 AM	1	1	2	1	1	1	0	0	1	1	1	1
11:45 AM	0	0	0	0	0	1	1	1	0	0	0	0
11:50 AM	0	0	0	0	0	0	0	0	0	0	0	0
12:00 PM	0	0	0	0	1	0	0	0	0	0	0	0
Total=	13	13	13	13	13	13	13	13	13	13	13	13
OvTime=	20	20	20	20	20	20	20	20	20	20	20	20
AvgWait=	18.46	15.38	1.54	2.31	0.00	1.15	10.38	2.31	16.15	11.92	11.54	
MaxWait=	20	20	20	20	0	15	15	15	20	25	20	
OvtFactor	1.08	1.13	1.34	1.32	1.39	1.35	1.20	1.32	1.09	1.12	1.13	

Performance of Calculated Overtime Factor as a predictor of simulated overtime

- Overtime factor correctly corresponds to the rank order of overtime for over 92% of the schedules simulated for each show rate.
- Regression analysis was performed on the same set of data to determine the effectiveness of Overtime Factor V and show rate S as predictors of simulated overtime. $F = 3.1825E-19$
 $R^2 = 0.9139$
- Show rate alone is not quite significant at the 0.05 level but V and the product $S * V$ are highly significant ($p \ll .001$).

The efficient frontier of simulation results



Minimizing total cost

$$C = \pi W + \omega O$$

- The optimal schedule that minimizes cost among those tested is at the point where π/ω has a value between the slopes of two adjacent line segments on the efficient frontier. (Ho and Lau, 1992)
- The set of schedules on the efficient frontier varies with show rate.
- Selection of the best schedule for a given show rate depends not only on the schedule and the show rate itself but also on the relative performance of the alternative schedules.

Recommendations from Simulation Results

- The simplest practical overbooked schedule 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.
- For all S tested, this schedule is always among the schedules that perform well and should be considered for implementation because they are not dominated by any others.

- Schedules with very tight uniform compression are not recommended for any show rate because they allow no catch-up time and even with stochastic no-shows, large accumulations of patient wait time occur.
- Block scheduling of multiple patients at the same time, especially in large block sizes, is not recommended unless $S \leq 0.50$.

- For $S = 0.90$, patient wait time can be avoided entirely by scheduling one extra appointment at the end of the clinic session, with about 18 minutes of overtime on average.
- If less overtime is desired, this can be accomplished by compressing 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.
- We recommend that these schedules be considered if patients and providers can accept these relatively small impacts for the benefit of adding capacity for one extra patient per provider per clinic session (two per day if sessions run both morning and afternoon) to be offered access to services.

Conclusions

- Wave Scheduling answers the question, Where to put overbooked patients in the schedule?
- Spreadsheet planning models develop schedules that improve performance.
- The heuristic predictor performs with a high degree of accuracy
- Wave scheduling provides new scheduling options that perform well in reducing patient wait time and provider overtime while balancing the two costs.
- Effectively increases access to services.

