

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/7167217>

# Designing Appointment Scheduling Systems for Ambulatory Care Services

Article in *Health Care Management Science* · March 2006

DOI: 10.1007/s10729-006-6279-5 · Source: PubMed

---

CITATIONS

335

---

READS

5,697

3 authors, including:



**Tugba Cayirli**

Ozyegin University

12 PUBLICATIONS 1,601 CITATIONS

[SEE PROFILE](#)



**Emre A Veral**

City University of New York - Bernard M. Baruch College

27 PUBLICATIONS 1,696 CITATIONS

[SEE PROFILE](#)

# Designing appointment scheduling systems for ambulatory care services

Tugba Cayirli · Emre Veral · Harry Rosen

Received: June 2005 / Accepted: October 2005/after two revisions  
© Springer Science + Business Media, Inc. 2006

**Abstract** The current climate in the health care industry demands efficiency and patient satisfaction in medical care delivery. These two demands intersect in scheduling of ambulatory care visits. This paper uses patient and doctor-related measures to assess ambulatory care performance and investigates the interactions among appointment system elements and patient panel characteristics. Analysis methodology involves simulation modeling of clinic sessions where empirical data forms the basis of model design and assumptions. Results indicate that patient sequencing has a greater effect on ambulatory care performance than the choice of an appointment rule, and that panel characteristics such as walk-ins, no-shows, punctuality and overall session volume, influence the effectiveness of appointment systems.

**Keywords** Scheduling/sequencing · Service operations · Health care · Simulation · Systems analysis

---

This research was supported by a Summer Research Grant from the Frank G. Zarb School of Business at Hofstra University.

---

T. Cayirli  
Hofstra University, Frank G. Zarb School of Business, 134  
Hofstra University, Department of Management, Entrepreneurship  
and General Business, Hempstead, NY 11549,  
e-mail: Tugba.Cayirli@hofstra.edu

E. Veral  
Baruch College, CUNY, Zicklin School of Business, 17 Lexington  
Avenue, Department of Management B9-240, New York,  
NY 10010,  
e-mail: Emre.Veral@baruch.cuny.edu

H. Rosen  
Baruch College, CUNY Zicklin School of Business, 17 Lexington  
Avenue Department of Management, B9-240 New York,  
NY 10010  
e-mail: Harry\_Rosen@baruch.cuny.edu

## 1. Introduction

Starting with the pioneering work of Bailey [1], there has been extensive research on outpatient-scheduling in health care. An analysis of the literature reveals that most studies consider appointment systems (AS) with no patient classification, assuming patients are homogeneous for scheduling purposes. This means that the scheduler assigns patients to available slots on a first-call, first-appointment basis. If there are certain classes of patients distinguishable in terms of various attributes such as service time distributions or arrival patterns, this raises the issue of whether or not an AS can be improved by recognizing such differences. In ambulatory care, some variables used for classifying patients include major problem, acute problem, acute problem follow-up and chronic problem [2].

Apart from outpatient services, patient classification has potential applicability in equipment-related medical services. For example, when scheduling appointments for CAT scans, patients may be classified by procedure type (such as head, spine, brain, chest), or by age, if pediatric and geriatric patients are known to require longer times to prepare compared to adult patients. Implementation of a sequence-based AS requires the scheduler to identify each slot by patient class.

The main objective of this study is to investigate the effect of using such information on patient class when designing an AS. Without loss of generality, our study uses a classification scheme of ‘new/return’ to analyze the effects of sequencing at the time of booking. Although other classification criteria such as ‘pediatric/adolescent/geriatric’, or ‘difficult/easy’ are possible, all pertain to a single common element, which has relevance for scheduling decisions: consultation time length.

Our second objective is to expand previous environmental assessment studies which address potential differences in the patient panels in terms of no-shows, service time variability

and overall session volume, by analyzing a wider range of factors and determining how and in what manner these affect the performance of AS. This study examines the effects of presence of walk-ins and patient punctuality – two factors largely neglected in the literature.

The use of patient classification for scheduling purposes has been considered by a number of studies. Some of the classification schemes addressed in these studies include new/return [3], variability of service times [4] and type of procedure [5]. A simulation study by Walter [6] investigated an application to a radiology department, where service times depend on factors such as patient's age, physical mobility and type of service. Details on these studies are omitted here for the sake of brevity; however interested readers can refer to Cayirli and Veral [7] for a complete review of the literature on outpatient appointment-scheduling.

This research differs from the earlier studies in a number of ways. First, each sequencing rule is evaluated based on several different underlying appointment rules, allowing us to investigate the possible interactions between sequencing rules and appointment rules. Second, the analysis is carried on a wider range of environments than previous studies, some of which focused on specific clinic applications. Lastly, another contribution is the empirical data collected from a hospital clinic with the goal of basing the simulation model on more realistic grounds.

Our analysis shows that there is an advantage to using additional information on patient class when scheduling patients in ambulatory health-care settings. The proposed sequence-based AS perform better than the traditional rules in the literature. As a result, practical guidelines are developed to help managers choose the best AS based on the specific characteristics of their clinic environments.

The remainder of this paper is organized as follows. Section 2 discusses the simulation model, empirical data analysis and the factor settings for decision and environmental factors in the experimental design. Results of the simulation experiments are reported in Section 3, followed by a summary of the findings and practical considerations in Section 4.

## 2. Research methodology

The experimental methodology for this study was implemented on a Gateway Pentium 4 (2.60GHz 512 MB RAM) computer using a discrete event simulation model written in GPSS/H language.

### 2.1. Decision factors

There are two decision factors in the experimental model: (i) **sequencing rule (SEQ)** determines the order in which calling patients are assigned to blocks based on a particular patient

classification scheme, and (ii) **appointment rule** determines the basic template of the AS by specifying the number of patients scheduled to each appointment slot (i.e., block size) and the length of appointment intervals. This study tests six sequencing rules and seven appointment rules, resulting in a total of forty-two AS.

#### 2.1.1. Sequencing rules (SEQ)

This paper uses a patient classification scheme of 'new/return' patients. New patients are defined as those who are totally new to the clinic; return patients are former patients that arrive with new problems or for follow up of an old problem. Empirical data collected in a variety of specialties reveal that the mean service time of new patients is usually higher than that of return patients [8], [9]. The wide range of applications of this scheme in practice, as well as the availability of empirical support, made this classification the best candidate for our analysis. However, the results of this study can be generalized to any classification scheme based on service time length. Thus, the scheme may also be called 'long/short' service times, if preferred.




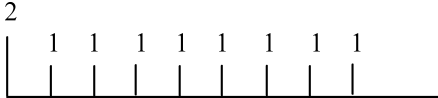
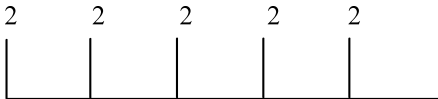
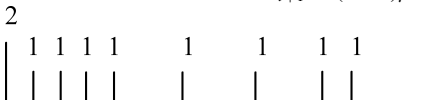
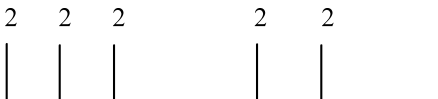
Six sequencing rules are tested: **FCFA** represents the base experimental setting with "no sequencing rule," patients receiving appointment slots on a first-call, first-appointment basis; **ALTER** orders new and return patients in an alternating pattern (RNRNRNR...); **NWBG** schedules new patients in the beginning, and return patients in the remaining part of the session, based on the expected percentage of each patient class (NNN...RRRR); **RTBG** schedules return patients in the beginning (RRRR...NNN); **NWBND** schedules new patients in the beginning and in the end (NN.RRR.NN) and **RTBND** schedules return patients in the beginning and in the end (RR.NNN.RR). These rules have a pattern that is identical to those tested by Klassen and Rohleder's [4] study where they addressed sequencing 'low/high-variance' patients.

#### 2.1.2. Appointment rules (RULE)

Table 1 summarizes the description and formulation of seven appointment rules included in this study. Calculations of appointment times for the  $i^{\text{th}}$  patient ( $t_i$ ) are based on the mean service times of patients ( $\mu$ ), the standard deviation of service times ( $\sigma$ ), parameters  $\beta_i$  and  $k_i$ . Multipliers  $\beta_i$  adjust how early or late appointment times are relative to the benchmark rule, and  $k_i$  determines which patients are scheduled to arrive early, or late, relative to the benchmark rule.

The benchmark rule is the individual-block/fixed-interval rule (**IBFI**), which calls patients individually at intervals equal to the mean service times of patients. We included Ho and Lau's [10] best rule, **OFFSET**, where the initial ( $k_1 - 1$ ) patients are scheduled earlier and the rest are

**Table 1** Appointment rules

Symbol	Description	Formulations <sup>1</sup>
<b>IBFI</b>	Individual-block/fixed-interval rule calls patients individually at intervals equal to the mean service times of patients	$t_1 = 0$ $t_i = t_{i-1} + \mu$ for $i > 1$
		
<b>OFFSET</b>	Individual-block/variable-interval rule, where initial $(k_1 - 1)$ patients are scheduled earlier, and the rest are scheduled later compared to IBFI	$t_i = (i - 1)\mu - \beta_1(k_1 - i)\sigma$ for $i \leq k_1$ , and $t_i = (i - 1)\mu + \beta_2(i - k_1)\sigma$ for $i > k_1$
		
<b>DOME</b>	Individual-block/variable-interval rule, where initial $(k_1 - 1)$ patients are scheduled earlier, patients $(k_1 + 1)$ through $(k_2 - 1)$ are scheduled later, and the rest earlier compared to IBFI	$t_i = (i - 1)\mu - \beta_1(k_1 - i)\sigma$ for $i \leq k_1$ , $t_i = (i - 1)\mu + \beta_2(i - k_1)\sigma$ for $k_1 < i < k_2$ , and $t_i = (i - 1)\mu - \beta_3(i - k_2)\sigma$ for $i \geq k_2$
		
<b>2BEG</b>	Individual-block/fixed-interval rule with an initial-block of two patients	$t_1 = t_2 = 0$ $t_i = t_{i-1} + \mu$ for $i > 2$
		
<b>MBFI</b>	Multiple-block/fixed-interval rule calls patients two-at-a-time with intervals set equal to twice the mean service time	$t_i = t_{i+1} = (i - 1)\mu$ for $i = 1, 3, 5, 7, \dots$
		
<b>2BGDM</b>	Combination of the 2BEG and the DOME rules	$t_i = t_{i+1} = (i - 1)\mu - \beta_1(k_1 - i)\sigma$ for $i = 1$ $t_{i+1} = (i - 1)\mu - \beta_1(k_1 - i)\sigma$ for $2 \leq i \leq k_1$ , $t_{i+1} = (i - 1)\mu + \beta_2(i - k_1)\sigma$ for $k_1 < i < k_2$ , and $t_{i+1} = (i - 1)\mu - \beta_3(i - k_2)\sigma$ for $i \geq k_2$
		
<b>MBDM</b>	Combination of the MBFI and the DOME rules	$t_i = t_{i+1} = (i - 1)\mu - \beta_1(k_1 - i)\sigma$ for $i = 1, 3, \dots, k_1$ $t_i = t_{i+1} = (i - 1)\mu + \beta_2(i - k_1)\sigma$ for $k_1 < i < k_2$ , and $t_i = t_{i+1} = (i - 1)\mu - \beta_3(i - k_2)\sigma$ for $i \geq k_2$ ; where $i, k_1, k_2$ are odd integers
		

<sup>1</sup>  $t_i$  is the appointment time for patient  $i$ ,  $\mu$  is the mean service time,  $\sigma$  is the standard deviation of service time.  $\beta_i$  are multipliers that adjust how early or late appointment times are relative to the benchmark rule (IBFI), and  $k_i$  are early/delay parameters that determine which patients are scheduled to arrive early or late relative to IBFI.

scheduled later compared to **IBFI**. During our pilot study, we tested the best-performing parameters reported in Ho and Lau [10], namely the  $(\beta_1, \beta_2, k_1)$  combinations of (0.15, 0.30, 5) and (0.25, 0.50, 5) for clinics with 10 patients. Both cases have  $\beta_1 < \beta_2$ , which result in variable intervals that are rel-

atively shorter for the initial patients compared to the rest (Note that the opposite is true when  $\beta_1 > \beta_2$ , and the rule results in fixed intervals when  $\beta_1 = \beta_2$ ). However, for the environments that we investigated, ( $\beta_2 = 0.50$ ) resulted in appointment times beyond the clinic end time. Overall, the

combination ( $\beta_1 = 0.15$ ,  $\beta_2 = 0.30$ ,  $k_1 = 5$ ) performed better, and therefore was chosen for our main experiments ( $k_1$  value was adjusted to 10 for  $N = 20$ ).

The **DOMÉ** rule is included based on findings of recent analytical studies, which observe a “dome” pattern in optimal solutions where appointment intervals gradually increase toward the middle and then decrease slightly at the end of the session [11–13]. The dome pattern is “roughly” represented using a formula parallel to that of the **OFFSET** rule (see Table 1 for formulations). Setting  $\beta_1 < \beta_2$  results in relatively shorter intervals for the initial ( $k_1 - 1$ ) patients, and longer intervals after patient  $k_1$ . The third multiplier,  $\beta_3$ , shortens subsequent appointment intervals after patient ( $k_2 - 1$ ) toward the end of the session. There are some restrictions on multipliers  $\beta_2$  and  $\beta_3$  to preserve the original patient sequence ( $t_i < t_{i+1}$ ). More specifically, two constraints exist:  $\beta_2 < \mu/(k_2 - k_1 - 1)\sigma$  and  $\beta_3 < \mu/\sigma$ . Pilot runs tested each parameter at two levels ( $\beta_1 = 0.15, 0.25$ ;  $\beta_2 = 0.30, 0.50$ ;  $\beta_3 = 0.05, 0.25$ ;  $k_1 = 3, 5$ ;  $k_2 = 7, 9$ ). As a result, the combination ( $\beta_1 = 0.15$ ,  $\beta_2 = 0.30$ ,  $\beta_3 = 0.05$ ,  $k_1 = 5$ ,  $k_2 = 9$ ) was chosen. There was no significant difference among  $\beta_3$  values tested at  $\alpha = 0.05$  based on Tukey’s test for pairwise comparisons. Therefore,  $\beta_3$  was arbitrarily set at 0.05. Similar to the **OFFSET** rule,  $k$ -values were adjusted to clinic size using  $k_1 = 10$  and  $k_2 = 18$  for clinics with 20 patients.

The **2BEG** rule, first introduced by Bailey [1], calls two patients at the beginning of the session, and the rest individually at fixed intervals. This rule is identified among the best performing appointment rules in the extensive comparative analyses of Ho and Lau [10] and Klassen and Rohleder [4]. This study also tests the multiple-block/fixed-interval rule (**MBFI**), which calls patients two-at-a-time. Soriano [14], Blanco White and Pike [15] and Cox et al. [3] find that multiple-block rules perform better for the particular clinics they studied. On the other hand, Ho and Lau [10] report that multiple-block rules perform worse than **OFFSET** and **2BEG**, when patients are assumed to arrive punctually. Given these conflicting results, this study aims at testing the **MBFI** rule more rigorously under a wider range of environmental factors.

The **2BGDM** and **MBDM** rules are variations of their original counterparts (i.e., **2BEG** and **MBFI**) with variable-intervals that follow a “dome” pattern. To the best of our knowledge, multiple-block rules with variable-intervals are investigated for the first time in the literature.

## 2.2. Environmental factors

This study investigates six environmental factors. These include the number of patients scheduled per clinic session, or the ‘clinic size’ ( $N$ ), the probability of no-shows ( $P_N$ ), the probability of walk-ins ( $P_W$ ), the coefficient of variation of

service times for return patients ( $CV_{Ret}$ ), the coefficient of variation of service times for new patients ( $CV_{New}$ ), and the mean punctuality of patients ( $Punc$ ).

Ho and Lau’s [10] assessment of environmental factors ( $N$ ,  $P_N$  and  $CV$ ) reports that among the three, no-show probability has the major effect on performance of an AS. A number of other simulation studies cite the importance of including  $N$  when comparing AS [1], [15–17]. In practice, the total number of patients that can be seen in a session will depend on the mean service times. For this reason, our simulation model keeps clinic session length fixed while varying  $N$ . This helps us represent a range of medical disciplines whose average service times vary by their nature.

The probability of walk-ins ( $P_W$ ) is measured as the number of patients who walk in without appointments as a percentage of all appointments. Few studies model walk-ins [16], [18–20]. As one would expect, walk-in rates vary across different specialties [21].

Punctuality of a patient is defined as the difference between a patient’s appointment time and actual arrival time, and it accounts for both earliness and lateness. Mean punctuality of a group of patients ( $Punc$ ) is the average of their individual punctualities, and may be zero although they were individually unpunctual [15]. The majority of the studies in the literature assume all patients are punctual, ignoring the possible effects of unpunctual patients on the performance of AS. Lastly, this study investigates the effects of service time variability for both new patients ( $CV_{New}$ ) and return patients ( $CV_{Ret}$ ), measured by the coefficient of variation.

## 2.3. Simulation model

Most simulation studies analyzed the environment of a specific clinic, thus their findings lacked generalized applicability. Others simulated hypothetical clinics with no empirical basis for identifying probability distributions that represent actual service times and patient arrival patterns. As mentioned by Shafer and Smunt [22], combining the power and flexibility of simulation with empirical data can be one of the most effective ways to help bridge the gap between academic rigor and managerial applicability. For this purpose, this study collected empirical data to establish the validity of some of the input parameters used in our simulation model. Further support came from earlier empirical studies, which led to realistic environments in the simulated clinics. In order to accommodate generalization of results, the hypothetical clinics included environmental factor settings beyond those observed in the clinic.

All combinations of factors were simulated for 100 replications, each of which represented averages of 100 clinic sessions (i.e., 10,000 clinic sessions simulated for each environment). Pilot runs showed that for this sample size, primary performance measures of doctor’s idle time, doctor’s

overtime, and patients' waiting times were accurate within an average  $\pm 30.51$  seconds at the 95 percent confidence level. The statistical analyses on empirical data, the experimental factor settings and the basic assumptions of the simulation model are discussed in the following sections.

### 2.3.1. Empirical data

This study collected data from a primary health care clinic in a New York metropolitan hospital that provides service to about 300,000 outpatients a year. The observed clinic had six doctors, and each doctor scheduled 10 patients for a half-day clinic session that lasted  $3\frac{1}{2}$  hours. Through observation, patient arrival times and appointment times were recorded to derive the distribution for the punctuality of patients. Similarly, consultation start and end times were recorded to obtain the service time distributions.

Generally, it is considered better to use a theoretical distribution rather than the empirical distribution function directly in the simulation model. Such an approach has the advantage of simulating cases beyond those observed empirically [23]. Kolmogorov-Smirnov test results showed normal distribution as a good fit to our punctuality data ( $\alpha = 0.05$ ). The distribution mean was  $-17$  minutes (on average, patients were seventeen minutes early), and standard deviation was 27 minutes. Swartzman's [19] study on patient arrival pattern also reported that normal distribution was a good fit to empirical data collected on patient punctuality (mean  $-2$  minutes).

Service time data were collected on the same doctor for both new and return patients, so that possible differences could be attributed to patient class, and not the doctor. Combining data across several doctors was purposefully avoided due to large differences observed in doctors' mean service times. Over a six-week period, 35 and 90 observations were collected for new and return patients, respectively. With this sample size, nonparametric Mann-Whitney-Wilcoxon test results showed that the difference in the mean service times of new and return patients was statistically significant at  $\alpha = 0.05$ . Kolmogorov-Smirnov test results indicated the lognormal distribution as the best fit to our service time data (at  $\alpha = 0.05$ ). Table 2 shows the summary statistics on primary data collected.

Secondary data on the percentage of no-shows and walk-ins were obtained from the monthly reports of clinics during January-February 2002. Overall, the percentage of no-shows was 0.38 with major differences observed among specialties. For example, colposcopy had zero no-shows, whereas pediatrics neurology had 67 percent no-show rate. A possible reason for the excessive no-shows in this hospital could be the unique patient population that included mostly immigrants. The clinic administrators were aware of this problem, and they were considering policies such as automated phone reminders. As for walk-ins, average rate was 16 percent across all clinics, and some clinics such as peripheral vascular diseases, had zero walk-ins. The highest walk-in rates were in pediatrics neurology and hepatology (up to 60 percent).

### 2.3.2. Simulated clinics

A  $3\frac{1}{2}$ -hour clinic session was used to simulate a typical half-day period. Clinic size was investigated at two levels ( $N = 10, 20$ ), which corresponded to mean service times ( $\mu = 21, 10.5$  minutes). These weighted averages resulted from combinations of ( $\mu_{New} = 30, \mu_{Ret} = 15$ ), and ( $\mu_{New} = 15, \mu_{Ret} = 7.5$ ), where the percentage of new patients was fixed at 0.40, and the ratio of the service time of new patients to that of return patients ( $\mu_{New}/\mu_{Ret}$ ) was fixed at 2.

In this study, we set the experimental  $CV_{Ret}$  and  $CV_{New}$  values at two levels: "low" 0.35 and "high" 0.70. The empirical CV-values reported in the literature range from about 0.35 to 0.85 [7]. Our empirical data collected from the hospital clinic indicated approximately similar values for new and return patients (see Table 2).

Earlier empirical results on patient punctuality reported that patients arrive early more often than late [3], [8], [15], [19], [21]. Two levels were simulated for the mean patient punctuality: patients are on average 0 minutes early, or 15 minutes early. In both cases, the standard deviation was fixed at 25 minutes, determined by our primary data collection.

Excessive no-show rates, such as those observed in the metropolitan hospital clinic, may indicate that clinic administrators are not making the best use of the policies successful in discouraging no-shows (e.g., reminders by phone,

**Table 2** Summary statistics on empirical data

Environmental Factor	Sample Size	Distribution ( $\mu, \sigma$ )	CV ( $\sigma/\mu$ )	[Min, Max]	Skewness
Punctuality of patients	110	Normal ( $-16.62, 27.07$ )	$-1.63$	$[-105, 80]$	$-0.337$
Service times for new patients	35	Lognormal ( $19.09, 6.85$ )	$0.360$	$[10, 40]$	$1.132$
Service times for return patients	90	Lognormal ( $15.50, 5.038$ )	$0.325$	$[6, 29]$	$0.707$

**Table 3** Environmental factor settings

Environmental Factor	Symbol	Levels	Settings
Number of patients scheduled per session	$N$	2	10, 20
Coefficient of variation for return patients	$CV_{Ret}$	2	0.35, 0.70
Coefficient of variation for new patients	$CV_{New}$	2	0.35, 0.70
Probability of walk-ins	$P_W$	2	0, 0.15
Probability of no-shows	$P_N$	2	0, 0.15
Mean punctuality of patients	$P_{unc}$	2	-15, 0 min.

penalties for failed appointments). Likewise, walk-ins may be encouraged to make appointments instead of simply showing up. Some clinics may totally deny access to walk-ins. For this reason,  $P_N$  and  $P_W$  were simulated at two levels (0 and 15 percent) in order to accommodate more common cases reported in the literature [5], [8], [15], [21]. Table 3 lists settings for each environmental factor. In sum, sixty-four hypothetical clinics were simulated in our study ( $2N \times 2CV_{Ret} \times 2CV_{New} \times 2P_W \times 2P_N \times 2P_{unc}$ ).

### 2.3.3. Basic Assumptions

Clinics represent single-server, single-phase queuing systems, where only the doctor's consultation service is considered. Patients are served once per visit. The single-server assumption holds for most ambulatory care settings, because sharing patients among multiple doctors is generally avoided in order to maintain continuity of care.

It is assumed that clinic sessions are independent – no patients from the morning session spill over to the afternoon, or vice versa. Based on results from our empirical data, service times are modeled using lognormal distribution, and patient punctuality is modeled using normal distribution. It is assumed that patients' punctuality is independent of their appointment times. On theoretical grounds, exponential distribution is used to model inter-arrival times of walk-ins. All walk-ins are admitted, regardless of the congestion in the clinic. For simplification, no-show probabilities of new and return patients are assumed to be identical. Furthermore, no-show probability is assumed to be independent of a patient's place in the session.

The simulated clinics use a queue discipline where patients are seen in the order of their appointment times. The presence of unpunctual patients and walk-ins complicates the administration of patient flow. Setting up priority rules is a subjective decision, and is more complex than it may at first appear. It is natural to assume that late patients will tolerate longer waiting times. However, in certain cases, severely penalizing a late patient may not be appropriate if the pa-

tient is late only a few minutes, or if the doctor is running late anyway. Sometimes the doctor will see the next available patient, who happens to arrive early, and the latecomer will lose her place automatically. In short, the effect of lateness is rather circumstantial and the risk of it becoming a problem increases with the magnitude of lateness. Our simulation model mimics a clerk's decision-making process by using a rule where the patient loses her place proportional to the magnitude of lateness (i.e., patient loses one place for every full  $\mu$  minutes of lateness). Walk-ins are given the lowest priority, yet they are not kept waiting inordinately long. A walk-in is forced to wait for at most three scheduled patients, and she is seen earlier if the clinic is empty at any point of time.

This study assumes that service times are i.i.d., yet in practice doctors may increase their service rates if they observe congestion in the waiting area. It is also postulated that the doctor is punctual, and there are no gap times between two consecutive consultations. It is believed that these factors are more controllable, and thus less interesting. Adjustments can be made to account for doctors' habitual lateness, by delaying the first appointment by average lateness time. Similarly, regular gap times (e.g., coffee breaks) can be treated as breaks, or they may be included as a part of consultation times (e.g., paperwork after each consultation).

### 2.4. Performance measures

The primary performance measures used in this study include patients' average waiting time (WAIT), doctor's average idle time per patient (IDLE), and doctor's average overtime per patient (OVER). Since our simulation model includes unpunctual patients, "true" waiting times can be negative when early patients are served before their appointment times ("true" waiting time is calculated by subtracting the greater of {appointment time, arrival time} from the service start-time). Negative "true" waiting times are truncated at zero when computing overall average waiting times of patients. IDLE is calculated by dividing total idle time in a session by the number of patients seen. Similarly, OVER is calculated by dividing total overtime {actual session end time minus scheduled end time} by the number of patients seen.

In an effort to combine the two doctor-related measures, overtime cost is fixed at 50 percent higher than idle time cost. The resulting measure is called IDLE&OVER. This approximate valuation reflects the fact that overtime is a greater concern in practice, since doctors may use idle time more productively during the day, with tasks such as consulting with colleagues, or reviewing patient charts. The following equation represents the total cost of the system based on the

**Table 4** ANOVA results on performance measures

WAIT				IDLE&OVER			
	Sum of Sq	DF	F-value		Sum of Sq	DF	F-value
Corrected Model	14395981	493	18718	Corrected Model	2057287	493	34604
Intercept	66723870	1	42770429	Intercept	8207761	1	68062560
SEQ	3642860	5	467019	SEQ	295091	5	489407
RULE	992956	6	106082	RULE	150303	6	207731
SEQ*RULE	182344	30	3896	SEQ*RULE	16614	30	4592
PW	5331532	1	3417546	N	1181332	1	9796155
PN	2351588	1	1507383	PUNC	131730	1	1092368
N	781981	1	501255	PN	73486	1	609380
CVRET	83262	1	53372	CVNEW	19036	1	157853
CVNEW	30247	1	19389	CVRET	13319	1	110449
PUNC	18253	1	11700	PW	931	1	7723
SEQ*PN	282714	5	36244	SEQ*N	23787	5	39450
SEQ*PW	116924	5	14990	SEQ*PUNC	10595	5	17572
SEQ*PUNC	75573	5	9689	SEQ*PW	6711	5	11130
SEQ*N	10150	5	1301	SEQ*PN	1659	5	2752
SEQ*CVNEW	5834	5	748	SEQ*CVNEW	528	5	876
SEQ*CVRET	5295	5	679	SEQ*CVRET	11	5	18
RULE*N	118859	6	12698	RULE*N	44503	6	61506
RULE*PN	52436	6	5602	RULE*PUNC	4760	6	6579
RULE*PUNC	47374	6	5061	RULE*PW	3991	6	5515
RULE*PW	9367	6	1001	RULE*PN	695	6	961
RULE*CVNEW	2313	6	247	RULE*CVNEW	102	6	141
RULE*CVRET	191	6	20	RULE*CVRET	19	6	27
R-Square	.972			R-Square	.985		

*Decision Factors:* RULE: appointment rule, SEQ: sequencing rule *Environmental Factors:* N: clinic size, PW: probability of walk-ins, PN: probability of no-shows, PUNC: mean punctuality of patients, CVNEW: coefficient of variation for new patients, CVRET: coefficient of variation for return patients.

relative valuation given to the cost of patient's waiting time ( $C_p$ ) and the cost of doctor's idle time ( $C_d$ ):

$$TC = (WAIT)C_p + [(IDLE) + 1.5(OVER)]C_d \quad (1)$$

### 3. Results and analysis

#### 3.1. Effects of decision factors and environmental factors

The simulation results are analyzed by Analysis of Variance (ANOVA), which evaluates the effect of main experimental factors and their interaction effects on each dependent variable. Results are based on average performance statistics across the sixty-four simulated environments. All statistical analysis is done using the SPSS software.

Table 4 reports the ANOVA results for dependent variables WAIT and IDLE&OVER. The model explains 97.2 percent of the total variation in WAIT (similarly,  $R^2 = 98.5\%$  for IDLE&OVER). All the main and two-way interaction effects are statistically significant at  $\alpha = 0.05$  (in fact, all  $p > F$  are less than 0.000). The ANOVA table lists factors in descending

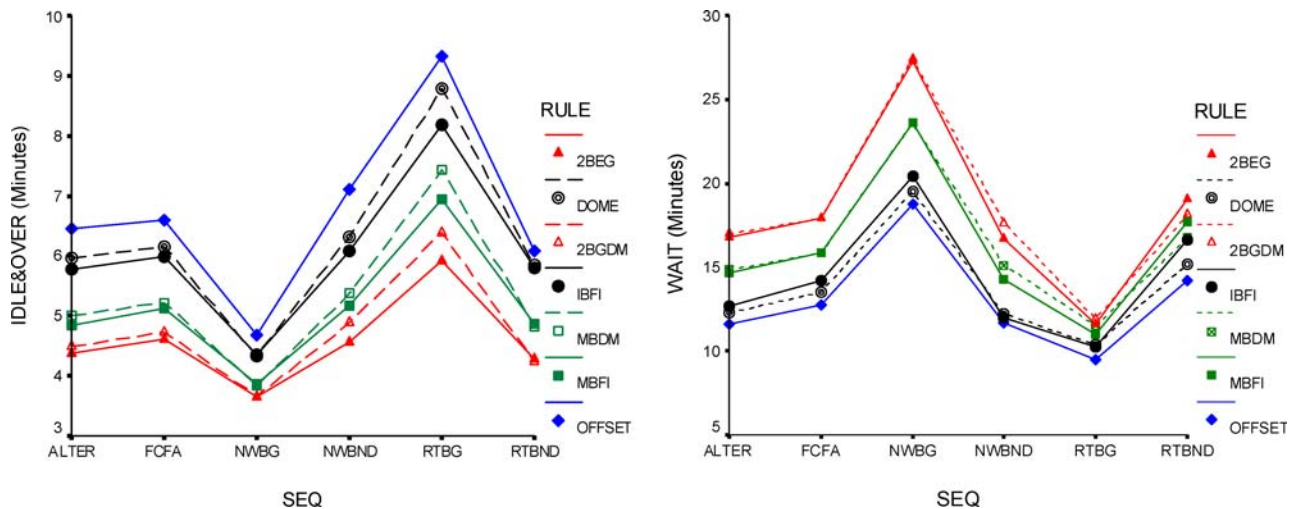
order of their sum of squares (and by extension, F-values), grouped separately for the main and interaction effects of decision factors and environmental factors. This approach facilitates identifying the most important factors within each group, as the sum of square values indicate the proportion of variation explained by each factor.

Three and higher-level interactions are omitted from Table 4, as these account for less than 1 percent of the total variation in dependent variables calculated by the sum of squared values. Also excluded from the analyses, are the two-way interactions between environmental factors (such as  $N*P_W$ ). Given that our main interest is to evaluate different AS, we focus on interaction effects between decision factors (SEQ\*RULE), and between decision factors and environmental factors (such as  $SEQ*P_N$ ). These are discussed in the following sections.

##### 3.1.1. Main effects of decision factors

Both decision factors, SEQ and RULE, are significant at 95 percent confidence level. A comparison of their sum of square values shows that SEQ explains a larger proportion of variability than RULE (Table 4). Given the assumptions built into





**Fig. 1** Interaction effects between decision factors (SEQ\**RULE*)

our model, this result suggests that the choice of sequencing rule is more critical than the choice of appointment rule when designing AS.

### 3.1.2. Interaction effect between decision factors

ANOVA results (Table 4) show that the interaction effect between decision factors, SEQ\**RULE*, is significant at  $\alpha = 0.05$  for both measures of WAIT and IDLE&OVER. Figure 1 shows the interaction plots for SEQ\**RULE*, which reveal only infatuating effects, and not cross effects. This means the effect of sequencing is similar on all appointment rules, and the effect of appointment rule is similar on all sequencing rules. For example, in Figure 1, a shift from **ALTER** to **FCFA** increases IDLE&OVER regardless of the underlying appointment rule used. On the other hand, if cross effects existed, this would imply that rankings of sequencing rules differ across different appointment rules (or rankings of appointment rules change based on the underlying sequencing approach used).

### 3.1.3. Main effects of environmental factors

Among the six environmental factors tested,  $P_W$ ,  $P_N$  and  $N$  emerge as the most important factors for WAIT based on sum of squares in Table 4. Similarly,  $N$ ,  $P_{unc}$  and  $P_N$  are the most important factors for IDLE&OVER. Overall,  $CV_{Ret}$  and  $CV_{New}$  are relatively less critical, even though significant at  $\alpha = 0.05$ .

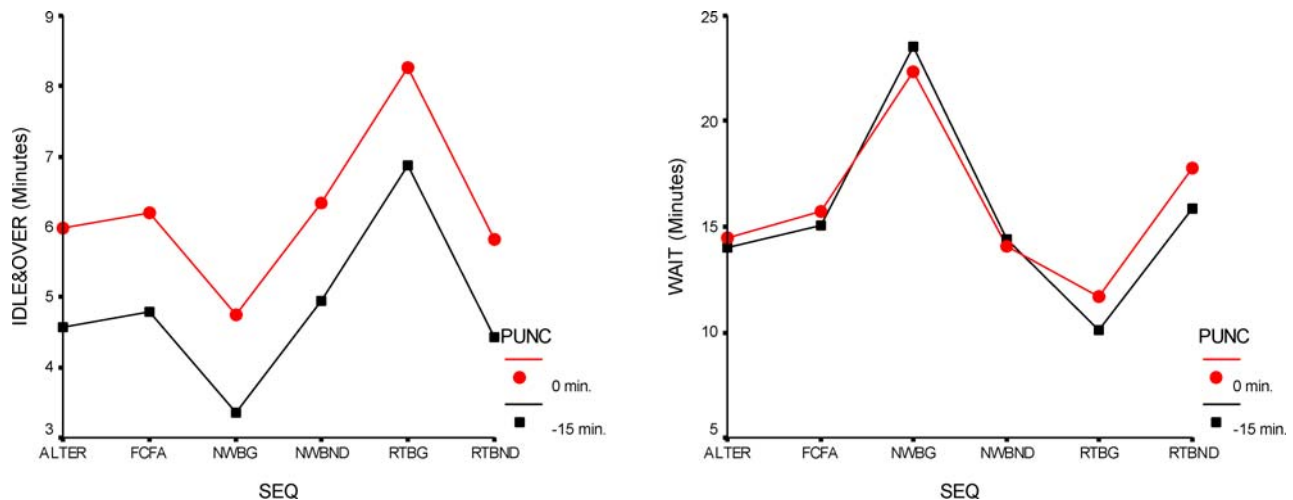
Analyzing the differences in means between factor levels leads to some findings on the effects of each environmental factor. As one would expect, an increase in no-shows increases idle time, and decreases doctor overtime and patient waiting time. The exact opposite effect is ob-

served for walk-ins. As the percentage of walk-ins increases, WAIT and OVER measures increase, and IDLE measure decreases. Higher earliness of patients (i.e.,  $P_{unc} = -15$  minutes) improves all three measures, yet the effect is weaker on WAIT (Recall that the WAIT measure ignores the portion of a patients' waiting time prior to the appointed time). High service time variability ( $CV$ ) deteriorates clinic performance on all measures. Last, an increase in clinic size ( $N$ ), which corresponds to shorter average service times, improves clinic performance for all criteria. This observation highlights the importance of designing effective AS for clinics/practices with relatively longer service times.

### 3.1.4. Interaction effects between decision factors and environmental factors

All two-way interactions of environmental factors with SEQ and *RULE* are significant at  $\alpha = 0.05$ . However, interactions with  $CV_{Ret}$  and  $CV_{New}$  emerge as the weakest (Table 4). Interaction plots are analyzed to check the nature of these interactions. Figure 2 illustrates the SEQ\* $P_{unc}$  interaction on IDLE&OVER and WAIT measures. Higher patient lateness (i.e.,  $P_{unc} = 0$  minutes) increases doctor idle time plus overtime, yet the effect is more or less uniform for all sequencing approaches, as observed by parallel lines. On the other hand, there are slight cross effects observed for WAIT measure. Generally, higher patient lateness leads to higher WAIT, except for the **NWBG** rule (change in **NWBND** is not significant at alpha level 0.05).

Discussion of these findings serves to further illustrate the general nature of interaction effects between environmental factors and decision variables. They are mostly due to the extent to which environmental factors affect the performance of sequencing and appointment rules. More importantly, they



**Fig. 2** SEQ\*Punc interaction

are not due to a cross effect, which would imply opposite effect of an environmental factor for different sequencing or appointment rules.

### 3.2. Comparison of appointment systems

As a result of our analysis of the main and interaction effects of environmental and decision factors in Section 3.1,  $CV_{Ret}$  and  $CV_{New}$  prove to have relatively less effects on the performance of AS. In the interests of (i) developing a more parsimonious model, (ii) creating an actionable set of guiding principles for the practitioners, and (iii) focusing on data ordinarily available to the practitioner, we exclude  $CV_{Ret}$  and  $CV_{New}$  from further analysis. Thus the total number of clinic environments is reduced to sixteen ( $2N \times 2P_W \times 2P_N \times 2P_{unc}$ ).

Efficient frontiers that plot the mean performance measures can be used for simultaneous assessment of multiple measures of interest. This is also a common approach used for comparing AS in appointment-scheduling literature. In Ho and Lau [10] and Klassen and Rohleder [4], AS are plotted with their corresponding IDLE and WAIT values, and the best performing AS that minimize both measures *simultaneously* identify the ‘efficient frontier’. Essentially, AS that are on a two-dimensional efficient frontier represent the most efficient trade-offs between two conflicting performance measures.

We plotted the efficient frontiers to compare forty-two AS (combinations of six sequencing rules and seven appointment rules) in terms of IDLE&OVER and WAIT. The general shape of the efficient frontier remains similar in all sixteen environments investigated. Thus, Figure 3, which shows the efficient frontier for a particular clinic environment ( $N = 10$ ,  $P_N = 0.15$ ,  $P_W = 0.15$ ,  $P_{unc} = -15$ ) is representative in portraying generalized findings. The figure repeats the same plot by SEQ and by RULE, the combina-

tion of which determines the specific AS. It is noticeable that groupings by sequencing rules are more clustered compared to groupings by appointment rules. This reinforces the ANOVA findings, which indicated that the choice of sequencing rule mattered more than the choice of appointment rule.

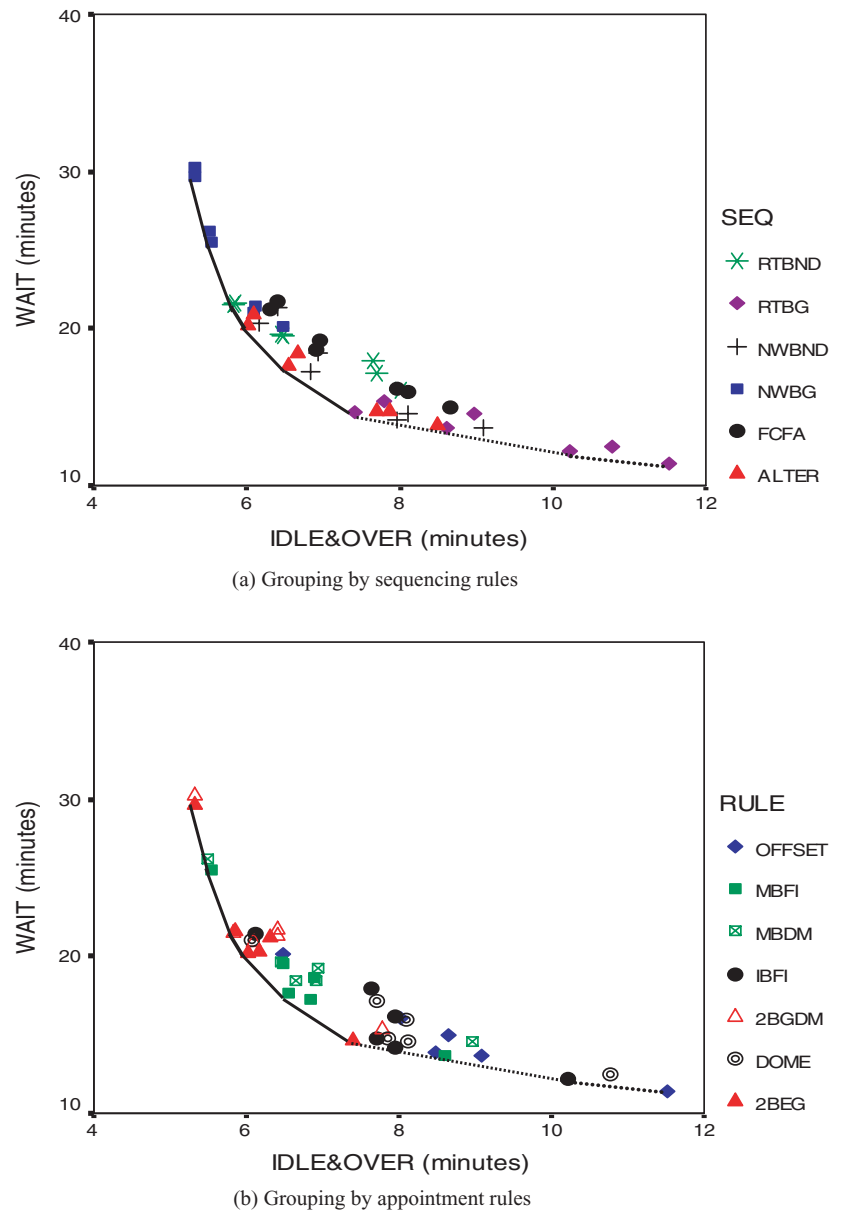
In Figure 3, only six out of the forty-two AS remain on the efficient frontier as best-performers. Table 5 lists the WAIT and IDLE&OVER values, as well as the corresponding  $C_d/C_p$  values at which each AS becomes the best choice. For example, **2BEG-NWBG** is the best choice when it is assumed that doctor’s time is at least 19 times more valuable than patient’s time. **IBFI-RTBG** and **OFFSET-RTBG** are classified as ‘infeasible’, because these are best choices for  $C_d/C_p$  ratios less than 1 - an unlikely assumption in practice (infeasible AS are indicated with a dashed-line in Figure 3). As the relative value given to doctor’s time (i.e.,  $C_d/C_p$  ratio) decreases, the preference shifts from **2BEG-NWBG** to **MBFI-NWBG**, **2BEG-RTBND**, **2BEG-ALTER**, **MBFI-ALTER** and **2BEG-RTBG**. Notice that the base sequencing rule, **FCFA**, stays above the efficient frontier indicating an inferior performance

**Table 5** Best appointment systems for environment: ( $N = 10$ ,  $P_N = 0.15$ ,  $P_W = 0.15$ ,  $P_{unc} = -15$ )

Appointment System	IDLE&OVER	WAIT	$C_d/C_p \geq$
2BEG-NWBG	5.315	29.666	19.09
MBFI-NWBG	5.533	25.523	13.71
2BEG-RTBND	5.851	21.595	6.74
2BEG-ALTER	6.015	20.242	4.75
MBFI-ALTER	6.557	17.664	3.52
2BEG-RTBG	7.403	14.686	1.00
IBFI-RTBG	10.224	12.210	INF
OFFSET-RTBG	11.521	11.378	INF

INF: Infeasible ( $C_d/C_p < 1$ ).

**Fig. 3** Efficient frontier for environment: ( $N = 10$ ,  $P_N = 0.15$ ,  $P_W = 0.15$ ,  $P_{unc} = -15$ ).



(Figure 3). This was true for all the sixteen environments tested.

Even though the shape of the efficient frontiers remains similar, the exact  $C_d/C_p$  values at which each AS becomes the best choice, varies from one environment to another. After our detailed analysis of sixteen efficient frontiers and their corresponding  $C_d/C_p$  values, we summarize our most important findings.

#### Sequencing rules:

- Under all environments, scheduling patients on a first-call, first-appointment (**FCFA**) basis performs worse than the sequence-based AS.
- Upper-left part of the efficient frontiers is consistently dominated by the **NWBG** rule, and the preference shifts to **ALTER** and then to **RTBG**, as the relative value given to patient's time increases.
- Scheduling long-consultation patients in the beginning of the session (i.e., **NWBG**) performs best when applied to appointment rules that utilize multiple blocks (i.e., **2BEG** and **MBFI**). However, in clinic environments with early arrivals and low no-show rates, these AS result in very high patient waiting times with only minor improvements in doctor idle time and overtime. Thus, they should be avoided unless doctor's time is deemed substantially more valuable than patient's time.

- Combining the approach of scheduling short-consultation patients in the beginning of the session (i.e., **RTBG**) with the **2BEG** rule, appears on the efficient frontiers as one of the best-performing AS. On the other hand, **RTBG** sequencing should not be combined with individual-block rules, as the resulting AS are generally infeasible. Few exceptions include environments characterized by high walk-ins, low no-shows, and/or early patients.
- The **NWBND** rule usually performs badly, or at best similar to the **ALTER** rule. The **RTBND** rule performs best only in combination with **2BEG** or **2BGDM** rules, and less so in larger clinics with  $N = 20$ .

#### *Appointment rules:*

- In terms of appointment rules, the preference shifts from **2BEG** to **MBFI**, and then to individual-block rules (i.e., **IBFI**, **DOME** and **OFFSET**), as patient's time becomes increasingly valuable.
- Among the appointment rules tested, **2BEG** has the strongest existence on the efficient frontiers. This rule performs best in combination with all sequencing approaches tested. **MBFI** performs best in combination with **ALTER** and **NWBG**, whereas the individual-block rules, **IBFI** and **OFFSET**, both perform best with **ALTER** and **RTBG**.
- The individual-block rules are more suited to clinics/specialties with shorter consultation times ( $N = 20$ ), especially when patients are known to arrive early, walk-ins are high, and no-shows are low. On the other hand, in clinics with longer consultation times ( $N = 10$ ), individual-block rules should be used only if walk-ins are high and no-shows are low, and only if patient's time is assumed as valuable as doctor's time. In the latter case, **MBFI** and **2BEG** generally perform better.
- The **OFFSET** rule, which results in lowest patient waiting time, and highest doctor idle time and overtime, is generally infeasible. It becomes the best choice when walk-ins are high and no-shows are low, yet under the restrictive assumption that patient's time is valued as high as doctor's time. The simpler **IBFI** rule appears as the best choice under a wider range of environments, compared to **OFFSET**.
- One of the goals of this study was to test the performance of variable-interval rules with "dome-shaped" intervals (i.e., **DOME**, **2BGDM** and **MBDM**). These rules performed either inferior to or insignificantly different than their fixed-interval counterparts, based on Tukey's test results at  $\alpha = 0.05$ .
- **MBFI**, identified as a poor performer in Ho and Lau's study [10], performs as one of the best appointment rules in this study. Most likely, the consequences of multiple blocks were not as severe when the assumption of punctual patients was relaxed in our simulation model.

#### 4. Summary of findings and practical considerations

The most relevant finding of this study to the practitioners involved in designing appointment systems (AS) is that sequencing decisions have a more pronounced impact on clinic performance than the choice of an appointment rule. New sequence-based AS introduced in this study were successful in improving patient waiting time, doctor idle time and overtime, without any trade-offs.

Among the six environmental factors we investigated, no-shows, walk-ins, clinic size and patient punctuality, emerged as the major factors affecting the performance, and the ultimate selection of an appointment system. Service time variability of patients proved to be less important, even though statistically significant. Limiting environments to these factors, forty-two appointment systems (combinations of six sequencing rules and seven appointment rules) were compared on the basis of mean performance measures plotted on efficient frontiers. The results of our analysis indicated that, placing new patients in the beginning of the session is preferred when doctor's idle time is assumed to be highly valuable compared to patients' time. At the other extreme, placing return patients in the beginning of the session is preferred when patients' time is highly valued. For each environment, alternating new and return patients performed the best in between these two extremes. In short, the simpler **NWBG**, **ALTER** and **RTBG** rules generally performed the best among the sequencing rules tested.

In terms of appointment rules, fixed-interval rules of **2BEG**, **MBFI** and **IBFI** dominated the efficient frontiers as best performers. Furthermore, our findings indicated that the individual-block rules are mostly suited to specialties with short consultation times. In fact, these rules should be avoided in clinics with long consultation times, unless walk-ins are high, no-shows are low, and patient's time is assumed to be equally valuable as doctor's time. On the other hand, rules that utilize multiple-blocks, **2BEG** and **MBFI**, appear among the best performers in *all* the sixteen environments investigated. Confirming our previous findings that indicated sequencing approaches are more important than appointment rules, the best choice among these appointment rules, depends on the combination with a particular sequencing rule.

The above results can be fashioned into a series of advisory statements that will provide better guidance to managers responsible for designing a wide variety of ambulatory care delivery systems. Decision-makers responsible for group practices, radiology centers, and hospital clinics will be able to make explicit choices of appointment and sequencing rules in light of readily observable characteristics of their patient populations. Different clinics could have different AS within the same organization. They will also be able to make

the all-important trade-off between the respective value of patient time and physician time.

## References

1. Bailey NT (1952) A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society* 14: 185–199
2. Arbitman DB (1986) A primer on patient classification systems and their relevance to ambulatory care. *The Journal of Ambulatory Care Management* 9: 58–81
3. Cox TF, Birchall JF, Wong H (1985) Optimizing the queuing system for an ear, nose and throat outpatient clinic. *Journal of Applied Statistics* 12: 113–126
4. Klassen KJ, (1996) Rohleder TR Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management* 14: 83–101
5. Vanden Bosch PM, Dietz CD (2000) Minimizing expected waiting in a medical appointment system. *IIE Transactions* 32: 841–848
6. Walter SD (1973) A comparison of appointment schedules in a hospital radiology department, *British Journal of Preventive and Social Medicine* 27: 160–167
7. Cayirli T, Veral E (2003) Outpatient-scheduling in health care: A review of the literature. *Production and Operations Management* 12: 519–549
8. Nuffield Provincial Hospitals Trust, *Waiting in Outpatient Departments: A Survey of Outpatient Appointment Systems* (Oxford University Press, London, 1965)
9. Partridge JW (1992) Consultation time, workload, and problems for audit in outpatient clinics, *Archives of Disease in Childhood* 67: 206–210
10. Ho C, Lau H (1992) Minimizing total cost in scheduling outpatient appointments. *Management Science* 38: 1750–1764
11. Wang PP (1997) Optimally scheduling N customer arrival times for a single-server system. *Computers & Operations Research* 24: 703–716.
12. Robinson LW, Chen RR (2003) Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Transactions* 35: 295–307
13. Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* 35: 1003–1016.
14. Soriano A (1966) Comparison of two scheduling systems. *Operations Research* 14: 388–397
15. Blanco White MJ, Pike MC (1964) Appointment systems in outpatients' clinics and the effect on patients' unpunctuality. *Medical Care* 2: 133–145
16. Vissers J, Wijngaard J (1979) The outpatient appointment system: Design of a simulation study. *European Journal of Operational Research* 3: 459–463
17. Yang KK, Lu ML, Quek SA (1998) A new appointment rule for a single-server, multiple-customer service system. *Naval Research Logistics* 45: 313–326.
18. Rising E, Baron R, Averill B (1973) system analysis of a university health service outpatient clinic. *Operations Research* 21: 1030–1047
19. Swartzman G (1970) The patient arrival process in hospitals: Statistical analysis. *Health Services Research* 5: 320–329
20. Swisher JR, Jacobson SH, Jun JB, Balci O (2001) Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations Research* 28: 105–125.
21. Fetter R, Thompson J (1966) Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research* 1: 66–90
22. Shafer SM, Smunt TL (2004) Empirical simulation studies in operations management: Context, trends, and research opportunities. *Journal of Operations Management* 22: 345–354
23. Law AM, Kelton WD (1991) *Simulation Modeling and Analysis* (McGraw-Hill, New York).