PRODUCTION AND OPERATIONS MANAGEMENT
Vol. 12, No. 4, Winter 2003
Printed in U.S.A.

OUTPATIENT SCHEDULING IN HEALTH CARE: A REVIEW OF LITERATURE*

TUGBA CAYIRLI AND EMRE VERAL
Hofstra University, Frank G. Zarb School of Business, 134 *Hofstra University,*
Department of Management, Hempstead, New York 11549
Baruch College, CUNY, Zicklin School of Business, 17 *Lexington Avenue, Department*
*of Management B*9–240, *New York, New York* 10010

This paper provides a comprehensive survey of research on appointment scheduling in outpatient services. Effective scheduling systems have the goal of matching demand with capacity so that resources are better utilized and patient waiting times are minimized. Our goal is to present general problem formulation and modeling considerations, and to provide taxonomy of methodologies used in previous literature. Current literature fails to develop generally applicable guidelines to design appointment systems, as most studies have suggested highly situation-specific solutions. We identify future research directions that provide opportunities to expand existing knowledge and close the gap between theory and practice.

(HEALTH CARE; SERVICE OPERATIONS; APPOINTMENT SCHEDULING; OUTPATIENT SERVICES; QUEUING SYSTEMS; SIMULATION)

1. Introduction

Health care providers are under a great deal of pressure to reduce costs and improve quality of service provided. In recent years, given the greater emphasis on preventive medicine practices and the shorter lengths of stay, outpatient services are gradually becoming an essential component in health care. Hospitals that cannot make their outpatient departments more cost-effective find themselves in financially unviable positions in this fast-growing industry (Goldsmith 1989).

Patient waiting times and waiting-room congestion are two of the few tangible quality elements. Well-designed appointment systems (AS) have the potential to increase the utilization of expensive personnel and equipment-based medical resources as well as reducing waiting times for patients. Surveys indicate that excessive waiting time is often the major reason for patients' dissatisfaction in outpatient services (Huang 1994), and reasonable waiting times are expected in addition to clinical competence (Jackson 1991).

The goal of this paper is to provide an extensive review of the literature on appointment scheduling for outpatient services. This topic has attracted the interest of many academicians and practitioners over the last 50 years, starting with the pioneering works of Bailey (1952) and Lindley (1952). This review focuses on outpatient scheduling, with limited reference to the most relevant literature on surgical scheduling, as there is some overlap between the two

* Received June 2000; revisions received October 2001 and October 2002; accepted April 2003.

1059-1478/03/1204/519$1.25

topics. The interested reader is referred to reviews of Magerlein and Martin (1978) and Przasnyski (1986) on the latter topic.

The remainder of this paper is organized as follows: in Section 2, we define and formulate the problem of outpatient scheduling; Section 3 focuses on performance criteria used to evaluate AS. In Section 4, we present a classification of AS that have been studied in the literature; followed by Section 5, where analysis methodologies are discussed. Finally, Section 6 presents conclusions and suggestions for future research directions.

2. Problem Definition and Formulation

The objective of outpatient scheduling is to find an appointment system for which a particular measure of performance is optimized in a clinical environment—an application of resource scheduling under uncertainty. The underlying problem applies to a wide variety of environments, such as general practice patient scheduling, scheduling patients for hemodialysis, radiology scheduling, surgical scheduling, etc. Literature on appointment scheduling can be classified into two broad categories: static and dynamic. In the static case, all decisions must be made prior to the beginning of a clinic session, which is the most common appointment system in health care. Thus, it is not surprising to see that most of the literature concentrates on the static problem. Some papers, however, also consider the dynamic case, where the schedule of future arrivals are revised continuously over the course of the day based on the current state of the system (Fries and Marathe 1981; Liao, Pegden, and Rosenshine 1993; Liu and Liu 1998b). This is applicable when patient arrivals to the service area can be regulated dynamically, which generally involves patients already admitted to a hospital or clinic.

Outpatient clinics can be regarded as queuing systems, which represent a unique set of conditions that must be considered when designing AS. The simplest case is when all scheduled patients arrive punctually at their appointment times and a single doctor serves them with stochastic processing times. The formulation gets more complicated as multiple doctors and multiple services are considered. Presence of unpunctual patients, no-shows, walk-ins, and/or emergencies may intervene to upset the schedule. Furthermore, doctors may be late to start a clinic session or they may be interrupted during the course of the day due to activities not directly related to consultation. These environmental factors are discussed in detail next.

2.1. *Number of Services*

Almost all studies in the literature model a single-stage system where patients queue for a single service. A few simulation studies investigate clinic environments where a patient may pass through facilities such as registration, pre-examination, post-examination, x-ray, laboratory, checkout, etc. (Rising, Baron, and Averill 1973; Cox, Birchall, and Wong 1985; Swisher, Jacobson, Jun, and Balci 2001). In such multi-stage models, the patient flow (transition) probabilities associated with each facility need to be specified for Markov Process modeling.

2.2. *Number of Doctors*

Most literature has focused on single-server systems for appointment scheduling. As simple as it may look at first glance, common practices indicate that doctors usually have their own list of patients in clinics. Although queuing theory proves that a single common queue results in shorter wait-times, in most medical services, systems designed around random assignment of doctors are undesirable, as they fail to provide a one-to-one doctor–patient relationship. Given this psychological effect, both practitioners and researchers generally employ independent queues for each doctor (Rising et al. 1973; Cox et al. 1985). On the other hand, some public clinics do not give appointments for specific patients, sending them to the first available doctor. This is the case in clinics studied by Babes and Sarma

(1991) in Algeria and Liu and Liu (1998a, 1998b) in Hong Kong. Representing an even more complex system, Swisher et al. (2001) also model assignment of different types of medical staff members, assigning each patient category a probability of requiring a particular type/skill of staff.

2.3. *Number of Appointments per Clinic Session*

Vissers (1979), Heaney, Howie, and Porter (1991)), and Meza (1998) report a positive relationship between waiting times and the number of appointments in a clinic session (N). Also, studies by Welch and Bailey (1952), Vissers and Wijngaard (1979), and Ho and Lau (1992) cite the importance of including this factor when comparing scheduling rule performance. In addition, the Ho and Lau (1992) study finds that the effect of N is mitigated by no-shows and variability of consultation times, and thus cannot be easily generalized.

2.4. *The Arrival Process*

The arrival characteristics of patients to the clinic are comprised of the following factors, which affect appointment system performance:

i. Unpunctuality of patients can be defined as the difference between a patient's appointment time and actual arrival time. Empirical evidence suggests that patients arrive early more often than late (Fetter and Thompson 1966; Villegas 1967; Babes and Sarma 1991; O'Keefe 1985; Brahimi and Worthington 1991b; Klassen and Rohleder 1996; Lehaney, Clarke, and Paul 1999). As Welch and Bailey (1952) point out, patient earliness may also be undesirable, since it creates excessive congestion in the waiting area.

Some authors model patient punctuality by fitting theoretical probability distributions to empirically derived histograms of patient arrival times relative to their appointment times (Blanco White and Pike 1964; Fetter and Thompson 1966; Swartzman 1970; Cox et al. 1985). Vissers and Wijngaard (1979) combine patient and doctor unpunctuality under one variable called "system earliness." In the queuing models of Mercer (1960, 1973), patient lateness is modeled as an independent random variable with a certain limit on maximum lateness. In all these studies, it is assumed that patients' unpunctuality is independent of their scheduled appointment times.

ii. Presence of no-shows is moderately studied in the literature using no-show probabilities (p) that range from 5 to 30 percent. Empirical data suggest differences among specialties in terms of no-show probabilities observed (Nuffield Provincial Hospitals Trust 1965). As might be expected, studies find that larger p's increase the risk that the doctor will stay idle and decrease the waiting time of patients. Ho and Lau's (1992) assessment of three environmental factors (no-show probability, variability of service times, and number of patients per clinic session) reveals that, among the three, no-show probability is the major one that affects the performance and the choice of an AS.

Given that no-shows pose important problems for health-care administrators, many studies have attempted to investigate possible variables (such as age, socioeconomic level, etc.) that might affect patient attendance, and some identify policies aimed at discouraging no-shows. Interested readers are referred to reviews of Deyo and Inui (1980) and Barron (1980). Schafer (1986) discusses some policies that are found to be useful in dealing with latecomers and no-shows in a private clinic.

iii. Presence of walk-ins (regular and emergency) is neglected in most studies. In the U.K., hospital clinics are primarily used for consultation services for patients referred to them by the general practitioner outside the hospital, and walk-ins are rarely accepted. Thus, the reported walk-in rates are very low in the Nuffield studies of England (1965). However, in the U.S., some clinics are the patient's general practitioner, and are responsible for the patient's total care, whether elective or emergent. Therefore, walk-ins must be anticipated and planned for in the administration of clinic sessions. Similar to no-shows, walk-in probabilities are observed to vary across specialties (Fetter and Thompson 1966; Shonick and Klein 1977; Field 1980).

Swartzman (1970) presents a statistical analysis of the arrival pattern based on data collected from a Michigan Hospital, and finds that arrival rates of emergency patients and walk-ins differ significantly throughout the day, but not from day to day. He concludes that the Poisson distribution offers an acceptable representation (i.e., inter-arrival times are distributed negative-exponentially). Similarly, Rising et al. (1973) model walk-ins using negative exponential distribution to represent inter-arrival times, with the mean value changed on an hourly basis to reflect the seasonal pattern. Walter (1973) finds that when the proportion of patients with appointments increase (that is, the probability of walk-ins decrease), efficiency improves through the reduction in either the doctor's idle time, the patients' waiting time, or both, depending on the number of patients seen in the session (N). Vissers and Wijngaard (1979) model the impact of no-shows and walk-ins on the mean and variance of consultation times. Swisher et al. (2001) use exponential arrival rate for walk-ins based on their observation of a family clinic. None of these studies models balking or reneging behavior of walk-ins.

In the outpatient literature, there is even less focus on emergencies. These are a special type of walk-ins that require immediate medical attention and may possibly preempt the current consultation. Fetter and Thompson (1966) and Rising et al. (1973) include non-preemptive emergencies in their simulation models.

iv. Presence of companions may also be included when modeling the arrival process. Companions are those who accompany a patient to the clinic (e.g., a patient's child, husband, wife, etc.). Although they do not receive the service, they do utilize the waiting room and disregarding them may lead to misleading results for determining the appropriate size of a clinic's waiting room area (Swisher et al. 2001). In that case, the probability of a patient arriving with companion(s) needs to be determined. Differences among specialties are highly possible; for example, in a pediatrics or mental health clinic, all patients are expected to arrive with at least one companion.

2.5. *Service Times*

Service (or consultation) time can be defined as the sum of all the times a patient is claiming the doctor's attention, preventing him/her from seeing other patients (Bailey 1952). The majority of the studies assume patients are homogeneous for scheduling purposes, and use independently and identically distributed (i.i.d.) service times for all patients. Other studies that consider AS with unique patient classes model independently and distinctly distributed (i.d.d.) service times. The general assumption of independence between the arrival and the service patterns may be questionable. In practice, doctors may increase their service rate, if only subconsciously, during peak hours knowing that there are many patients waiting. This is observed to be the case in a number of studies (Bailey 1952; Rockart and Hofmann 1969; Rising et al. 1973; Babes and Sarma 1991).

A variety of service time distributions are chosen in the studies (see Appendix A). Some use empirical data collected from the clinics investigated, and the frequency distributions of observed service times display forms that are unimodal and right-skewed (Welch and Bailey 1952; Jackson 1964; Rising et al. 1973; Buchan and Richardson 1973; Cox et al. 1985; Brahimi and Worthington 1991b; Meza 1998). Most analytical studies use Erlang or exponential service times to make their models tractable.

The coefficient of variation, which is the standard deviation divided by the mean ($CV = \sigma/\mu$), is a commonly used measure for the variability of consultation times. Empirical studies report CV values that range from approximately 0.35 to 0.85 (Bailey 1952; Blanco White and Pike 1964; Rising et al. 1973; O'Keefe 1985; Brahimi and Worthington 1991b; Meza 1998).

Denton and Gupta (2001) find that optimal solutions, although mostly dependent on mean and variance, may exhibit some dependence on higher moments such as skewness. On the other hand, some report that the *relative* performance of AS is not affected by skewness and kurtosis, but only by the mean and variance (Ho and Lau 1992; Yang, Lu, and Quek 1998).

Robinson and Chen (2001) show that the means of the service times can be removed from the formulation without affecting the problem.

A number of studies report that high variability of service times deteriorates both the patients' waiting times and the doctor's idle time (Bailey 1952; Blanco White and Pike 1964; Vissers and Wijngaard 1979; Ho and Lau 1992; Klassen and Rohleder 1996; Denton and Gupta 2001). Similarly, in his analysis of the effects of *CV*, Wang (1997) indicates that, the larger the *CV*, the smaller the optimal appointment intervals, and the higher the costs due to uncertainty created in the system.

In general, studies that evaluate the effect of service-time duration find that shorter mean consultation times result in lower patient waiting times (Bailey 1952; Blanco White and Pike 1964; Walter 1973). Support mechanisms that provide rapid access to clinical information (internal medical records, lab reports, etc.) may be used to reduce the mean and the variability of consultation times (Dexter 1999).

Bailey (1952) reports that the performance of the system is very sensitive to even small changes in appointment intervals. Thus, it is also important to tailor AS to individual doctors, as some studies find that doctor style is a predictor of consultation time.

2.6. *Lateness and Interruption Level of Doctors*

Doctors' unpunctuality, measured as lateness to first appointment, is considered by Blanco White and Pike (1964), Fetter and Thompson (1966), Vissers (1979), Mahachek and Knabe (1984), Babes and Sarma (1991), and Liu and Liu (1998a, 1998b). Agreement among all studies is that patient waiting times are highly sensitive to this factor. If the doctor does not start the clinic on time, a delay factor builds up from the start that ripples throughout the clinic session.

Another doctor-related factor is the interruption level (also called the "gap times"). These include all activities during the session that may require doctor's attention, such as interactions with support staff, phone calls, writing up notes, comfort breaks, etc., which interrupt consultation. Rising et al. (1973) and Lehaney et al. (1999) include non-preemptive gap times in their simulation model by assuming that interruptions occur only in between consultations.

Game theory may be useful in modeling patient and doctor arrivals by considering the conflicting interests of both parties. It is likely that patients arrive early to "beat" the system or arrive late knowing that they will have to wait anyway. Similarly, doctors may arrive late, being afraid that the first patient will be late. There should be either some sort of mechanisms to enforce punctuality, or the AS should be designed to account for all parties' behavior (Van Ackere 1990). One might expect that when clinics are run under more credible AS, both patients and doctors will become more punctual.

2.7. *Queue Discipline*

Almost in all studies, it is assumed that arriving patients are served on a first-come, first-served (FCFS) basis. Given punctual patients, this queue discipline is identical to serving patients in the order of their appointment times. However, unpunctuality may cause changes in the actual order of seeing patients, as doctors would not keep idle waiting for the next appointment in the presence of other waiting patients.

A clinic which deals with walk-ins, emergencies, and/or second consultations, (i.e., those patients returning from the lab, x-ray, etc. visited after an initial consultation) needs to set a priority rule, which determines the order in which these patients will be seen. In general, the first priority is given to emergencies, followed by second consultations, then scheduled patients; the lowest priority is given to walk-ins that are seen on a FCFS basis (Rising et al. 1973; Cox et al. 1985). In practice, it is not uncommon for patients to be called in the order of arrival even when there is an AS, probably because of the ease of administration. However, this may destroy the whole purpose of an AS, and may lead to patients ignoring appointments and coming earlier than necessary. It is more fair if the scheduler maintains a policy of calling

TABLE 1

Problem Definition and Formulation

1. Nature of Decision-Making
   1.1 Static
   1.2 Dynamic
2. Modeling of Clinic Environments
   2.1 Number of services (Single or Multi-stage)
   2.2 Number of doctors (Single or Multi-server)
   2.3 Number of appointments per clinic session
   2.4 Arrival process (Deterministic or Stochastic)
      2.4.1 Punctuality of patients
      2.4.2 Presence of no-shows
      2.4.3 Presence of regular and emergency walk-ins (Preemptive or Non-preemptive)
      2.4.4 Presence of companions
   2.5 Service times (Empirical or Theoretical distribution)
   2.6 Lateness of doctors and their interruption levels (i.e. gap times) (Preemptive or Non-preemptive)
   2.7 Queue discipline (FCFS, by appointment time, by priority)

patients in the order of appointments, while trying to fit in walk-ins and late patients as early as possible.

Table 1 summarizes the relevant factors that are encountered in appointment scheduling environments.

3. Measures of Performance

There is a variety of performance criteria used in the literature to evaluate AS (see Table 2). Studies often list results in terms of the mean waiting time of patients *E(W)*, and the mean idle time of doctor *E(I)*, and/or the mean overtime of doctor *E(O)*, but a "reasonable"

TABLE 2

Performance Measurements Used in the Literature

1. Cost-Based Measures
   Mean total cost calculated using relevant combinations of:
   1.1 Waiting time of patients
   1.2 Flow time of patients
   1.3 Idle time of doctor(s)
   1.4 Overtime of doctor(s)
2. Time-Based Measures
   2.1 Mean, maximum, and frequency distribution of patients' waiting time
   2.2 Mean, variance, and frequency distribution of doctor's idle time
   2.3 Mean, maximum and standard deviation of doctor's overtime
   2.4 Mean and frequency distribution of patients' flow time
   2.5 Percentage of patients seen within 30-minutes of their appointment time
3. Congestion Measures
   3.1 Mean and frequency distribution of number of patients in the queue
   3.2 Mean and frequency distribution of number of patients in the system
4. Fairness Measures
   4.1 Mean waiting time of patients according to their place in the clinic
   4.2 Variance of waiting times
   4.3 Variance of queue sizes
5. Other
   5.1 Doctor's productivity
   5.2 Mean doctor utilization
   5.3 Delays between requests and appointments
   5.4 Percentage of urgent patients served
   5.5 Likelihood of patients receiving the slots they requested
   5.6 Clinic effectiveness

trade-off level between them is to be decided subjectively by the decision-maker. One can give them relative weights in terms of the cost of patients' waiting time (C_p), cost of doctor's idle and overtime (C_d, C_o). Then the objective becomes minimizing the expected total cost of the system represented as:

$$Min\ E(TC) = E(W)C_p + E(I)C_d + E(O)C_o \tag{1}$$

3.1. *Cost-Based Measures*

Studies use different subsets or variations of the cost function shown in Equation 1. The majority include only the patients' waiting time and the doctor's idle time. Others use patients' flow time instead of patients' waiting time (see Appendix A for details). The general cost function assumes a linear relationship between the waiting cost and the waiting time of the patient. However, as pointed out by Klassen and Rohleder (1996), a system where one patient waits 40 minutes is not the same as one in which 20 patients wait 2 minutes each. And the fact that relative costs may differ from one patient to another complicates the issue further. In the literature, studies assume identical waiting costs for all patients. When modeling unpunctual patients and/or walk-ins, the assumption of homogeneous waiting costs may need to be relaxed. Late patients may consider some additional waiting as normal, being partly their own fault. Similarly, walk-ins may tolerate longer waits compared with scheduled patients. For regular patients, there might be a threshold over which patients' tolerance declines steeply. Some survey results indicate that tolerance diminishes after about 30 minutes (Westman, Andersson, and Fredriksson 1987; Huang 1994). In the U.K., hospitals are rated each year according to a national standard set by the Ministry of Health that requires 75 percent of the patients to be seen within 30 minutes of their appointment time (Department of Health 1991).

From a decision-making point of view, it is sufficient to come up with relative values for these costs. For example, estimates of C_d/C_p and/or C_o/C_p ratios, but not the actual monetary values of C_d, C_p, and C_o are needed. The relative cost ratios of C_d/C_p considered in the studies range from 1 to 100. As Fries and Marathe (1981) point out, it is easier to estimate the costs relative to the server, which are usually available via standard cost accounting, but the costs of waiting involve a different type of analysis where the issues of goodwill, service, and "costs to the society" place a value on patients' waiting time. Keller and Laughhunn (1973) divide the annual salary of the doctor by the hours worked per year to estimate C_d and use the minimum wage to reflect the opportunity cost of the patients' waiting time. It is generally assumed that $C_d > C_p$; this is because C_d includes not only the cost of the idle doctor but also the cost of the idle facility (Yang et al. 1998).

3.2. *Time-Based Measures*

It is usually desirable to evaluate waiting time, idle time, and overtime measures separately, as there may be a maximum acceptable level for each. A common approach is to calculate the "*true*" waiting time of patients by subtracting the greater of {appointment time, arrival time} from the consultation start time. This excludes any waiting prior to appointment time, because additional waiting due to early arrival is voluntary and is not a consequence of the AS. True waiting times will be negative if patients are served before their appointment times, which may help the decision-maker to capture information regarding the benefit patients receive by being seen earlier. However, if one wants to focus on positive waiting times only, then negative values need to be truncated at zero. *Flow-time* is another patient-related measure, which is the total time a patient spends in the clinic, including the service time. Since patients generally do not mind time spent in service, most of the literature focuses on waiting time, rather than the flow time.

Idle time of a doctor is the total time during the clinic session when s/he is not consulting because there are no patients waiting to be seen. *Overtime* is calculated as the positive difference between the "desired" completion time of the clinic session and the actual end of service for the last patient. The desired end time for the clinic may be set by accounting for the additional tasks the doctor needs to complete before s/he can leave the office (e.g., writing patient charts, meetings with colleagues, etc.). Yet, it is possible that these tasks are partially handled during the course of the day whenever the doctor stays idle. In general, the negative overtime value can be considered as a part of idle time.

3.3. *Congestion Measures*

Congestion in the clinic hurts service quality from many different perspectives. Apart from taking up valuable space, when queues get excessively long, doctors may increase their service rate or they may be forced to call back some patients at another time. Main measure of congestion is the mean number of patients in the queue (or system).

3.4. *Fairness Measures*

Some studies pay attention to the "fairness" issue, which is the uniformity of performance of an AS across patients. In fixed-interval AS, each successive patient is expected to have, on average, a longer wait time due to the congestion that tends to build up over time. Not only do waiting times increase, but also consultation times tend to decrease as doctors speed up when they progressively fall behind schedule (Heaney et al. 1991). Therefore, patients at the end of the clinic session generally get the worst combination of long waiting times and truncated consultation times, unless an adjustment is made to the AS to account for this phenomenon. Bailey (1952) measures the mean waiting times of patients according to their place in the clinic session (1st, 2nd, etc.); Yang et al. (1998) measure the uniformity of waiting times, and Cox et al. (1985) compare AS based on the variance of queue sizes over the duration of the clinic session.

3.5. *Other Measures*

Other measures used to evaluate AS include doctor's productivity (i.e., number of patients seen in a session), mean doctor utilization, delays between requests and granted appointments, percentage of urgent patients served, and likelihood of patients receiving the slots that they requested. Swisher et al. (2001) use a measure called "clinic effectiveness," which encompasses both clinic profits (revenues and expenses) and patient waiting time on a dollar scale. Table 2 classifies the most commonly used performance measures used in the literature.

4. Designing an Appointment System

The AS design can be broken down into a series of decisions regarding: (1) the appointment rule, (2) the use of patient classification, if any, and (3) the adjustments made to reduce the disruptive effects of walk-ins, no-shows, and/or emergency patients.

4.1. *Appointment Rules*

The appointment rule used to schedule patients can be described in terms of three variables:

i. block-size (n_i) is the number of patients scheduled to the *i*th block. Patients can be called individually, in groups of constant size, or in variable block sizes.

ii. begin-block (n_1), also called the *initial block*, is the number of patients given an identical appointment time at the start of a session.

iii. appointment interval (a_i) is the interval between two successive appointment times, also called *"job allowance."* Appointment intervals can be constant or variable. A common practice is to set them equal to some function of the mean (and sometimes the standard deviation) of consultation times.

Any combination of these three variables (n_i, n_1, a_i) is a possible appointment rule. So far, the following appointment rules have been investigated in the literature (see Figure 1).

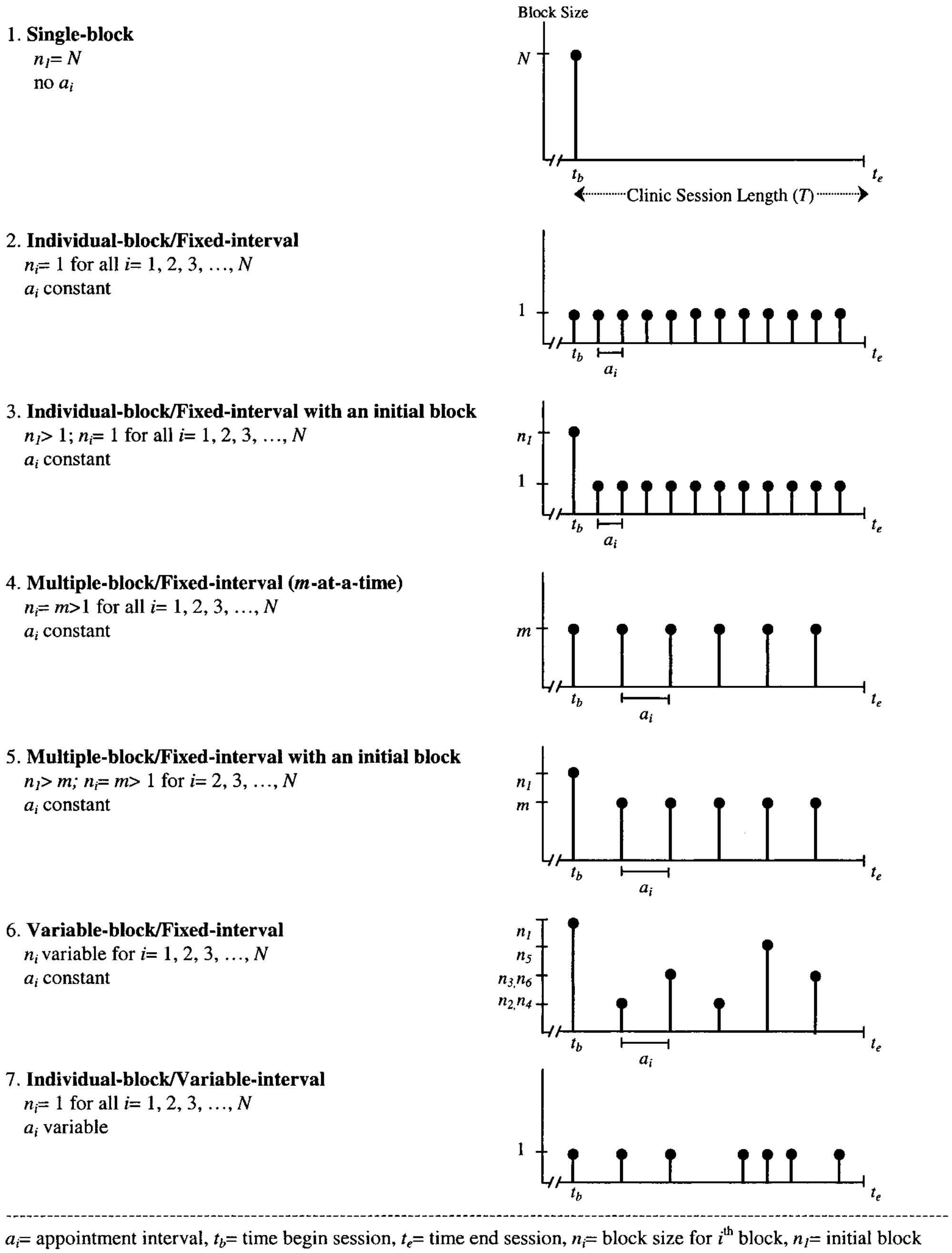

1. Single-block
$n_1 = N$
no a_i
Block Size
N
t_b
t_e
Clinic Session Length (T)
2. Individual-block/Fixed-interval
$n_i = 1$ for all $i = 1, 2, 3, \ldots, N$
a_i constant
1
a_i
3. Individual-block/Fixed-interval with an initial block
$n_1 > 1$; $n_i = 1$ for all $i = 1, 2, 3, \ldots, N$
a_i constant
n_1
4. Multiple-block/Fixed-interval (m-at-a-time)
$n_i = m > 1$ for all $i = 1, 2, 3, \ldots, N$
a_i constant
m
5. Multiple-block/Fixed-interval with an initial block
$n_1 > m$; $n_i = m > 1$ for $i = 2, 3, \ldots, N$
a_i constant
6. Variable-block/Fixed-interval
n_i variable for $i = 1, 2, 3, \ldots, N$
a_i constant
n_1
n_5
n_3, n_6
n_2, n_4
7. Individual-block/Variable-interval
$n_i = 1$ for all $i = 1, 2, 3, \ldots, N$
a_i variable
a_i= appointment interval, t_b= time begin session, t_e= time end session, n_i= block size for i^{th} block, n_1= initial block

FIGURE 1. Appointment rules in the literature (adapted from Fries and Marathe 1981).

1. Single-block rule assigns all patients to arrive as a block at the beginning of the clinic session. For example, all morning patients are scheduled for 9:00 a.m. and they are seen on a first-come, first-served basis. This is the most primitive form of AS, where patients are assigned a “date-only,” rather than a specific appointment slot. Clearly, single-block systems will lead to excessive waiting times for patients, while ensuring that doctors do not stay idle. This was the common practice in most clinics in the 1950s, when the research on outpatient scheduling initiated. Thus, we see that most of the earlier studies praise the advantages of individual appointments, pioneering the shift from single-block to individual-block systems (Lindley 1952; Bailey 1952; Welch 1964; Fry 1964; Johnson and Rosenfeld 1968; Rockart and Hofmann 1969). Single-block systems are still used, mostly in public clinics, probably because they require the least administrative effort. Babes and Sarma (1991) investigate a public clinic in Algeria that uses a single-block AS.

2. *Individual-block/Fixed-interval rule* assigns each patient unique appointment times that are equally spaced throughout the clinic session. A number of studies investigate this type of an appointment rule (Fetter and Thompson 1966; Klassen and Rohleder 1996; Rohleder and Klassen 2000).

3. *Individual-block/Fixed-interval rule with an initial block* is a combination of the previous rule with an initial group of n_1 patients ($n_1 > 1$) called at the start of the clinic session. The goal is to keep an inventory of patients so that the doctor's risk of staying idle is minimized if the first patient arrives late or fails to show up. Bailey (1952, 1954) is the first to suggest an individual-block system with two patients assigned at the beginning of the session and the rest scheduled at intervals equal to the mean consultation time ($n_1 = 2$, $n_i = 1$, $a_i = \mu$). Jansson (1966), Blanco White and Pike (1964), Brahimi and Worthington (1991b), Ho and Lau (1992), and Klassen and Rohleder (1996) evaluate this rule in their comparative analyses.

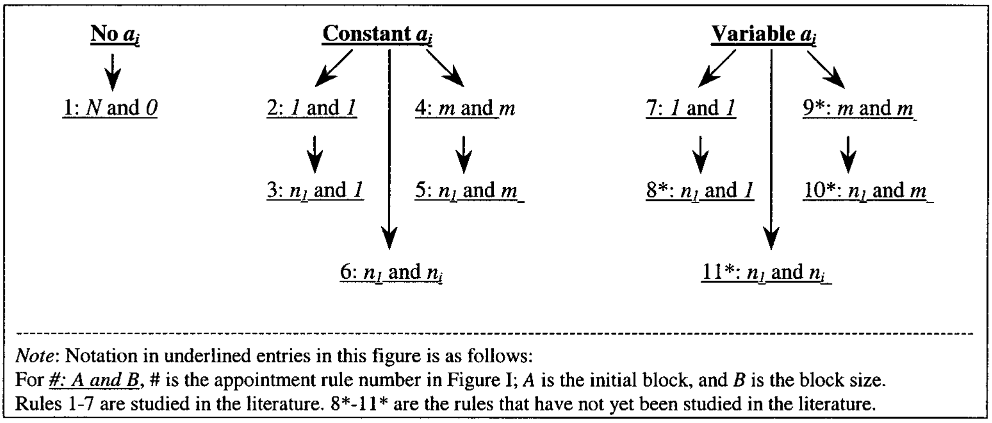
4. *Multiple-block/Fixed-interval rule* is one in which groups of *m* patients are assigned to each appointment slot with appointment intervals kept constant. Soriano (1966) studies an appointment system where patients are called two-at-a-time with intervals set equal to twice the mean consultation time ($n_i = 2$, $a_i = 2\mu$). Blanco White and Pike (1964) and Cox et al. (1985) find that multiple-block rules perform the best in their particular environments. There is a need for more rigorous research that will investigate under what circumstances multiple-block rules might perform better than individual-block rules. As the Nuffield Trust (1965) indicates, it is possible that block-booking is more suitable when the mean consultation times are short, such that patients called for the same time block do not experience excessive waits. There is also some practical advantage in terms of giving patients "rounded" appointment times, such as calling four patients every 15 minutes rather than one every 3.75 minutes (Walter 1973).

5. *Multiple-block/Fixed interval rule with an initial block* is simply a variation of the above system with an initial block ($n_1 > m$). Cox et al. (1985) is the only study that investigates this particular type of rule.

6. *Variable-block/Fixed-interval rule* allows different block sizes during the clinic session, while keeping appointment intervals constant. Villegas (1967), Rising et al. (1973), Fries and Marathe (1981), Liao et al. (1993), Liu and Liu (1998a, 1998b), and Vanden Bosch, Dietz, and Simeoni (1999) investigate this rule in their studies.

7. *Individual-block/Variable-interval rule* is one in which customers are scheduled individually at varying appointment intervals. Ho and Lau (1992) introduce a number of variable-interval rules and test their performance against traditional ones using simulation. They find that, among the rules they tested, increasing appointment intervals toward the latter part of the session improves performance the most. Some recent analytical studies show that, for i.i.d. service times and uniform waiting costs for all patients, optimal appointment intervals exhibit a common pattern where they initially increase toward the middle of the session and then decrease. This is referred to as the "dome" shape, studied by Wang (1997), Robinson and Chen (2001), and Denton and Gupta (2001). In addition, Pegden and Rosenshine (1990), Yang et al. (1998), and Vanden Bosch and Dietz (2000) are some recent studies that analyze individual-block/variable-interval rules.

Figure 2 presents a generalization structure of appointment rules. Rules 1 through 7 are the ones examined in the literature (as summarized in Figure 1). Appointment rules that have not yet been studied in the literature include individual-block/variable-interval rule with an initial block, multiple-block/variable-interval rule with and without an initial block, and variable-block/variable-interval rule (rules 8* through 11*). Note that rule 7 can be considered as subsuming the variable-block rules 6 and 11*, since it is possible to set $a_i = 0$. Also, there are special cases of rule 7, such as the ones studied in Ho and Lau (1992, 1999), which may



Note: Notation in underlined entries in this figure is as follows:
For *#: A and B*, # is the appointment rule number in Figure I; *A* is the initial block, and *B* is the block size.
Rules 1-7 are studied in the literature. 8*-11* are the rules that have not yet been studied in the literature.

FIGURE 2. Generalized structure for appointment rules.

be considered as an intermediate between rules 2 and 7, where $a_1 = a_2 = \ldots = a_k$ are different than $a_{k+1} = a_{k+2} = \ldots = a_{N-1}$ for $k < N - 1$.

4.2. *Patient Classification*

In the majority of the studies, patients are assumed to be homogeneous and they are scheduled on a first-call, first-appointment (FCFA) basis. When there are patient groups (classes) that are known to be distinct in terms of various attributes (e.g., service time characteristics, arrival patterns, costs of waiting, etc.), then this raises the issue whether an AS can be improved by recognizing such differences.

In outpatient scheduling, patient classification can be used for two purposes: to sequence patients at the time of booking; and/or to adjust the appointment intervals based on the distinct service time characteristics of different patient classes. Since the schedule has to be ready in advance and the arriving requests are handled dynamically, the use of patient classification in outpatient settings is somewhat limited. A realistic application requires that the patients are classified into a manageable number of groups and that they are assigned to pre-marked slots when they call for appointments. In the literature, some of the classification schemes used for scheduling purposes include new/return, variability of service times (i.e., low/high-variance patients), and type of procedure. These factors are discussed in Cox et al. (1985), Klassen and Rohleder (1996), Rohleder and Klassen (2000), Lehaney et al. (1999), Lau and Lau (2000), and Vanden Bosch and Dietz (2000). In an application to a radiology department, Walter (1973) investigates the possibility of improving the AS by dividing patients with similar exam times into different sessions. It is found that examination times depend on factors such as patient's age, physical mobility, and type of service. For example, older patients with limited mobility (trolley, wheelchair) require, on average, considerably more time than the younger and walking patients (see Section 5 for detailed discussions of papers).

There is also relevant literature on surgical scheduling literature, which recognizes heterogeneous patients in the context of operating room. In this case, the scheduler estimates surgery durations for *every* procedure individually, and assigns them start times, given the desired sequencing rule (e.g., FCFS, random, in the order of increasing/decreasing mean or variance of service times, etc.). Unlike in outpatient scheduling, the scheduler has a complete list of all requests for the day, and patient availability is guaranteed. Nevertheless, some of the most pertinent papers, such as Charnetski (1984) and Weiss (1990), are summarized in Section 5.

When evaluating AS that use patient classification, several issues need to be considered: First of all, such AS are less flexible than those that assign patients on a FCFS basis, as they limit the number of alternative appointment times that can be offered to patients. For example, a clinic may restrict all new patients to be seen before 10:00 a.m. and all return patients from 10:00 a.m. to 12:00 noon. In this case, it is perfectly possible that a return patient insists on a 9:00 a.m. slot, even though that is reserved for a new patient. Furthermore, even if the ratio of new to return patients may be known over longer term, daily ratios may fluctuate such that the predetermined AS may fail to meet the demand or fulfill the quota of a particular day. When this happens, the scheduler has two options; s/he can either assign a new (return) patient to a different slot, defeating the goal of sequencing, or s/he can postpone that patient to another day. In the latter case, some near-future slots may be left vacant, and delays between the time of request and the appointment may increase. Rohleder and Klassen (2000) address these issues by using secondary performance measures, such as the likelihood of patients receiving the slots that they requested. Vanden Bosch and Dietz (2000) evaluate AS based on the number of slots left vacant, when the scheduler tries to keep delays between requests and appointments at minimum.

4.3. *Adjustments for No-Shows, Walk-Ins, Urgent Patients, Emergencies, and/or Second Consultations*

Whenever relevant, no-shows, walk-ins, urgent patients, and/or emergencies need to be planned for, during the design of an AS. In clinics where second consultations occur frequently, such as in orthopedics, some allowance should be made for the additional demand imposed on doctors (Older 1966). Even though many administrative mechanisms are found to be effective in reducing the likelihood of patients to break their appointments (such as reminders by mail or phone prior to appointment dates, fees for failed appointments, etc.), it is not entirely possible to eliminate no-shows (Barron 1980). On the other hand, strong links are found between a tendency to attend without an appointment and lower social class and perception of urgency by Taylor (1984) and Virji (1990). These findings suggest that a clinic that denies access to walk-ins may further disadvantage these groups. Therefore, in general, a better approach is to anticipate no-shows and walk-ins, and adjust the AS in order to reduce their disruptive effects.

Blanco White and Pike (1964) consider adjustments for no-shows, only. They use simulation to analyze the effects of adding extra patients to make-up for the anticipated average number of no-shows, and find that such an adjustment can considerably improve system performance. Fetter and Thompson (1966) illustrate that it is dangerous to assume walk-ins and no-shows cancel out each other, since they rarely occur in the same volume or at the same time within a session. Therefore, they suggest that the patient load (i.e., percent of available appointments filled) be adjusted based on the expected number of walk-ins and no-shows. Vissers and Wijngaard (1979) introduce a procedure that finds the revised mean and the revised variance of consultation times based on the expected probabilities of no-shows and walk-ins. Using simulation, they illustrate that their method leads to an adequate approximation. In a later paper, Vissers (1979) simulates two options for dealing with no-shows: adding extra patients spread out evenly during the session (called overbooking) vs. shortening appointment intervals proportionally. He finds that the latter approach is slightly better perhaps because of its sustained effect throughout the clinic.

Pierskalla and Brailer (1994) suggest that an AS which considers the stochastic variation of walk-ins (regular/emergency) separately from the stochastic variation of no shows will better achieve improvements in performance. For unplanned walk-in patients, adjustment requires either leaving open slots or setting appointment intervals relatively longer. The former case requires a secondary decision to identify which particular slot(s) to leave open. In their case study, Rising et al. (1973) show that, when walk-ins exist, scheduling appointments to complement the arrival pattern of walk-in patients can smooth the patient flow. They also model second consultations where patients may be sent for an x-ray before seeing the doctor again. Klassen and Rohleder (1996) investigate the best position for leaving open slots for “urgent” patients, who need to be seen within 24 hours. They find no conclusive results; if more urgent slots are left earlier in the session, average patient waiting time is lower and fewer urgent patients are served; whereas if

TABLE 3
Designing an Appointment System

1. Appointment Rule
   - 1.1 Block size
     - 1.1.1 Individual
     - 1.1.2 Multiple
     - 1.1.3 Variable
   - 1.2 Appointment interval
     - 1.2.1 Fixed
     - 1.2.2 Variable
   - 1.3 Initial block
     - 1.3.1 With
     - 1.3.2 Without
   - 1.4 Any combination of the above
2. Patient classification
   - 2.1 None (i.e., all patients assumed homogeneous)
   - 2.2 Use patient classification for:
     - 2.2.1 Sequencing patients at the time of booking
     - 2.2.2 Adjusting appointment intervals to match service time characteristics of patient classes
     - 2.2.3 Any combination of the above
3. Adjustments
   - 3.1 For no-shows
     - 3.1.1 None
     - 3.1.2 Overbooking extra patients to predetermined slots
     - 3.1.3 Decreasing appointment intervals proportionally
   - 3.2 For walk-ins, second consultations, urgent patients, and/or emergencies
     - 3.2.1 None
     - 3.2.2 Leaving predetermined slots open
     - 3.2.3 Increasing appointment intervals proportionally
   - 3.3 Any combination of the above

more slots are left later, doctor idle time is lower and more urgent patients are served. Fetter and Thompson (1966) include emergency breaks in the AS. In practice, when the doctor leaves for an emergency, private clinics usually call patients who have appointments to reschedule (Schafer 1986). This is rarely an option for hospital clinics where patients usually end up waiting longer, unless emergencies are accounted for or handled with other resources.

Table 3 summarizes the relevant decision areas and environmental factors that need to be considered for designing appointment systems.

5. Analysis Methodologies

Research methodologies in appointment scheduling literature can be classified as analytical, simulation-based, or case study, depending on the health-care environment on which they focus, and the assumptions they make.

5.1. *Analytical Studies*

The analytical approaches to the study of AS include queuing theory and mathematical programming methods. Most of the earlier queuing models assume steady-state behavior, which is never reached in a real clinic environment with a small and finite number of patients. Lindley (1952) addresses a G/G/1-type queuing model with a single-server where interarrival times between customers and service times are given by arbitrary distributions. He establishes an elementary relationship between the waiting times of successive customers, which enables him to derive the waiting time distributions of customers. In the conclusion of his paper, he shows that the system improves dramatically when customer arrivals are scheduled at regular intervals as opposed to random arrivals. Jansson (1966) studies a D/M/1 queuing model and derives the total cost distribution function (i.e., waiting and idle costs) for the *k*th customer. The optimal initial block (n_1) and the constant appointment interval (a_i) are determined for a given C_d/C_p ratio so that the mean total cost is minimized. Soriano (1966) compares the steady-state waiting time distribution functions of individual and multiple-block/fixed interval AS for various load factors, assuming deterministic arrivals and gamma service times. Mercer (1960) allows for late arrivals, using a general distribution for lateness. He obtains the steady-state queue length distributions for a single-server system with exponential service times. It is assumed that the patient either arrives at her/his scheduled interval or not at all. Mercer (1973) extends the study to batch arrivals, multi-stage services, and general service times studying a number of different queuing models.

Fries and Marathe (1981) study variable-block/fixed interval AS and compare results with single-block and multiple-block/fixed interval systems. They use dynamic programming to determine the optimal block sizes (n_i) for the next period given that the number of patients remaining to be assigned is known. They present an approximate method to apply the dynamic results to generate a schedule for the static version. Weiss (1990) is the first to address the problem of jointly determining the optimal start times of surgical procedures and the optimal order of those procedures. He presents analytical results for general service times for $N = 2$ and a heuristic solution for larger problems when the goal is to minimize the weighted cost of surgeon waiting time and OR idle time. Regarding the sequencing problem, he proves that the optimal order of two procedures is in increasing variances when service times are exponential or uniform. For larger N-values, the study uses simulation to compare a number of sequencing rules.

Brahimi and Worthington (1991a) study finite capacity multi-server queuing models with nonhomogeneous arrivals (arrival rate dependent on time) and general discrete service time distributions. Their Markov-chain-based algorithm computes time-dependent distributions of the number of customers in the system, from which several key performance measures can be derived. They extend the method to provide approximate results for continuous service time distributions and present results for the transient behavior of systems with constant arrivals. Pegden and Rosenshine (1990) study an S(N)/M/1 model which assumes finite number of scheduled arrivals with distinct inter-arrival times and exponential service times. They prove that the mean waiting time is a convex function of the inter-arrival times for $N \leq 4$, and develop a Markov-chain based procedure to compute the optimal appointment intervals. In a later study, Liao et al. (1993) constrain customer arrivals to fixed lattice of times with k intervals and N number of patients. They use dynamic programming to determine the optimal block sizes when service times are Erlang. The dynamic solution is used as a lower bound to solve the static problem by a branch-and-bound algorithm, which is restricted to small-scale problems. Vanden Bosch et al. (1999) propose a fathoming approach to solve the same problem with customers constrained to lattice points. They show that their fathoming algorithm is more efficient than Liao et al.'s (1993) method for larger problems.

Wang (1993) considers both the static and the dynamic case for a single-server system with exponential service times where the goal is to minimize the weighted sum of customer flow time and system completion time. He shows that customer flow times can be represented by a phase-type distribution which enables using the matrix method to derive the expected flow times. The optimal appointment times are then calculated using a recursive procedure. The results show that the optimal appointment intervals are not constant, but dome-shaped. Wang (1997) extends the study to any service time distribution that can be approximated with a phase-type distribution.

Liu and Liu (1998b) study a queuing system with multiple doctors, where doctor arrival times are random. They develop a dynamic programming formulation to optimally find the block sizes and use the results from the dynamic case to solve the static problem. They compare the performance of schedules obtained using the approximation method to the best ones found by exhaustive simulation. Lau and Lau (2000) address two problems relevant to outpatient and surgical scheduling: (1) How to determine the total system cost given a particular appointment scheduling rule; and (2) How to determine the optimal schedule given a particular sequence of arrivals. They present an efficient procedure to solve the first problem when service times are nonidentically and generally distributed. This leads to solving the second problem by examining a large number of AS and finding the optimal one using a search procedure. They evaluate the accuracy of this approximate method by comparing results with those obtained by simulation. As they note, solving the second problem efficiently can enable one to solve the relevant problem of finding the optimal sequence, which still remains an open issue.

Robinson and Chen (2001) study finding the optimal appointment times when the sequence of *N* patients has already been specified. They formulate the problem as a stochastic linear program, and solve it using Monte-Carlo integration. They use the "dome" structure of the optimal policy as the basis to develop a simple heuristic that adjusts appointment intervals by the relative valuation of C_d/C_p ratio. They show that its performance is robust with regards to distributional misspecification. Denton and Gupta (2001) present a two-stage stochastic linear programming model to determine the optimal appointment intervals, and apply a decomposition approach to solve it for general i.i.d service times. Similar to Wang (1993, 1997) and Robinson and Chen (2001), they show that the optimal intervals are dome-shaped, and note that this is more pronounced for higher values of C_d/C_p ratio.

5.2. *Simulation Studies*

An advantage of simulation modeling over analytical approaches is the ability to model complex outpatient queuing systems and represent environmental variables, such as server or customer-related attributes. Studies conduct simulation experiments to evaluate the performance of alternative AS and/or understand the relationship between various environmental factors and various performance measures. Also, a number of generic simulation modeling packages are developed that enable health care planners and administrators to assess the effectiveness of alternative AS for their particular clinics (Katz 1969; Paul and Kuljis 1995).

Bailey's (1952) is the first study to analyze an individual-block AS at a time when most hospitals were still using single-block systems. He used a manual Monte-Carlo simulation technique in his search for the best initial block (n_1) and appointment interval (a_i) for clinics with a variety of *N*-values. As a result, he concludes that an individual-block/fixed interval AS with an initial block of two patients leads to a reasonable balance between patient waiting time and doctor idle time. This is known as the "Bailey's rule," and it is widely studied in the literature. Blanco White and Pike (1964) relax the assumptions on patient and doctor punctuality when examining the effects of initial block (n_1), number of patients called together (n_i), and appointment interval (a_i). They find that different AS perform better for the two clinics investigated, which face different levels of patient unpunctuality. In their Yale studies, Fetter and Thompson (1966) conduct simulation experiments to analyze the effects of several key variables, such as unpunctuality of patients, lateness of doctors, no-show rates, walk-in rates, appointment scheduling intervals, and patient loads. Their results confirm the importance of doctor punctuality and stress the role of a realistic clinic load in the efficient operation of clinics. Vissers and Wijngaard (1979) reduce the variables essential for modeling AS to five: mean consultation time, coefficient of variation of consultation times, standard deviation of patient's punctuality, number of appointments per session, and mean "system earliness." "System earliness" includes all factors that decrease the risk of idle time of doctors, such as patients' earliness, doctor's lateness, block-booking ($n_i > 1$), initial block-booking ($n_1 > 1$), and setting the appointment intervals smaller than the mean consultation time. In another related study, Vissers (1979) extends the analysis to various *N*-values and develops a heuristic to select a suitable AS, given these five key variables and an acceptable balance between waiting time and idle time.

Charnetski (1984) uses simulation to study the problem of assigning time blocks to surgeons on a first-come, first-served basis when the goal is to balance the waiting cost of the surgeon and the idle cost of the facilities and operation room personnel. The proposed heuristic recognizes that different types of procedures have different service time distributions and sets job allowances based on the mean and the standard deviation of the *individual* procedure times.

Compared with earlier environmental assessment studies, Ho and Lau (1992, 1999) and Ho, Lau, and Li (1995) are the most comprehensive, where they evaluate 50 appointment rules under various operating environments. They introduce a number of individual-block/variable-interval AS and test their performance against the traditional rules. Their best performing variable-interval rule allows patients to arrive in shorter intervals in a session's earlier part, and in larger intervals later on. They conclude that there is no rule that will perform well under all circumstances and propose a simple heuristic to choose an appointment rule for a clinic given p, CV, N, and C_p/C_d ratio. Their procedure for finding the "efficient frontier" provides a unified framework for comparing the performance of AS. They find that p, CV, and N affect AS performance in the order of decreasing importance.

Klassen and Rohleder (1996) introduce AS that classify patients based on their expected service time variability and use simulation to compare alternative ways of sequencing "low" and "high" variance patients when appointment intervals are kept constant. They find that the AS that schedules low-variance patients at the beginning of the session (called the LVBEG rule) performs better than Ho and Lau's best performing rules. They also model urgent patients. In a later study, Rohleder and Klassen (2000) consider the possibility that the scheduler can make an error when classifying patients, and moreover the possibility that s/he cannot sequence patients perfectly when some patients insist on particular slots. They find that the LVBEG rule still performs well under these more realistic assumptions.

Liu and Liu (1998a) study a variable-block/fixed interval AS for a multi-server queuing system where doctors may arrive late. They develop a simulation search procedure to determine the number of patients to schedule to each block (n_i) that will minimize the total cost of patient flow-time and doctors' idle time. Using the properties of the best rules, derived after simulating various environmental factors (number of doctors, no-show probability, number of appointment blocks, and C_d/C_p), they propose a simple procedure to find an appointment rule for a given environment. Yang et al. (1998) propose a heuristic that is presented as a mathematical function of the mean and the standard deviation of consultation times, and the "planning constant k," where k is calculated for a particular clinic environment (i.e., combination of CV, p, N, and C_d/C_p) using a regression model. This rule explicitly tries to be more "fair" by increasing appointment intervals toward the end of the session to avoid compounded waiting times. They use simulation to compare the performance of the heuristic to the best rules proposed by Ho and Lau (1992).

Swisher et al. (2001) provide a discrete-event (visual) simulation model, which can be utilized for decision-making in outpatient services. They apply this model to a family practice clinic and show that the results are very sensitive to changes in the patient mix, patient scheduling, and staffing levels. Regarding scheduling, they only study the effect of changing the time of day a certain patient category is scheduled, rather than comparing different appointment rules.

Our review reveals that, in general, simulation research fails to report the variance-reduction techniques employed and/or the statistical significance of the results. Other simulation studies, although not directly addressing the problem on hand, also offer useful insights on the general design and analysis of outpatient clinics in regards to staffing requirements, facility size/layout, etc. (Stafford and Aggarwal 1979; Taylor III and Keown 1980).

5.3. *Case Studies*

In case studies, the researchers analyze a particular outpatient clinic, make recommendations for improving the existing system, and sometimes evaluate the results of actual implementation. Even though case studies offer valuable insights into how real outpatient clinics function, their major drawback is the lack of generalization.

Villegas (1967) reports a study in the general medicine clinic of an outpatient department, where he experiments with actual practices. He compares the performance of a number of variable-block/fixed-interval AS in terms of patients' waiting times and doctor's idle time. Williams, Covert, and Steele (1967) use simulation to analyze a university clinic to improve patient and doctor scheduling. They show that, when a multiple-block schedule is used, as opposed to a single-block system, patient waiting times decrease substantially with no decrement in staff utilization. Johnson and Rosenfeld (1968) study factors affecting patient waiting times in eight New York City Hospitals using an observational approach. They conclude that the AS in use is a major determinant of waiting times, and both individual and multiple-block systems outperform single-block systems. In their analysis of Massachusetts General Hospital, Rockart and Hofmann (1969) observe that, when clinics shift to individual-block systems, where patients are given unique appointments and are assigned to specific doctors, both doctors and patients behave more punctually and no-show rates decline. Based on the data collected from the same hospital, Hofmann and Rockart (1969) study variables that affect no-show rates in outpatient clinics. One key factor is the "request to appointment" interval; the more time patients have to wait for an appointment, in general, the greater the percentage of no-shows.

Walter (1973) uses simulation to model the queuing system in a radiology department and explores the effects of varying n_1, m, N, CV, and the ratio of patients with appointments (r). He also investigates the effects of dividing the clinic session into more homogeneous groups, and finds that even a simple grouping of inpatients and outpatients results in substantial improvement in doctors' idle time. Rising et al. (1973) use simulation modeling to investigate an outpatient clinic at the University of Massachusetts, with the goal of improving patient and physician scheduling. They suggest scheduling patients in a way that complements the daily and hourly arrival pattern of walk-ins, resulting in a smoothing of the actual arrival rate. Their implementation results suggest improvements in terms of reduced clinic overtime and less waiting time for walk-ins. In a simulation model of an ear, nose, and throat clinic, Cox et al. (1985) evaluate a number of AS with alterations of parameters (a_i, n_1, n_i) and various ways of sequencing new/return patients at the time of booking. They validate their simulation model by comparing results with those observed in real life. When implemented, their proposed rules achieve improved patient flow times, uniform queue sizes, and uniform work rates for doctors. Mahacheck and Knabe (1984) use simulation to analyze alternative operational decisions with respect to patient scheduling, staffing requirements, patient-mix, and facility size. The rules evaluated involve a patient classification scheme of new vs. return patients.

O'Keefe (1985) uses a mainly qualitative approach in his analysis of the operations of three outpatient departments in the U.K. The proposed AS which uses a patient classification of new vs. return patients is rejected by staff who prefer to keep AS simple and uniform across the institution. Similarly, the author faces an enormous resistance by doctors who refuse to change their old habits. This case study is a good illustration of the fact that the real appointment-scheduling problem is primarily a "political" one. Babes and Sarma (1991) investigate a clinic in Algeria which uses a single-block AS, where patients are assigned a certain date with no appointment times specified. Under these conditions, the problem is reduced to determining the number of patients per clinic session (N) and the number of doctors (S) that will optimize the cost performance of the system. They initially apply steady-state queuing theory models of type M/G/S. However when results turn out to be very different than those observed in real operation, they use simulation modeling. They examine the sensitivity of performance parameters to N, S, the lateness of doctors, and the mean service time. Brahimi and Worthington (1991b) apply their previous work on time-dependent-queuing model (Brahimi and Worthington 1991a) to the problem of improving AS in seven clinics in the U.K. They compare alternative systems based on a number of performance measures, and as a result, they suggest an individual-block/fixed-interval system with an initial block of three patients. They observe an improvement in patients' waiting times after the new AS is implemented. Huarng and Lee (1996) use simulation to model the outpatient department of a local hospital in Taiwan that uses no AS, with the goal of improving waiting times and doctor utilization. The authors report that they could not implement an individual-block AS because of staff resistance. Instead, they recommend extending the doctor's work hours in order to better match demand and supply.

Bennett and Worthington (1998) use a systems approach where they consider the interaction of outpatient department with other units in the hospital. They observe that overbooking and scheduling of excessive follow-up appointments create a major capacity problem. Authors' recommendations could not be implemented successfully as the required changes in the behavior of doctors could not be enforced. Lehaney et al. (1999) propose an AS that sorts patients in ascending order of consultation times, similar to sequencing jobs by the shortest processing time (SPT) rule in job shop scheduling. Even though they discuss this approach in the context of an outpatient setting, in a real application the scheduler has to assign a slot to a calling patient without knowing whether the next one will require a shorter or a longer consultation time (strict-ordering schemes are more suited to surgical scheduling, as discussed previously in Section 4). As the authors acknowledge, it is also not practical to come up with reliable estimates of *individual* consultation times. Authors encourage end-user participation in simulation model building in order to increase acceptability of the model, its results, and eventually its implementation. This approach is called "soft-simulation," which combines simulation and soft systems methodology. When implemented as suggested, the new AS improves the performance of the clinic in terms of patient waiting times.

Vanden Bosch and Dietz (2000) examine scheduling/sequencing policies for a specific primary clinic, which uses a classification scheme based on patients' past appointment history or type of procedure (called type A, B, or C patients). This is the first attempt to study the best patient-mix and sequence over several days. They present an analytical approach to solve the static problem where all patients that need to be scheduled for the day are known in advance. They find that there is no easy rule for the optimal sequence; it is difficult to generalize any results on ordering patients by service time means or variances. Furthermore, due to enumeration required, the optimal solution cannot be determined except for very small problems. For the more complex problem of finding the schedule/sequence when requests arrive dynamically, they develop a heuristic policy and test its performance using simulation.

Table 4 summarizes the various research methodologies that address appointment scheduling problems as observed in the literature.

6. Conclusions and Future Research Directions

This paper reviews the literature dealing with outpatient scheduling in health care. Studies have analyzed appointment systems (AS) for effectively regulating the flow of patients so that both patient waiting times and doctor idle times are minimized. Today, despite many published theoretical work, the impact on outpatient clinics has been very limited. The main goal of future research should be to close this gap between theory and practice.

First of all, most studies analyze the environment of a specific clinic, thus their findings lack

TABLE 4
Analysis Methodologies

1. Analytical Studies
    1.1 Queuing theory
    1.2 Mathematical Programming
        1.2.1 Dynamic Programming
        1.2.2 Nonlinear Programming
        1.2.3 Stochastic Linear Programming
2. Simulation Studies
    2.1 Environmental assessment (Which factors affect which performance measures?)
    2.2 Comparison of the performance of alternative AS
3. Case Studies
    3.1 Observational
    3.2 Real-life experimentation of alternative AS
    3.3 Quantitative modeling (simulation, queuing model, etc.) for alternative systems design with or without after-implementation analysis.

generalized applicability. An emerging conclusion is that there is no AS that will perform well under all circumstances, and each situation must be individually considered before an AS can be recommended. The biggest challenge for future research will be to develop easy-to-use heuristics that can be utilized to choose the best AS for individual clinics. Thus, rigorous research is required to reveal the dynamics of a wider range of environmental factors on AS.

Second, there is a need for more realistic representation of outpatient clinics. Queuing models studied in the literature dominantly represent single-server, single-phase systems. The single-server assumption may hold for most cases, since doctors usually have their own list of patients, and sharing patients among multiple doctors is generally avoided in order to maintain continuity of doctor–patient relationship. Multi-phase models that depict the interaction of clinics with various supporting facilities (e.g., lab, x-ray, etc.) are more realistic than the commonly used single-level analysis of individual services. From a research design perspective, more empirical data will be useful for identifying probability distributions that represent actual service times. Even though most analytical studies use exponential service times to make their methods tractable, empirical evidence shows that this assumption is too restrictive and unrealistic. There is still a void in the literature in terms of more realistic arrival patterns that incorporate unpunctual patients, walk-ins, and emergencies. It is important that future studies include these factors and examine the sensitivity of various performance measures. Studying the impact of walk-in seasonality (regular and emergency) may also be an interesting research area, which has a practical bearing on certain practices such as radiology (summer fractures and winter colds), pulmonary specialties (seasonal asthma and allergy agents), etc.

Third, future studies should use multiple measures of performance to evaluate AS, possibly including “fairness” measures as well. Apart from the dominantly used mean measures of performance, it may be important to look at the breakdown of waiting times throughout the clinic session to ensure homogeneity across appointment slots. The common assumption of linearity between waiting cost and waiting time is no longer capturing the complexity of different patients’ attitudes toward waiting. It may be more realistic to use different waiting costs and cost functions for heterogeneous patients.

Fourth, there are still many AS that have not been fully explored. Studies mostly focus on the first decision level in AS design, which is to find the best appointment rule. Recently, there is a lot of interest in variable-interval appointment systems, and future research may continue to investigate other variations with multiple and variable blocks. Some previous findings suggest that it may be advantageous to use patient classification when scheduling outpatients, i.e., relax the assumption of homogeneous patients. So far, studies have only considered the use of patient classification for sequencing patients at the time of booking. However, further improvements are likely when appointment intervals are also tailored according to different patient types. No rigorous research exists which investigates possible approaches to adjusting the AS in order to minimize the disruptive effects of no-shows, walk-ins, and/or emergencies. In short, the biggest challenge for future research will be to find new AS that will improve system performance over a wide range of measures with no trade-offs.

Finally, there is a lack of emphasis on the real-life performance of AS implemented as a result of studies. Discussions on implementation issues reveal how misleading it can be to view the problem as a “pure optimization” problem. Practical issues such as the ease of use of the AS, or implications on modifying physicians’ behavior need to be considered in order to achieve the ultimate goal of improving “real systems.” It may also be interesting to determine what are the most commonly used AS in practice.

Today, health care industry is facing an increasingly competitive marketplace. Patients’ expectations are changing, and surveys indicate that patients choose among providers by their ability to honor their appointment times as well as medical proficiency. Therefore, health care administrators cannot ignore the consequences of poorly designed AS and concentrate only on costs. It is also arguable today that doctor’s time is much more valuable than that of patients, and relative costs of waiting versus idle time to the society need to be reevaluated.

APPENDIX A

Summary of Studies

I—ANALYTICAL STUDIES

Study	Service Time Distribution μ = mean σ = standard deviation C_v = coefficient of variation = σ/μ	Patient Unpunctuality	No-shows p = no-show probability	Walk-ins	Doctors' Lateness and Interruption Level (Gap Times)	Queue Discipline	Performance Measurement	Number of doctors (S) Number of patients per session (N) Duration of session (T)	Appointment System n_i = block size for i^{th} block n_I = initial block a_i = appointment interval
Lindley (1952)	General	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Mean and variance of patient waiting times (2) Probability of not having to wait	$S = 1$	Individual-block/ fixed-interval AS $a_i = h\mu$ where h = 1.2 to 4
Mercer (1960, 1973)	Exponential, General	General lateness distribution Patient scheduled for rth interval arrives (r − 1, r) or not at all; also (r − 1, r + 1)	$p > 0$	None	Punctual	FCFS	(1) Frequency distribution of queue length	Multiple $S \geq 1$	Individual-block/ fixed-interval AS
Jannson (1966)	Exponential $C_v = 1$	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Frequency distribution of total cost of patients' waiting and doctor's idle time	$S = 1$	Individual-block/ fixed-interval AS with an initial-block Solve for optimal n_1 and $a_i = h\mu$ where h = 1.1, 2, 3, 5
Soriano (1966)	Gamma $C_v = 0.50$	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Waiting time distributions of patients	$S = 1$ $N = 8$	Multiple-block/ fixed interval AS $n_i = 2$ and $a_i = 2h\mu$ where h = 1.01 to 2

Fries and Marathe (1981)	Negative exponential	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Mean waiting time of patients (2) Mean idle time (3) Mean overtime of doctor	$S = 1$ $N = 24$	Variable-block/ fixed-interval AS Solve for n_i given a_i constant
Pegden and Rosenshine (1990)	Exponential	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of patients' waiting time and doctor's availability	$S = 1$ $N \leq 3$	Individual-block/ variable interval Solve for optimal a_i where $n_i = 1$
Liao, Pegden, Rosenshine (1993)	Erlang	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of patients' waiting time and doctor's overtime	$S = 1$ $N \leq 12$	Variable-block/ fixed-interval AS Solve for optimal n_i given a_i constant
Brahimi and Worthington (1991a)	Discrete service time distribution is used to approximate general continuous service times	Punctual (for constant arrivals)	None: $p = 0$	None	Punctual	FCFS	(1) Frequency distribution of mean number of customers in the system (2) Frequency distribution of the probability of all servers being busy	Multiple $S = 4$	No specific AS
Wang (1993)	Exponential	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of patients' flow times and doctor's completion time	$S = 1$ $N = 2, 3, 4, 9$	Individual-block/ variable interval Solve for optimal a_i where $n_i = 1$
Wang (1997)	Coxian, Exponential $C_v = 0.85$, 1 and 1.27	Punctual Also, considers the lateness of the first and the last patient	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of patients' flow times and doctor's completion time	$S = 1$ $N < 50$	Individual-block/ variable interval Solve for optimal a_i *where* $n_i = 1$
Liu and Liu (1998b)	Uniform $C_v = 0.33$ Exponential $C_v = 1$ Weibull $C_v = 5$ $\mu = 10$ min.	Punctual	$p = 0, 0.10$	None	Distribution for available service capacity is determined by (1) Poisson approximation, (2) Simulation	FCFS	(1) Total cost of patients' waiting time, doctor's idle time and doctor's overtime $C_d/C_p = 0.5$–1000	Multiple $S = 3$ N is a decision variable	Variable-block/ fixed-interval AS Solve for optimal n_i where a_i. constant What-if analysis to determine a_i

APPENDIX A (*cont'd*)

I—ANALYTICAL STUDIES (cont'd)

Study	Service Time Distribution μ = mean σ = standard deviation C_v = coefficient of variation = σ/μ	Patient Unpunctuality	No-shows p = no-show probability	Walk-ins	Doctors' Lateness and Interruption Level (Gap Times)	Queue Discipline	Performance Measurement	Number of doctors (S) Number of patients per session (N) Duration of session (T)	Appointment System n_i = block size for i^{th} block n_l = initial block a_i = appointment interval
Vanden Bosch, Dietz & Simeoni (1999)	Erlang	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of patients' waiting times and doctor's overtime	$S = 1$ N general	Variable-block/ fixed-interval AS Solve for optimal n_i given constant $a_i = \mu$
Lau & Lau (2000)	General	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of patients' waiting time and doctor's idle time	$S = 1$ $N < 30$	Individual-block/ variable-interval Solve for optimal a_i
Robinson and Chen (2001)	Generalized Lambda distribution fitted to data collected on surgery times	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of patients' waiting time and doctor's idle time C_d/C_p ranges from 1 to 100	$S = 1$ $N = 3, 5, 8, 12, 16$	Individual-block/ variable-interval Solve for optimal a_i
Denton and Gupta (2001)	General; illustrated for Uniform, Gamma and Normal ($C_v = 0.236$)	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Total cost of customers' waiting, server's idle and overtime	$S = 1$ $N = 3, 5, 7$	Individual-block/ variable-interval Solve for optimal a_i

II—SIMULATION STUDIES

Bailey (1952) Welch & Bailey (1952)	Pearson Type III fitted to empirical data C_v = 0.51–0.62 $\mu = 5$ min for $N = 25$ $\mu = 6.25$ for $N = 20$ $\mu = 8.33$ for $N = 15$ $\mu = 12.5$ for $N = 10$	Punctual	None: $p = 0$	None	Punctual	FCFS	(1) Frequency distribution of patients' waiting times (2) Frequency distribution of doctor's idle time (3) Frequency distribution of the number of patients in the queue (4) Mean and standard deviation of clinic completion time (5) Mean waiting time of patients according to their place in the clinic. $C_d/C_p = 37.5$	$S = 1$ $N = 10, 15, 20, 25$ $T = 125$ min.	Individual-block/fixed-interval AS w/an initial-block $n_1 = 2$ $n_2 = 1$ for $i = 2, \ldots, N$ $a_i = \mu$
Blanco White and Pike (1964)	Pearson Type III fitted to empirical data C_v = 0.44–0.70 $\mu = 2.5$ *min for* $N = 60$ $\mu = 3$ for $N = 50$ $\mu = 5$ for $N = 30$ $\mu = 7.5$ for $N = 20$ $\mu = 15$ for $N = 10$	Pearson Type VII fitted to empirical distribution for unpunctuality (mean = 0) based on empirical data	$p = 0, 0.09, 0.19$ Patient load is adjusted according to the observed no-shows.	None	Doctor is late to first appointment 0, 5, 10, 15, 20 min.	FCFS	(1) Mean waiting time of patients (2) Mean idle time of doctors (3) % of patients seen within 30 min. of appointment time	$S = 1$ $N = 10, 20, 30, 40, 50, 60$ $T = 150$ min.	*For punctual case:* Individual-block/fixed-interval AS w/ an initial-block $n_1 = 2$ or 3, $a_i = \mu$ *For unpunctual case:* Multiple-block/Fixed interval AS $a_i = T/10$, $n_i = m$

APPENDIX A (*cont'd*)

II—SIMULATION STUDIES (cont'd)

Study	Service Time Distribution μ = mean σ = standard deviation C_v = coefficient of variation = σ/μ	Patient Unpunctuality	No-shows p = no-show probability	Walk-ins	Doctors' Lateness and Interruption Level (Gap Times)	Queue Discipline	Performance Measurement	Number of doctors (S) Number of patients per session (N) Duration of session (T)	Appointment System n_i = block size for i^{th} block n_l = initial block a_i = appointment interval
Fetter and Thompson (1966)	Empirically collected service times for walk-ins (μ = 9.8 min.) and scheduled patients (μ = 12.6 min.)	Early-late times are allowed to go to maximum 5 minutes.	Observed p-values range from 0.04 to 0.22 depending on clinic type, mean is 0.14	Observed probabilities range from 0.07–0.58 with the mean 0.38	Doctor is late to first appointment 0, 30, 60 min.	FCFS Walk-ins are assigned to first available doctor, Emergencies are nonpreemptive	(1) Mean waiting time of patients (walk-ins vs scheduled patients) (2) Mean idle time of doctors (3) doctors' productivity, i.e. # of patients seen per session	$S = 3$ $N = 26$	Individual-block/ fixed-interval AS a_i = 15 min. and a_i = 20 min.
Vissers and Wijngaard (1979) Vissers (1979)	General C_v = 0.25, 0.5, 0.75, 1, 1.25	Modeled under a new variable called "system earliness" Mean punctuality = 0 to 3μ.	This factor is incorporated by revising the mean and variance of service times	This factor is incorporated by revising the mean and variance of service times	Doctor lateness is modeled under "system earliness"	FCFS	(1) Mean waiting time of patients	$S = 1$ N = 10, 20, 30, 40, 50, 60	Individual-block & Multi-block/ fixed interval AS's with or without an initial block.
Ho and Lau (1992, 1995, 1999)	Uniform C_v = 0.2, 0.5 Exponential C_v = 1	Punctual	p = 0, 0.10, 0.20	None	Punctual	FCFS	(1) Mean waiting time of patients (2) Mean idle time of doctors C_d/C_p = 19.2–29.1	$S = 1$ N = 10, 20, 30	Individual-block/ variable-interval $n_i = 1$ a_i is variable

Klassen and Rohleder (1996) Rohleder & Klassen (2000)	Lognormal dist $\mu = 8, 10, 12$ min. $\sigma = 5, 10$ $C_v = 0.5, 1.0$ In the later study, a client's service time variance is randomly chosen from a lognormal distribution (mean 7.5 min. standard deviation 3.75, 7.5 min.)	Punctual	$p = 0.05$	None	Punctual	FCFS for regular patients Poisson distribution is used to generate urgent calls with a rate of 2 per session	(1) Total cost of patients' waiting and doctor's idle time with C_d/C_p ratio chosen as 1 (2) Mean waiting time of patients (3) Mean idle time of doctor (4) Mean and maximum completion times (5) % of urgent clients served In the later study add: (6) Mean max. waiting time (7) % of waits < 10 min. (8) % of patients who eceive the slot requested	$S = 1$ $T = 210$ min. N varies from 19 to 21, depending on urgent calls received	Individual-block/ fixed-interval $a_i = 10$ min. Two slots left open for urgent calls Use patient classification to sequence low and high-variance patients in various ways
Liu and Liu (1998a)	Uniform $C_v = 0.58$ Exponential $C_v = 1$ Weibull $C_v = 2.236$ $\mu = 10$ min.	Punctual	$p = 0, 0.10, 0.20$	None	Doctors' arrival times uniform over [0, 0] and [0, 6]	FCFS	(1) Total cost of patients' flow-times and doctor's idle time	Multiple $S = 2, 3, 5$ $N = 46$	Variable-block/ fixed-interval AS Solve for optimal n_i given a_i
Yang, Lau & Quek (1998)	Gamma $C_v = 0.2, 0.4, 0.6, 0.8, 1.0$ $\mu = 1$	Punctual	$p = 0, 0.05, 0.10, 0.15, 0.20$	None	Punctual	FCFS	(1) Total cost of patients' waiting and doctor's idle time $C_d/C_p = 1$ to 100. (2) Variance of doctor idle time (3) Variance of patients' waiting times	$S = 1$ $N = 10, 15, 20, 25, 30$	Individual-block/ variable-interval Solve for a_i
Swisher, Jacobson, Jun and Balci (2001)	Exponential	Punctual	None: $p = 0$	Exponential	Punctual	FCFS	(1) Clinic effectiveness (2) Mean doctor utilization (3) Mean overtime	Multiple $S \geq 1$	Scheduling various patient categories on different periods of the day

APPENDIX A (*cont'd*)

III—CASE STUDIES

Study	Service Time Distribution μ = mean σ = standard deviation C_v = coefficient of variation = σ/μ	Patient Unpunctuality	No-shows p = no-show probability	Walk-ins	Doctors' Lateness and Interruption Level (Gap Times)	Queue Discipline	Performance Measurement	Number of doctors (S) Number of patients per session (N) Duration of session (T)	Appointment System n_i = block size for i^{th} block n_l = initial block a_i = appointment interval
Walter (1973)	Gamma $C_v = 0.31–0.86$ $\mu = 2.5$ min for $N = 60$ $\mu = 15$ for $N = 10$	Punctual	None: $p = 0$	Ratio of scheduled to walk-in patients (r) analyzed for $r = 0.5, 1, 2$	Punctual	FCFS	(1) Mean waiting time of patients (2) Mean idle time of doctors	$N = 10, 20, 30, 40, 50, 60$ $T = 150$ min.	Individual-block/fixed-interval AS w/an initial-block $n_1 = 2, 3, 4$; $a_i = \mu$ Multiple-block/Fixed interval AS $m = 2, 3$ Patient classification as outpatients and inpatients
Rising, Baron, Averill (1973)	Separate empirical distributions are derived for scheduled patients ($\mu = 13$, $C_v = 0.75$), walk-ins and ($\mu = 10$, $C_v = 0.51$), second-consultations ($\mu = 5$, $C_v = 0.84$).	Punctual	None: $p = 0$	Negative exponential interarrival times for walk-ins and emergencies (nonpreemptive).	Gap times per day per doctor (in min) 0, 20, 40, 60, 80	1st priority given to emergencies or patients returning from X-ray, etc. 2nd scheduled patients, and 3rd walk-ins.	(1) Mean waiting time of scheduled patients, walk-ins, and second-service patients (2) Frequency distribution of waiting times for all types of patients	Number of doctors scheduled hourly is a decision variable Maximum $S = 7$ N is around 100 $T = 480$ min.	Variable-block/fixed-interval AS Find n_i that best complement the observed arrival pattern of walk-ins $a_i = 60$ min.

Mahacheck & Knabe (1984)	Poisson $\mu = 11$ min.	Punctual	None: $p = 0$		Doctor lateness is modeled based on actual times observed	FCFS	(1) Average flow time of patients (2) Doctor utilization (3) Clinic end time	S and N are decision variables $S = 1, 2, 3$ $N = 20–35$	Individual-block fixed interval AS that uses new/return classification $n_1 = 1$, a_i constant
Cox, Birchall, Wong (1985)	Empirical distribution: X_5^2 for new patients, X_3^2 for return patients, and truncated Normal distribution for the hearing test N (8, 12)	Empirical distribution of arrival time relative to appointment time: Exponential is fitted for late patients, and a probability density function of form ($a + bx$) is fitted for early patients.	None: $p = 0$ Empirically observed p's range from 0.10 to 0.30	None	Increase service times to allow for "gap times"	For consultation: Highest priority given to patients returning from a hearing test, For hearing test: FCFS	(1) Queue sizes at each time of session (2) % doctor is busy (3) Frequency distribution of flow time of patients (4) Frequency distribution of total idle time of doctors (5) Frequency distribution of total waiting time of patients	Multiple 2 doctors, and 2 audiometricians $T = 210$ min. Two clinics: $N_1 = 40$, $N_2 = 63$	Multiple-block fixed interval AS that sequences new/return patients $n_i = 3$, $a_i = 15$ min. Multiple-block/ fixed interval AS w/an initial block $a_i = 15$ min. $n_1 = 3$, $n_i = 5$ for $i = 2, \ldots, N$
O'Keefe (1985)	The distribution for one department is close lognormal. C_v's range from 0.58–0.75.	Observe that 75% of the patients arrive early	Observed $p = 0.05$	None	Observe that doctors are usually late	FCFS	(1) Mean waiting time of patients	Multiple $S = 1, 2, 3$ Three clinics: $N_1 = 68$, $N_2 = 41$, $N_3 = 132$	Individual-block/ fixed-interval AS w/an initial-block $a_i = 15$ min.
Babes and Sarma (1991)	Weibull	Arrivals are Poisson	None: $p = 0$	None	Punctual vs. when doctors start 30 min. late to first appointment	FCFS	(1) Mean number of patients in queue (2) Mean waiting times in queue and in system (3) Time clinic session ends (4) % doctor busy	S and N are decision variables Multiple $S = 1, 2, 3$	Single-block AS Patients are scheduled no appointment times, just the date.

APPENDIX A (*cont'd*)

III—CASE STUDIES (cont'd)

Study	Service Time Distribution μ = mean σ = standard deviation C_v = coefficient of variation = σ/μ	Patient Unpunctuality	No-shows p = no-show probability	Walk-ins	Doctors' Lateness and Interruption Level (Gap Times)	Queue Discipline	Performance Measurement	Number of doctors (S) Number of patients per session (N) Duration of session (T)	Appointment System n_i = block size for i^{th} block n_l = initial block a_i = appointment interval
Brahimi and Worthington (1991b)	Discrete approximation to empirical distributions $C_v = 0.36–0.61$ $\mu = 5.4–7.2$ min.	Punctual	$p = 0, 0.10$	None	Punctual	FCFS	(1) Mean number of patients in the system (2) Mean waiting time in the system (3) % doctor is busy	$S = 1$ Seven clinics: N ranges from 11 to 23	Individual-block/fixed-interval AS w/an initial-block $n_1 = 3$, $a_i = 5$ min. $n_i = 1$ for $i = 2, \ldots, N$
Huarng and Lee (1996)	Exponential Dermatology: $\mu = 1.82$ min., General surgery: $\mu = 2.82$ min., General medicine: $\mu = 2.78$ min.	Punctual	None	None	None	FCFS	(1) Mean waiting time (2) Mean and maximum idle time (3) Mean and maximum # of patients in the queue (4) Mean doctor utilization (5) Mean number of patients served	Multiple $S = 2$ $N = 80–335$ $T = 240–480$ min.	Individual-block/variable-interval $a_i = 1.05\mu$ vs. Single-block system currently used
Lehaney, Clarke, Paul (1999)	Not specified $\mu = 11$ min.	Punctual	$p = 0$	None	Gap times modeled as nonpreemptive in the simulation model	FCFS	(1) Mean patient waiting time (2) Frequency distribution of waiting times (3) # of patients in queue	Multiple $S = 3$ $N = 11$	AS that sequences patients in ascending order of processing times
Vanden Bosch & Dietz (2000)	Truncated Erlang & truncated gamma based on empirical service times $C_v = 0.046$ to 0.163.	Punctual	$p = 0.05–0.11$	None	Punctual	FCFS	(1) Total cost of patients' waiting times and doctor's overtime $C_d/C_p = 3$ (2) Minimize # of vacant slots, while ensuring that delays between requests and appointments are <5 days.	$S = 1$ $T = 170$ min. $N = 6$	Individual-block/variable-interval a_i variable, multiples of 10 min. Use patient classification based on problem type

References

Babes, M. and G. V. Sarma (1991), "Out-Patient Queues at the Ibn-Rochd Health Center," *Journal of the Operational Research Society*, 42, 10, 845–855.

Bailey, N. (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.

———. (1954), "Queuing for Medical Care," *Applied Statistics, A Journal of the Royal Statistical Society* 3, 137–145.

Barron, W. M. (1980), "Failed Appointments: Who Misses Them, Why They Are Missed, and What Can Be Done?" *Primary Care*, 7, 4, 563–574.

Bennett, J. C. and D. J. Worthington (1998), "An Example of a Good but Partially Successful OR Engagement: Improving Outpatient Clinic Operations," *Interfaces*, 28, 5, 56–69.

Blanco White, M. J. and M. C. Pike (1964), "Appointment Systems in Outpatients' Clinics and the Effect on Patients' Unpunctuality," *Medical Care* 2, 133–145.

Brahimi, M. and D. J. Worthington (1991a), "The Finite Capacity Multi-Server Queue with Inhomogeneous Arrival Rate and Discrete Service Time Distribution and Its Application to Continuous Service Time Problems," *European Journal of Operational Research*, 50, 3, 310–324.

——— and ——— (1991b), "Queuing Models for Out-patient Appointment Systems: A Case Study," *Journal of the Operational Research Society*, 42, 9, 733–746.

Buchan, I. C. and I. M. Richardson (1973), *Time Study of Consultations in General Practice*, Scottish Health Service Studies No: 27, Scottish Home and Health Department, Edinburgh.

Charnetski, J. R. (1984), "Scheduling Operating Room Surgical Procedures With Early and Late Completion Penalty Costs," *Journal of Operations Management*, 5, 1, 91–102.

Cox, T. F., J. F. Birchall, and H. Wong (1985), "Optimizing the Queuing System for an Ear, Nose and Throat Outpatient Clinic," *Journal of Applied Statistics*, 12, 113–126.

Denton, B. and D. Gupta (2001), "A Sequential Bounding Approach for Optimal Appointment Scheduling," Unpublished working paper, DeGroote School of Business, McMaster University, Hamilton, Ontario.

Department of Health (1991), *The Patient's Charter*, H. M. S. O., London, U.K.

Dexter, F. (1999), "Design of Appointment Systems for Preanesthesia Evaluation Clinics to Minimize Patient Waiting Times: A Review of Computer Simulation and Patient Survey Studies," *Anesthesia Analgesia*, 89, 925–931.

Deyo, R. A. and T. S. Inui (1980), "Dropouts and Broken Appointments: A Literature Review and Agenda for Future Research," *Medical Care*, 18, 11, 1146–1157.

Fetter, R. and J. Thompson (1966), "Patients' Waiting Time and Doctors' Idle Time in the Outpatient Setting," *Health Services Research*, 1, 66–90.

Field, J. (1980), Problems of Urgent Consultations within an Appointment System," *Journal of the Royal College of General Practitioners*, 30 (March), 173–177.

Fries, B. and V. Marathe (1981), "Determination of Optimal Variable-Sized Multiple-Block Appointment Systems," *Operations Research*, 29, 2, 324–345.

Fry, J. (1964), "Appointments in General Practice," *Operational Research Quarterly*, 15, 3, 233–237.

Goldsmith, J. (1989), "A Radical Prescription for Hospitals," *Harvard Business Review*, 67, 3, 104–111.

Heaney, D. J., J. G. Howie, and A. M. Porter (1991), "Factors Influencing Waiting Times and Consultation Times in General Practice," *British Journal of General Practice*, 41, 315–319.

Ho, C. and H. Lau (1992), "Minimizing Total Cost in Scheduling Outpatient Appointments," *Management Science*, 38, 12, 1750–1764.

———, ———, and J. Li (1995), "Introducing Variable-Interval Appointment Scheduling in Service Systems," *International Journal of Production & Operations Management*, 15, 6, 59–69.

——— and ——— (1999), "Evaluating the Impact of Operating Conditions on the Performance of Appointment Scheduling Rules in Service Systems," *European Journal of Operational Research*, 112, 3, 542–553.

Hofmann, P. B. and J. F. Rockart (1969), "Implications of the No-Show Rate for Scheduling OPD Appointments," *Hospital Progress*, 50, 8, 35–40.

Huang, X. (1994), "Patient Attitude Towards Waiting in an Outpatient Clinic and Its Applications," *Health Services Management Research*, 7, 2–8.

Huarng, F. and M. H. Lee (1996), "Using Simulation in Out-Patient Queues: A Case Study," *International Journal of Health Care Quality Assurance*, 9, 6, 21–25.

Jackson, A. R. (1991), "A Waiting Time Survey in General Practice," *Australian Family Physician*, 20, 12, 1744–1750.

Jackson, R. R. P. (1964), "Design of an Appointment System," *Operational Research Quarterly*, 15, 219–224.

Jansson, B. (1966), "Choosing a Good Appointment System: A Study of Queues of the Type (D/M/1)," *Operations Research*, 14, 292–312.

Johnson, W. L. and L. S. Rosenfeld (1968), "Factors Affecting Waiting Time in Ambulatory Care Services," *Health Services Research*, 3, 4, 286–295.

Katz, J. (1969), "Simulation of Outpatient Appointment Systems," *Communications of the ACM*, 12, 215–222.

KELLER, T. F. AND D. J. LAUGHHUNN (1973), “An Application of Queuing Theory to a Congestion Problem in an Outpatient Clinic,” *Decision Sciences*, 4, 379–394.

KLASSEN, K. J. AND T. R. ROHLEDER (1996), “Scheduling Outpatient Appointments in a Dynamic Environment,” *Journal of Operations Management*, 14, 2, 83–101.

LAU, H. AND A. H. LAU (2000), “A Fast Procedure for Computing the Total System Cost of an Appointment Schedule for Medical and Kindred Facilities,” *IIE Transactions*, 32, 9, 833–839.

LEHANEY, B., S. A. CLARKE, AND R. J. PAUL (1999), “A Case of Intervention in an Outpatients Department,” *Journal of the Operational Research Society*, 50, 9, 877–891.

LIAO, C., C. D. PEGDEN, AND M. ROSENSHINE (1993), “Planning Timely Arrivals to a Stochastic Production or Service System,” *IIE Transactions*, 25, 5, 63–73.

LINDLEY, D. V. (1952), “The Theory of Queues with a Single Server,” *Proceedings Cambridge Philosophy Society*, 48, 277–289.

LIU, L. AND X. LIU (1998a), “Block Appointment Systems for Outpatient Clinics with Multiple Doctors,” *Journal of the Operational Research Society*, 49, 12, 1254–1259.

——— AND ——— (1998b), “Dynamic and Static Job Allocation for Multi-Server Systems,” *IIE Transactions*, 30, 9, 845–854.

MAGERLEIN, J. M. AND J. B. MARTIN (1978), “Surgical Demand Scheduling: A Review,” *Health Services Research*, 13, 4, 418–433.

MAHACHEK, A. R. AND T. L. KNABE (1984), “Computer Simulation of Patient Flow in Obstetrical/Gynecology Clinics,” *Simulation*, 43 (August), 95–101.

MERCER, A. (1960), “A Queuing Problem in Which Arrival Times of The Customers are Scheduled,” *Journal of the Royal Statistical Society Series B*, 22, 108–113.

——— (1973), “Queues with Scheduled Arrivals: A Correction, Simplification and Extension,” *Journal of the Royal Statistical Society Series B*, 35, 1, 104–116.

MEZA, J. P. (1998), “Patient Waiting Times in a Physicians’ Office,” *The American Journal of Managed Care*, 4, 5, 703–712.

NUFFIELD PROVINCIAL HOSPITALS TRUST (1965), *Waiting in Outpatient Departments: A Survey of Outpatient Appointment Systems*, Oxford University Press, London.

O’KEEFE, R. (1985), “Investigating Outpatient Departments: Implementable Policies and Qualitative Approaches,” *Journal of the Operational Research Society*, 36, 8, 705–712.

OLDER, A. E. (1966), “British Hospital Journal and Social Service,” *British Hospital Journal and Social Service Review*, (March 11), 448–452.

PAUL, R. J. AND J. KULJIS (1995), “A Generic Simulation Package for Organizing Outpatient Clinics” in *Proceedings of the 1995 Winter Simulation Conference*, C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman (eds.), Association for Computing Machinery, Baltimore, 1043–1047.

PEGDEN, C. D. AND M. ROSENSHINE (1990), “Scheduling Arrivals to Queues,” *Computers & Operations Research*, 17, 4, 343–348.

PIERSKALLA, W. P. AND D. J. BRAILER (1994), “Applications of Operations Research in Health Care Delivery,” Ch. 13 in *Operations Research in the Public Sector*, S. M. Pollock, M. H. Rothkopf, and A. Barnett (eds.); Vol. 6 of *Handbooks in Operations Research and Management Science*, North-Holland, New York.

PRZASNYSKI, Z. H. (1986), “Operating Room Scheduling: A Literature Review,” *AORN Journal*, 44, 1, 67–79.

RISING, E., R. BARON, AND B. AVERILL (1973), “A System Analysis of a University Health Service Outpatient Clinic,” *Operations Research*, 21, 5, 1030–1047.

ROBINSON, L. W. AND R. R. CHEN (2001), “Scheduling Doctors’ Appointments: Optimal and Empirically-Based Heuristic Policies,” Unpublished working paper. Johnson Graduate School of Management, Cornell University, Ithaca, New York.

ROCKART, J. F. AND P. B. HOFMANN (1969), “Physician and Patient Behavior Under Different Scheduling Systems in a Hospital Outpatient Department,” *Medical Care*, 7, 6, 463–470.

ROHLEDER, T. R. AND K. J. KLASSEN (2000), “Using Client-Variance Information to Improve Dynamic Appointment Scheduling Performance,” *Omega*, 28, 3, 293–302.

SCHAFER, W. B. (1986), “Keep Patients Waiting? Not in My Office,” *Medical Economics*, 63, 10 (May 12), 137–141.

SHONICK, W. AND B. W. KLEIN (1977), “An Approach to Reducing the Adverse Effects of Broken Appointments in Primary Care Systems: Development of a Decision Rule Based on Estimated Conditional Probabilities,” *Medical Care*, 15, 5, 419–429.

SORIANO, A. (1966), “Comparison of Two Scheduling Systems,” *Operations Research*, 14, 388–397.

STAFFORD, E. F. AND S. C. AGGARWAL (1979), “Managerial Analysis and Decision-Making in Outpatient Health Clinics,” *Journal of Operational Research Society*, 30, 10, 905–915.

SWARTZMAN, G. (1970), “The Patient Arrival Process in Hospitals: Statistical Analysis,” *Health Services Research*, 5, 4, 320–329.

SWISHER, J. R., S. H. JACOBSON, J. B. JUN, AND O. BALCI (2001), “Modeling and Analyzing a Physician Clinic Environment Using Discrete-Event (Visual) Simulation,” *Computers & Operations Research*, 28, 2, 105–125.

Taylor, B. (1984), "Patient Use of Mixed Appointment Systems in an Urban Practice," *British Medical Journal*, 289 (November 10), 1277–1278.

Taylor III, B. W. and A. J. Keown (1980), "A Network Analysis of an Inpatient Outpatient Department," *Journal of Operational Research Society*, 31, 169–179.

Van Ackere, A. (1990), "Conflicting Interests in the Timing of Jobs," *Management Science*, 36, 8, 970–984.

Vanden Bosch, P. M., C. D. Dietz, and J. R. Simeoni (1999), "Scheduling Customer Arrivals to a Stochastic Service System," *Naval Research Logistics*, 46, 549–559.

——— and ——— (2000), "Minimizing Expected Waiting in a Medical Appointment System," *IIE Transactions*, 32, 9, 841–848.

Villegas, E. L. (1967), "Outpatient Appointment System Saves Time for Patients and Doctors," *Hospitals, J. A. H. A.*, 41, 52–57.

Virji, A. (1990), "A Study of Patients Attending Without Appointments in an Urban General Practice," *British Medical Journal*, 301 (July 7), 22–26.

Vissers, J. (1979), "Selecting a Suitable Appointment System in an Outpatient Setting," *Medical Care*, 17, 12, 1207–1220.

——— and J. Wijngaard (1979), "The Outpatient Appointment System: Design of a Simulation Study," *European Journal of Operational Research*, 3, 6, 459–463.

Walter, S. D. (1973), "A Comparison of Appointment Schedules in a Hospital Radiology Department," *British Journal of Preventive and Social Medicine*, 27, 160–167.

Wang, P. P. (1993), "Static and Dynamic Scheduling of Customer Arrivals to a Single-Server System," *Naval Research Logistics*, 40, 345–360.

——— (1997), "Optimally Scheduling *N* Customer Arrival Times for a Single-Server System," *Computers & Operations Research*, 24, 8, 703–716.

Weiss, E. N. (1990), "Models for Determining Estimated Start Times and Case Orderings in Hospital Operating Rooms," *IIE Transactions*, 22, 2, 143–150.

Welch, J. D. (1964), "Appointment Systems in Hospital Outpatient Departments," *Operational Research Quarterly*, 15, 224–232.

——— and N. Bailey (1952), "Appointment Systems in Hospital Outpatient Departments," *The Lancet*, 1105–1108.

Westman, G., S. Andersson, and P. Fredriksson (1987), "Waiting Room Time in the Assessment of an Appointment System in Primary Care," *Scandinavian Journal of Primary Health Care*, 5, 35, 35–40.

Williams, W. J., R. P. Covert, and J. D. Steele (1967), "Simulation Modeling of a Teaching Hospital Outpatient Clinic," *Hospitals, J. A. H. A.*, 41 (November 1), 71–75.

Yang, K. K., M. L. Lu, and S. A. Quek (1998), "A New Appointment Rule for a Single-Server, Multiple-Customer Service System," *Naval Research Logistics*, 45, 313–326.

Emre Veral is associate professor of operations management at Baruch College, Zicklin School of Business. He obtained his Ph.D. in industrial management at Clemson University. His research interests include health care operations, scheduling systems, and international business codes and compliance. He serves as Senior Research Fellow at the International Center for Corporate Accountability, housed at Baruch College. He is a member of Decision Sciences Institute, and the Production and Operations Management Society. His publications have previously appeared in *Decision Sciences, Journal of Operations Management, Production and Inventory Management Journal, Computers and Industrial Engineering*, and various trade journals.

Tugba Cayirli is a Special Assistant Professor of Operations Management at Hofstra University. She is currently a Ph.D. candidate in Management Planning and Information Systems at the Graduate School and University Center of the City University of New York. Her research interests include health care applications of management science, scheduling, queuing systems, and quality management.