Minimizing Total Cost in Scheduling Outpatient Appointments

Article *in* Management Science · December 1992

DOI: 10.1287/mnsc.38.12.1750 · Source: RePEc

2 authors, including:

Hon-Shiang Lau
000
138 PUBLICATIONS **4,676** CITATIONS

SEE PROFILE

MANAGEMENT SCIENCE
Vol. 38, No. 12, December 1992
Printed in U.S.A.

MINIMIZING TOTAL COST IN SCHEDULING OUTPATIENT APPOINTMENTS

CHRWAN-JYH HO AND HON-SHIANG LAU
Department of Management, Oklahoma State University, Stillwater, Oklahoma 74078

This paper considers various rules for scheduling appointments for medical clinic outpatients and investigates their ability to minimize a weighted sum of medical personnel's and patients' idle-time costs. It is shown that the idle times incurred by any given rule are affected by the following three "environmental factors" (in decreasing order of importance): the probability of no-show, the coefficient of variation of service times, and the number of patients per clinical session. Theoretically, an appropriate scheduling rule can be identified only if one knows the values of these parameters and the ratio between the medical personnel's and patients' idle-time costs.

Under environments characterized by 27 different combinations of the three environmental factors, the performance of nine scheduling rules are evaluated using simulation. Some of the rules evaluated are original to this study. The results are presented in the form of "efficient frontiers," together with a simple procedure for identifying the best scheduling rule for given environmental-parameter values. This rule-identification procedure is shown to be easily adaptable for circumstances with limited knowledge about the environmental factors; it also reveals that the simple Bailey-Welch individual-appointment rules are surprisingly robust.
(SCHEDULING; HEALTH CARE; SIMULATION)

1. Introduction

Under currently soaring health care costs and rising public discontent with health care quality, it becomes increasingly important to consider productivity improvements in the health care industry. This paper considers one aspect of reducing cost and improving service in outpatient clinics.

In scheduling medical appointments for outpatients, it is now well recognized that the idle times of patients should also be considered (see, e.g., Bailey 1952, Fries and Marathe 1981). Therefore, patients' appointments should be scheduled with the objective of minimizing some function of the idle costs of both the patients and the medical personnel (hereafter "m.p."). Our purpose here is to (i) compare the cost performance of various simple appointment-scheduling rules under a variety of operating environments, (ii) identify the major factors determining the cost performance of scheduling rules, and (iii) develop a simple procedure for prescribing a rule for a given environment. Some of the kinds of rules we consider are original to this paper.

Problem Definition

Let π and ω denote the unit idle cost (\$/unit time) of the patients and m.p., respectively (i.e., time-cost relationships are linear). Assume that each patient i arrives punctually at an appointed time A_i. Let t_i, b_i and e_i be, respectively, the length of service time, the time at which service begins, and the time at which service ends for patient i. Since t_i is stochastic, so are b_i and e_i. Without loss of generality, let $A_1 = b_1 = 0$. For $i > 1$, we have

$$b_i = \max(A_i, e_{i-1}), \qquad e_i = b_i + t_i. \tag{1}$$

Patient i's idle time is then

$$P_i = \max(0, b_i - A_i). \tag{2}$$

Following earlier works, the m.p.'s waiting time incurred just before patient i's arrival, i.e.,

0025-1909/92/3812/1750\$01.25

$$M_i = \max(0, A_i - e_{i-1}), \tag{3}$$

is considered the m.p.'s "idle time." Environments where M_i is not entirely unproductive can be handled by adjusting the idle-cost estimate π accordingly. For a clinical session with N patients scheduled, the stochastic total idle time of the patients (P) and m.p. (M) are, respectively:

$$P = \sum_{i=1}^{N} P_i \qquad \text{and} \qquad M = \sum_{i=1}^{N} M_i. \tag{4}$$

The general problem is to determine the values of the A_i's to optimize some objective function $f(P_i, M_i; i = 1 \text{ to } N)$. This study adopts the same objective used in all earlier works; i.e., to minimize $E(C)$, the expected cost of total patients' and m.p.'s idle time per clinical session:

$$E(C) = \pi \cdot E(P) + \omega \cdot E(M). \tag{5}$$

After the literature review in §2, several factors that help us better understand the problem are discussed in §3 and 4. The simulations are described in §5 and 6, and results are presented in §7 and 8. Managerial implementations of our results are discussed in §9, and the concluding §10 gives a summary and suggestions for further investigations.

2. Brief Literature Review

Let μ_t and σ_t denote t_i's mean and standard deviation, respectively. Also, $cv(t) = \sigma_t/\mu_t$. Without loss of generality, $\mu_t = 1$ is assumed throughout.

Bailey (1952, 1954), Welch (1964) and Welch and Bailey (1952) proposed the following "individual appointment scheduling rules": scheduling k patients to arrive at the start of a session, then schedule patients thereafter at intervals equal to the average service time; i.e.,

$$\text{set} \quad A_1 = A_2 = \cdots A_k = 0; \quad \text{then for } i > k, \text{ set} \quad A_i = A_{i-1} + \mu_t. \tag{6}$$

They used manual simulation to study the performance of these rules under different environments. These environments are characterized by different combinations of the following "environmental factors": (i) $cv(t)$ is one of four possible values (ranging from 0.53 to 0.71); and (ii) k in equation (6) is either 1, 2, 3, 4, 5 or 6. The following environmental factors are fixed: (i) N (no. of patients per session) = 25, and (ii) the t_i's are gamma distributed. They concluded that using $k = 2$ best compromises the conflicting needs of minimizing patients' and m.p.'s idle time.

Jansson (1966) considered a scheduling rule similar to Bailey-Welch's:

$$\text{set } A_1 \text{ to} \quad A_k = 0; \quad \text{then for } i > k, \text{ set} \quad A_i = A_{i-1} + I \tag{7}$$

where I (set at values less than μ_t) is a constant interval between scheduled appointments. For exponential t_i's, he derived (i) expressions for computing P_i and M_i in (2) and (3); and (ii) the limiting expressions for P_i and M_i for $i \to \infty$ when $\mu_t < I$.

In contrast to individual appointment rules, White and Pike (1964) and Soriano (1966), among others, have studied "block appointment rules" that can be characterized as:

Divide the session into k "blocks," and schedule $n = N/k$ patients to arrive at the beginning of each block; i.e.,

$$\text{set } A_{in+1} \text{ to} \quad A_{(i+1)n} = iT/k \quad \text{for} \quad i = 0 \text{ to } (k-1) \tag{8}$$

where T is the session length (in time units), e.g., assume $N = 30$, $T = 36$ and $k = 3$, then A_1 to $A_{10} = 0$; A_{11} to $A_{20} = 12$; A_{21} to $A_{30} = 24$.

Assuming exponentially distributed t_i's and load factor < 1 (i.e., $T > N\mu_t$, T defined in equation 8), Soriano (1966) derived and compared the limiting (i.e., $i \to \infty$) cost functions for the cases of $n = 1$ and 2, and recommended that patients be scheduled in blocks of 2. White and Pike (1964) arrived at a similar conclusion using simulation, but they also considered the effect of patients' and m.p.'s punctuality. Fries and Marathe (1981) considered a more sophisticated block appointment rule in which the number of patients n_j scheduled to arrive at the beginning of each block j need not be a constant for all j. For a given number of blocks (k), they developed a dynamic programming procedure to find the value of $[n_1, n_2, \ldots, n_k]$ that minimizes the expected total costs.

All the above studies assume homogeneous patients; i.e., the t_i's follow a single statistical distribution. In the slightly different context of scheduling medical operations (instead of patients), Charnetski (1984) recognizes that different types of operations have different time distributions, and he studied the scheduling rule:

$$\text{set} \quad A_1 = 0; \quad \text{then for } i > 0, \text{ set} \quad A_{i+1} = A_i + m_i + hs_i \tag{9}$$

where m_i and s_i are, respectively, the mean and standard deviation of the durations of type-i operations and h is a parameter to be fixed by (say) the systems analyst. Charnetski (1984) showed how: (i) h affects the facility's and users' expected idle costs; and (ii) the optimal h-value can be determined for different objectives.

Our study contributes to the above literature by: (i) considering a much larger set of scheduling rules; (ii) comparing their performance under different operating environments; and (iii) identifying the relevant environmental factors and summarizing how these factors affect the comparative performance of the various scheduling rules.

3. Characterizing Factors of a Clinical Session

Factors that may affect the performance of an appointment scheduling rule are considered below.

Probability Distribution of Service Times t_i's

Bailey's (1952, 1954) and Welch's (1964) simulations used gamma distributed t_i's with $cv(t)$ from 0.51 to 0.62. O'Keefe (1985) found empirically that t_i's have cv from 0.58 to 0.7, but their distributions have a variety of shapes (skewness and kurtosis) and do not follow simple two-parameter density functions such as gamma or normal (let alone exponential). In contrast, Jansson (1966), Soriano (1966) and Fries and Marathe (1981) assumed t_i to be exponential, obviously because their analytical approach will be intractable otherwise. The one-parameter exponential distribution requires $cv = 1$ exactly, and earlier empirical data indicates that this requirement is too high and restrictive.

Besides having different levels of cv, the t_i-distributions can also have various levels of skewness and kurtosis. In a more comprehensive simulation study by Ho and Lau (1987), t_i is modeled by a four-parameter family of distributions, which enables t_i's cv, skewness and kurtosis to be varied independently and systematically. Their study shows that the relative performance of appointment scheduling rules is affected only by t_i's cv, but not by t_i's skewness and kurtosis. Therefore, uniform and exponential distributions will be used to model t_i's of different cv's in this study.

Punctuality and No-Show

Punctuality of patients and m.p. was considered by, among others, Bailey (1952), Welch and Bailey (1952), White and Pike (1964), Fetter and Thompson (1966), Katz (1969) and O'Keefe (1985), but they reported quite different patterns of behavior. One expects punctuality behavior to depend very much on norms and culture, which vary considerably with time and locality. In other words, it will be difficult to model punctuality

realistically. Also, as Bailey (1952) and Welch and Bailey (1952) pointed out, patients may arrive late because they know they will have a long waiting time anyway, but punctuality can be expected under a credible appointment system. Therefore, patients and m.p. will be assumed punctual in this study.

No-show, on the other hand, is often due to unexpected developments unrelated to the appointment schedule. Earlier studies reported no-show probabilities as high as 20%. No-show cannot be handled in the analytical approaches of Jansson (1966) and Soriano (1966), but it can easily and will be incorporated in our simulations.

The Effect of the Number of Patients per Session, N

This subsection will demonstrate the effect of the environmental factor N on the performance of appointment-scheduling rules. This effect was largely overlooked in earlier studies; e.g., Bailey (1952, 1954) fixed $N = 25$.

Consider the scheduling rule called the "benchmark" rule hereafter:

$$\text{set} \quad A_1 = 0; \quad \text{then for } i = 2 \text{ to } N, \text{ set} \quad A_i = A_{i-1} + \mu_t \tag{10}$$

(with this schedule, both P and M defined in equation 5 will be zero if all the t_i's assume the deterministic value of μ_t). Assume that the t_i's are uniformly distributed with $\mu_t = 1$ and $cv(t) = 0.5$ (i.e., t_i's range is 0.134 to 1.866). The simulated means and standard deviations of P_i, M_i and e_i for using rule (10) are given in Table 1 for selected values of i from 1 to 30. From the complete set of values one obtains the following:

$$\begin{aligned} &N = 10, \quad E(P) = 5.75, \quad E(M) = 0.952, \quad E(P)/E(M) = 6.04 \\ &N = 20, \quad E(P) = 18.48, \quad E(M) = 1.488, \quad E(P)/E(M) = 12.42 \\ &N = 30, \quad E(P) = 35.71, \quad E(M) = 1.893, \quad E(P)/E(M) = 18.86. \end{aligned} \tag{11}$$

The ratios $E(P)/E(M)$ shown in (11) indicate that an appointment schedule deemed suitable for a certain N-value may not be suitable for other N-values. For example, if the clinical/societal cost structures are such that $\omega/\pi = 12$, the scheduling rule defined in (9) is adequate for this ω/π-ratio only if the sessions are run with $N = 20$ patients each. However, the rules causes too much m.p.'s idle time if $N = 10$, and too much patient's idle time if $N = 30$.

4. Use of the "Steady State" and Constant *I*

Table 1 brings out two other important points, whose realization leads to the proposal of potentially good scheduling rules.

Relevance of the "Steady State"

Table 1 shows that the distributions of P_i and M_i change with i. Particularly, $E(P_i)$ increases but $E(M_i)$ decreases with i. This is expected from basic queuing theory: with the benchmark rule defined in (10), the interarrival and service times are equal, i.e., the queue's load factor is 1. Therefore, the expected customer's waiting time increases with time without bound, while the expected facility idle time decreases. Noting that: (i) many "good" rules (see §6) set interarrival time to equal service time; and (ii) there are seldom more than 20 or 30 patients per session; there is little point talking about the "steady-state distribution of P_i's or M_i's when $i \to \infty$." However, several earlier studies (e.g., Jansson 1966 and Soriano 1966) concentrated on deriving such steady-state properties.

Necessity of Restrictions on the Scheduled Interarrival Interval I

In order for a steady-state to exist, Jansson (1966) and Soriano (1966) considered only environments where the load factor < 1 (i.e., as stated in §2, $\mu_t < I$ for individual

TABLE 1
Mean and Standard Deviation of P_i and M_i

Patient No.	$E(P_i)$	$\alpha(P_i)$	$E(M_i)$	$\sigma(M_i)$
1	0.000	0.000	0.000	0.000
2	0.222	0.281	0.215	0.282
3	0.362	0.411	0.144	0.239
5	0.576	0.583	0.098	0.203
7	0.749	0.720	0.081	0.189
9	0.905	0.843	0.067	0.181
11	1.032	0.943	0.064	0.171
15	1.255	1.124	0.056	0.162
20	1.488	1.301	0.046	0.147
25	1.707	1.465	0.040	0.138
30	1.903	1.619	0.035	0.128

appointment systems and $T > N\mu_t$ for block appointment systems). Perhaps for similar reasons, Bailey (1952, 1954) and Welch (1964) considered only schedules with $I = \mu_t$. Furthermore, all earlier studies considered only scheduling rules with constant I. However, Table 1 illustrates clearly that we are unavoidably dealing with the transient state, since different expected idle times are associated with patients arriving at different parts of the session. There is therefore no reason why we should not consider schedules with $I < \mu_t$ or with I_i's that vary with i; as explained below, such schedules may be beneficial.

The "Variable-I" Concept

In Table 1, note that $E(P_i)$ increases but $E(M_i)$ decreases with i. That is, patients with earlier appointments tend to have shorter waiting times—a widely-held belief among public-clinic patients. This "unfair" phenomenon can be corrected by using scheduling rules with variable scheduled interarrival intervals I_i's, with perhaps $I_i < \mu_t$ for the patients scheduled for a session's earlier part, and $I_i > \mu_t$ for the patients in a session's latter part. Compared to the benchmark schedule in Table 1, the proposed variable-I schedule requires patients in a session's earlier part to arrive earlier, and patients in a session's latter part to arrive later; this will reduce the variation of the expected idle times with arrival sequence i. Such variable-I rules *may* also reduce the total idle-time cost over an entire session. As far as we can ascertain, such variable-I rules have not been considered in the management science and health services literatures.

5. The Simulations

The preceding section shows that in studying clinical appointment scheduling, one is primarily interested in the transient-state behavior of a queue with non-Erlangian interarrival/service times. While the problem can be easily studied using simulation, analytical methods will be intractable.

Keeping in mind the observations in §3 and 4, sessions ("environments") with the following characteristics are considered in our simulations:

(i) Service-time Distribution: $\mu_t = 1$ time unit, $cv(t) = 0.2$ & 0.5, uniformly distributed, $cv(t) = 1.0$, exponentially distributed.
(ii) No. of Patients per Session: $N = 10$, 20 or 30.
(iii) No-show probability: $\rho = 0.0$, 0.1 or 0.2.
(iv) Punctuality: patients and m.p. are all punctual.

For each of the 27 different environments (3 cv-values $\times$ 3 N-values $\times$ 3 ρ-values), the cost performance of 50 different appointment scheduling rules were simulated and reported in Ho and Lau (1987). Each of the 50 rules is selected for its ease of imple-

mentation, similarity to the rules considered in the past, and/or its "potential" based on the observations made in the preceding sections. In this paper we concentrate on examining the results of the eight rules that performed "best" (the "efficient frontier" performance criterion explained in §7 is used in this and all following statements involving comparison of rules' performance). However, some of the remaining 42 "variations" are also mentioned briefly.

All simulations were made for 10,000 sessions. Pilot simulations indicate that at this sample size, the simulated values of $E(P)$ and $E(M)$ are accurate to within $\pm 1\%$ at the 95% confidence level, and this accuracy level is adequate for our purpose of comparing the performance of various scheduling rules under various environments.

6. The Appointment Scheduling Rules

We consider here nine rules that reflect the various strategies discussed in §2 to 4.

Rule 1. Set $A_1 = A_2 = 0$; then for $i > 2$, set $A_i = A_{i-1} + \mu_t$. This is the original Bailey-Welch rule with two patients at the start of a session.

One seemingly promising group of variations are the rules

$$A_1 = A_2 = 0; \qquad \text{then for } i > 2, \text{ set} \qquad A_i = A_{i-1} + \mu_t + k\sigma_t \qquad (k > 0). \quad (12)$$

These variations delay the scheduled arrivals of the third and later patients; the motivation for doing this is to reduce the "unfair" additional expected waiting time of the patients scheduled to arrive in a session's latter part, as depicted in Table 1. Another seemingly promising group if variations are:

$$A_1 = A_2 = A_3 = 0; \qquad \text{then for } i > 3, \text{ set} \qquad A_i = A_{i-1} + \mu_t + k\sigma_t \qquad (k \geq 0) \quad (13)$$

which are counterparts to Rule 1 and its variation defined in (12), but with 3 patients at the start of each session. However, both groups of variations were shown in Ho and Lau (1987) to be inferior to Rule 1.

Rule 2. Set $A_1 = 0$, $A_2 = 0.2$, $A_3 = 0.6$; then for $i > 3$, set $A_i = A_{i-1} + \mu_t$.

Rule 3. Set $A_1 = 0$, $A_2 = 0.3$, $A_3 = 0.6$, $A_4 = 0.9$; then for $i > 4$, set $A_i = A_{i-1} + \mu_t$.

Rule 4. Set $A_1 = 0$, $A_2 = 0.5$, $A_3 = 1.0$, $A_4 = 1.5$; then for $i > 4$, set $A_i = A_{i-1} + \mu_t$.

Rule 2 is a special case of a group of rules defined by

$$A_1 = 0, \qquad A_2 = k_2, \qquad A_3 = k_3;$$
$$\text{then for } i > 3, \qquad A_i = A_{i-1} + \mu_t + k\sigma_t \qquad (k \geq 0). \quad (14)$$

Similarly, Rules 3 and 4 are special cases of a similar group:

$$A_1 = 0, \qquad A_2 = k_2, \qquad A_3 = k_3, \qquad A_4 = k_4;$$
$$\text{then for } i > 4, \qquad A_i = A_{i-1} + \mu_t + k\sigma_t \qquad (k \geq 0). \quad (15)$$

These groups represent further modifications of Rule 1, (12) and (13) by scheduling the 2nd, 3rd (and 4th) patient to arrive earlier than what the benchmark schedule (10) calls for, but not as early as time 0 as required by the original Bailey-Welch rules. Various values of k_2, k_3, k_4 and k were tested in Ho and Lau (1987), and the specific values of k_i's and k shown in Rules 2 to 4 performed best.

Rule 5. Set $A_1 = A_2 = A_3 = A_4 = 0$; then for $i > 4$, set $A_i = A_{i-1} + \mu_t$. This is simply Rule 1 but with four patients at the start of a session.

Ho and Lau (1987) were unable to obtain better rules from the following generalization of Rule 5:

$$\text{Set} \qquad A_1 = A_2 = A_3 = A_4 = 0;$$
$$\text{then for } i > 4, \text{ set} \qquad A_i = A_{i-1} + \mu_t + k\sigma_t \qquad (k \geq 0). \quad (16)$$

Rule 6. Set $A_i = (i-1)\mu_t - k\sigma_t$, $(k = 0.1)$. If $k = 0$, this reduces to the benchmark schedule (10). Using $k > 0$ causes the patients to arrive earlier (compared to the benchmark) by $k\sigma_t$ time units. Different values of k in the modification step were tried, and $k = 0.1$ was found to perform best.

Rules 7 *and* 8. They are special cases of the general form:

$$\text{First set } A_i = (i-1)\mu_t; \quad \text{then for } i \le K, \text{ modify as } A_i = A_i - k_1(K-i)\sigma_t;$$
$$\text{and for } i > K, \text{ modify as } A_i = A_i - k_2(K-i)\sigma_t \qquad (17)$$

which is one form of implementing the "variable-I" concept stated in §4. The first step in (17) gives the benchmark schedule, the modification causes patients #2 to #$(K-1)$ to arrive earlier by $k_1(K-i)\sigma_t$ time units, but causes patients #$(K+1)$ to #N to arrive later by $k_2(K-i)\sigma_t$ time units. Also, the magnitudes of modification are smaller for patients arriving towards the middle of a session. These modifications are achieved by the term $(K-i)$, whose positive value decreases as i increases as long as $i < K$, but it has a negative value whose magnitude increases with i when $i > K$. Also, (k_1/k_2) controls the ratio between the rate of earliness imposed on the first K patients and the rate of lateness imposed on the remaining patients.

Ho and Lau (1987) considered many different combinations of the parameters K, k_1 and k_2 in (17); the following combinations performed best:

Rule 7. $k_1 = 0.15$, $k_2 = 0.3$, $K = 5$ in (17);

Rule 8. $k_1 = 0.25$, $k_2 = 0.5$, $K = 5$ in (17).

Note that although Rules 6 to 8 (and also expressions 12 to 16) appear complicated, they are no more difficult to implement than the simpler rules such as Rule 1. To implement any rule, the A_i's (i.e., the schedule) need to be computed only once, the appointments clerk then simply reads off the printed schedule to the patients each day; the rule's mathematical definition is never used again.

Rule 9. Set $A_i = A_{i+1} = (i-1)\mu_t$, $i = 1, 3, 5, 7 \cdots$. This is the block appointment rule with two patients per block, as advocated in Soriano (1966) and White and Pike (1964). This is NOT one of the "good" rules identified in Ho and Lau (1987), but the reason for its inclusion will be clarified below.

7. The Efficient Frontier of the Rules' $E(P) - E(M)$ Points

Consider the case of $cv(t) = 0.5$, $N = 20$ and $\rho = 0.0$. The values of $E(P)$ and $E(M)$ for Rules 1 to 9 are given in Table 2. For Rule 1, its $\{E(P), E(M)\} = \{26.319, 0.758\}$ is plotted as point "1" in Figure 1. The $\{E(P), E(M)\}$-values of Rules 2 to 9 are similarly plotted in Figure 1. Points "1" to "8" (corresponding to Rules 1 to 8) form a piecewise-linear "efficient frontier" (borrowing the well-known term and concept from the theory of financial portfolio). Most of the other rules in Ho and Lau's (1987) set of 50 rules have $E(P) - E(M)$ points that lie on the upper right-hand side of this efficient frontier; they are, therefore, "inferior" rules dominated by the frontier rules. In this paper, only one such inferior rule (Rule 9) is included; Figures 1 to 3 illustrate Rule-9's off-frontier position.

The points "1" to "8" can also be thought of as the lower boundary of the "feasible region" in the graphical two-variable cost-minimizing linear programming (LP) procedure. Therefore, the procedure for identifying the best scheduling rule is the same as the LP graphical procedure for minimizing $(ax + by)$; i.e., one compares the slope (a/b) of the isocost lines (i.e., lines for $ax + by =$ any constant) with the slopes of the various segments of the feasible-region's lower boundary. Since the objective here is to minimize $\omega E(M) + \pi E(P)$ (see equation 5), the slope of the isocost lines is ω/π. The slope of the frontier segment (say) "5 − 3" in Figure 1 is, using the entries in Table 2,

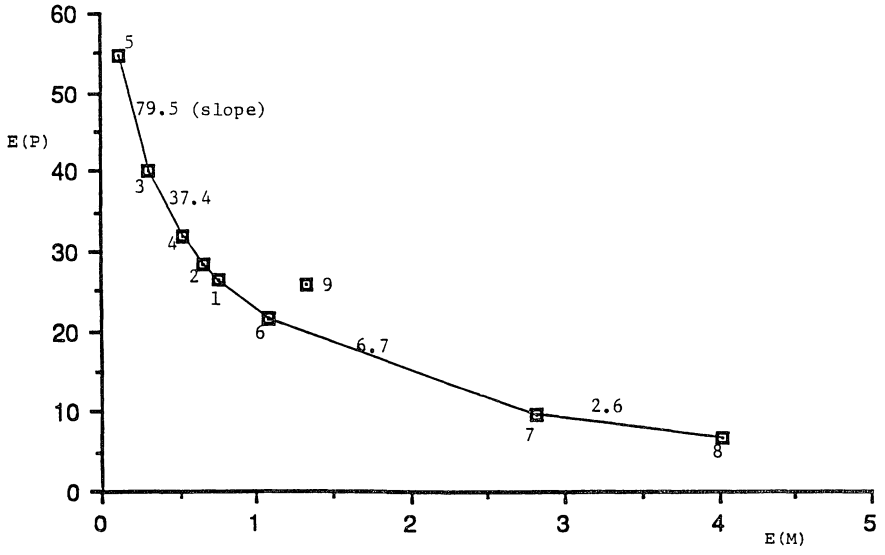

FIGURE 1. $E(P) - E(M)$ Points for Nine Rules
$cv = 0.5, \quad N = 20, \quad \rho = 0.0$

$$g(5, 3) = \frac{54.85 - 40.23}{0.307 - 0.123} = 79.5,$$

which is shown adjacent to the segment in Figure 1. (Since all slopes are actually negative, the negative sign is omitted.) The slopes of the other frontier segments (shown on the segments' left in Figure 1) are similarly computed. These slopes indicate that, if ω/π is between 79.5 and 37.4, Rule 3 should be used, but if ω/π is between 37.4 and 29.1, Rule 4 should be used. As in the LP graphical procedure, the cost minimizing rule will always

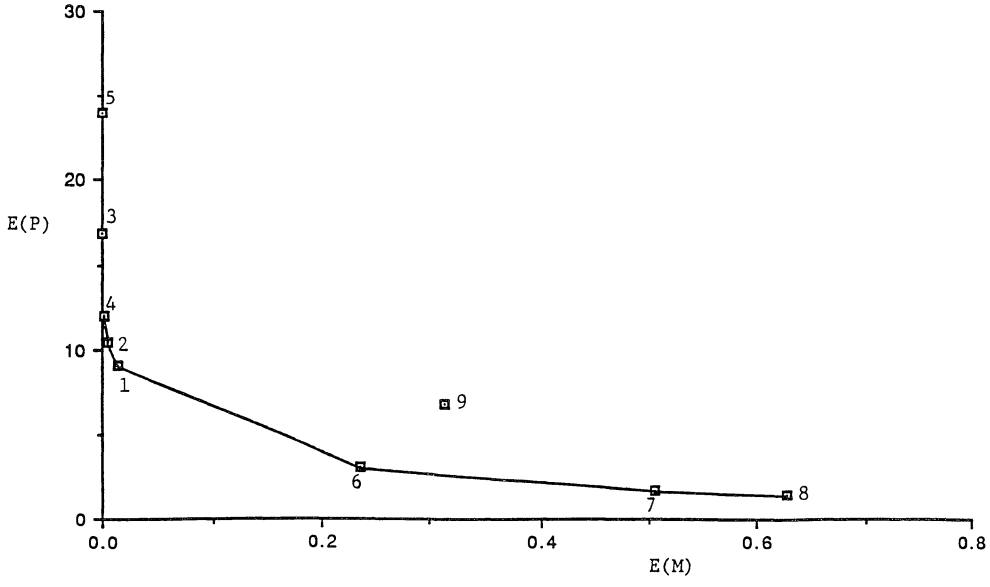

FIGURE 2. $E(P) - E(M)$ Points for Nine Rules
$cv = 0.2, \quad N = 10, \quad \rho = 0.0$

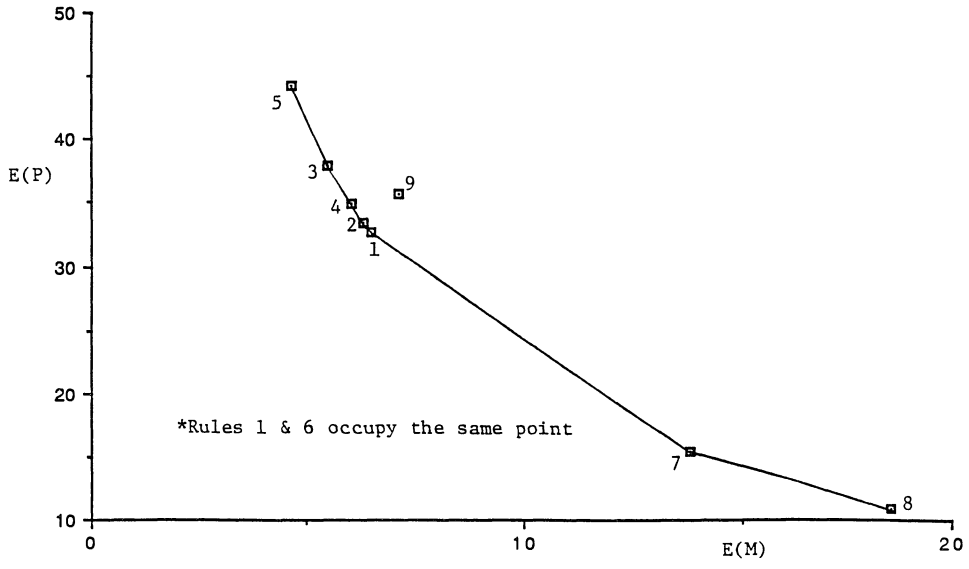

FIGURE 3. $E(P) - E(M)$ Points for Nine Rules
$cv = 1.0, \quad N = 30, \quad \rho = 0.0$

be one of the "corner rules" along the efficient frontier in Figure 1. Point "9" (for Rule 9), which is above the efficient frontier, is never attractive under any ω/π-ratio.

The efficient frontier concept reveals the flaw of an intuitive approach suggested in some earlier works. To illustrate, assume that the clinical environment is judged to have a ω/π-ratio of 43.7; noting that Rule-2's $E(P)/E(M) = 28.4/0.65$ (from Table 2) $= 43.7$, it might appear that Rule 2 is most appropriate for this environment. However, the preceding paragraph shows that Rule 3 is the appropriate one, even though $E(P)/E(M) = 129.7$ for Rule 3. To verify this, the objective-function value obtained by using Rule 3 is $40.2\pi + 0.31\omega$ (from Table 2) $= 40.2\pi + 0.31 \times (43.7\pi) = 53.75\pi$, which is less than the objective-function value of using Rule 2, i.e., $28.4\pi + 0.65\omega$ (from Table 2) $= 28.4\pi + 0.65 \times (43.7\pi) = 56.81\pi$.

The important point to note here is that, given an environment's ω/π-ratio, the suitability of a certain rule cannot be judged by its $E(P)/E(M)$-ratio alone, contrary to intuition and earlier suggestions. A rule's suitability must be determined by considering the performance of the available alternative rules on the efficient frontier. This conclusion parallels the case of minimizing a linear objective function $(ax + by)$ subject to a lower-bounded feasible region: the optimal corner point (x_p, y_p) is NOT the point such that $y_p/x_p = a/b$; instead, a/b should have a value between the slopes of the two adjacent feasible-region boundary segments joining at the point (x_p, y_p).

Figures 2 and 3 are counterparts to Figure 1, but for environments $\{cv = 0.2, N = 10, \rho = 0\}$ and $\{cv = 1, N = 30, \rho = 0.2\}$, respectively. Similar figures for the other 23 environments are available from the authors; they have the same patterns exhibited in Figures 1 to 3.

TABLE 2

Values of $E(P)$ and $E(M)$ for Rules 1 to 9 $cv(t) = 0.5, N = 20, \rho = 0.0$

Rule	1	2	3	4	5	6	7	8	9
$E(P)$	26.3	28.4	40.2	31.9	54.85	21.4	9.9	6.7	25.85
$E(M)$	0.76	0.65	0.31	0.53	0.12	1.08	2.81	4.03	1.33

Criterion for Selecting the Nine Rules under Consideration

Ho and Lau (1987) considered 50 scheduling rules and plotted graphs similar to Figures 1 to 3 for the 27 different (cv, N, ρ) combinations. From those 50 rules, Rules 1 to 8 are selected here because they are on the Ho and Lau (1987) efficient frontiers for many more environments than other rules; and even in environments where they are not on the frontier, they are always very close to it. Rules whose corresponding $E(P) - E(M)$ points often lie far above the frontier's upper left are "poor" rules for the purpose of minimizing $E(C)$ (see equation 5). Among those rules revealed as "poor" in Ho and Lau (1987), we chose the widely-used Rule 9 as an illustration here.

Note that Rule 1 is the simple Bailey-Welch's 2-initial-patient rule, and Rule 9 is the "blocks-of-2" rule studied by, e.g., Soriano (1966) and White and Pike (1964); they have the status of "standard rules" in the appointment-scheduling literature, but their original advocates considered their performance only in specific environments (i.e., specific values of cv, N and/or ρ). It is interesting that under diversified environments, Rule 1 turns out to be very robust (always on or close to the frontier), whereas Rule 9 is practically always "poor;" i.e., the $E(P) - E(M)$ point for Rule 9 consistently lies significantly above the frontier in all 27 $\{cv, N, \rho\}$ environments. Our simulation results (omitted here) show that other block appointment rules with more than two patients per block are even poorer.

Use of Table 3: *Corner Rules and Segment Slopes for* 27 *Environments*

As a space-saving alternative to Figures 1 to 3, the efficient frontiers of all 27 environments are summarized in Table 3 by each segment's corner rules and slope (in decreasing order).

To illustrate, consider the environment $\{cv = 0.5, N = 20, \rho = 0.0\}$. If Figure 1 is not available, and (say) $\omega/\pi = 6$, one can simply scan the corresponding cell in Table 3 and note the entries: $\{(7, 6)\ 6.7\}$ and $\{(8, 7)\ 2.6\}$; i.e., segments (7, 6) and (8, 7) have slopes of 6.7 and 2.6, respectively. Hence the cost minimizing rule is Rule 7, which can be easily verified with Figure 1.

Similarly, for the environment $\{cv = 0.2, N = 10, \rho = 0.0\}$, with $\omega/\pi = 6$ unchanged, the appropriate rule is Rule 6 according to both Table 3 and Figure 2. For the environment $\{cv = 1, N = 30, \rho = 0.2\}$ and $\omega/\pi = 6$, the appropriate rule is Rule 3.

8. Performance of Different Rules in Different Environments

A glance at Table 3 reveals that the order of appearance of Rules 1 to 8 from left to right along the efficient frontier for different environments remain essentially the same as follows:

$$5, 3, 4, 2, 1, 6, 7, 8. \tag{18}$$

(The only minor exception is that, at $cv = 0.1$, Rule 1 is not on any frontier, and Rule 4 is not on some frontiers.) This means that for decreasing ω/π ratios, one selects the appropriate rule by moving from left to right on the list given in (18). This is because rules towards the right on (18) have lower $E(P)/E(M)$-ratios. However, it is important to note from Table 3 that the actual matching ω/π-ratios for a given rule vary widely with an environment's characteristics. As an extreme example, Table 3 shows that under the environment $\{cv = 0.2, N = 10, \rho = 0\}$, Rule 3 should only be used if $\omega/\pi > 4799$, but under the environment $\{cv = 1, N = 10,$ and $\rho = 0.2\}$, Rule 3 becomes appropriate when ω/π is as small as 4.0. In fact, for this latter environment, if ω/π exceeds 6.6, one should develop rules with higher $E(P)/E(M)$-ratios than any of the eight "good" rules considered here (an issue reconsidered in §10).

TABLE 3

Corners and Slopes of Frontiers

(*The two rule numbers of a segment are given in parentheses, followed by the segment's slope*)

cv = 0.2												
	$N = 10$			$N = 20$			$N = 30$		
$p = 0.0$		$p = 0.1$		$p = 0.2$	$p = 0.0$		$p = 0.1$		$p = 0.2$	$p = 0.0$		$p = 0.1$		$p = 0.2$
(3, 5)	9999.9	(3, 5)	61.8	(3, 5)	16.1	(3, 5)	8096.5	(3, 5)	35.2	(3, 5)	10.7	(3, 5)	1931.9	(3, 5)	28.1	(3, 5)	9.8
(4, 3)	4799.0	(4, 3)	26.7	(4, 3)	8.6	(4, 3)	563.8	(4, 3)	19.7	(4, 3)	7.1	(4, 3)	348.4	(4, 3)	18.0	(4, 3)	6.9
(2, 4)	397.8	(2, 4)	14.7	(2, 4)	5.1	(2, 4)	157.9	(2, 4)	12.5	(2, 4)	4.7	(2, 4)	132.0	(2, 4)	12.3	(2, 4)	4.7
(1, 2)	152.9	(1, 2)	10.8	(1, 2)	4.2	(1, 2)	92.6	(1, 2)	10.3	(1, 2)	4.1	(1, 2)	90.5	(1, 2)	10.4	(1, 2)	4.1
(6, 1)	27.1	(6, 1)	6.7	(6, 1)	3.1	(6, 1)	31.4	(6, 1)	7.1	(6, 1)	3.1	(6, 1)	35.7	(6, 1)	7.3	(6, 1)	3.1
(7, 6)	4.8	(7, 6)	2.2	(7, 6)	1.2	(7, 6)	6.7	(7, 6)	2.5	(7, 6)	1.3	(7, 6)	8.4	(7, 6)	2.7	(7, 6)	1.3
(8, 7)	2.1	(8, 7)	1.1	(8, 7)	0.7	(8, 7)	2.6	(8, 7)	1.3	(8, 7)	0.7	(8, 7)	2.8	(8, 7)	1.3	(8, 7)	0.7
cv = 0.5																	
	$N = 10$			$N = 20$			$N = 30$		
$p = 0.0$		$p = 0.1$		$p = 0.2$	$p = 0.0$		$p = 0.1$		$p = 0.2$	$p = 0.0$		$p = 0.1$		$p = 0.2$
(3, 5)	108.7	(3, 5)	26.6	(3, 5)	11.1	(3, 5)	79.5	(3, 5)	22.1	(3, 5)	9.6	(3, 5)	77.3	(3, 5)	21.4	(3, 5)	9.4
(4, 3)	32.8	(4, 3)	12.7	(4, 3)	6.4	(4, 3)	37.4	(4, 3)	13.4	(4, 3)	6.3	(4, 3)	41.4	(4, 3)	14.0	(4, 3)	6.4
(2, 4)	25.1	(2, 4)	9.7	(2, 4)	4.0	(2, 4)	29.1	(2, 4)	10.8	(2, 4)	4.9	(2, 4)	34.0	(2, 4)	11.7	(2, 4)	5.0
(1, 2)	14.0	(1, 2)	6.4	(1, 2)	3.3	(1, 2)	19.2	(1, 2)	7.8	(1, 2)	3.8	(1, 2)	23.5	(1, 2)	8.7	(1, 2)	4.0
(6, 1)	11.4	(6, 1)	5.5	(6, 1)	3.0	(6, 1)	15.1	(6, 1)	6.5	(6, 1)	3.3	(6, 1)	18.2	(6, 1)	7.1	(6, 1)	3.4
(7, 6)	4.7	(7, 6)	2.5	(7, 6)	1.5	(7, 6)	6.7	(7, 6)	3.1	(7, 6)	1.6	(7, 6)	8.4	(7, 6)	3.5	(7, 6)	1.7
(8, 7)	2.0	(8, 7)	1.2	(8, 7)	0.7	(8, 7)	2.6	(8, 7)	1.4	(8, 7)	0.8	(8, 7)	2.8	(8, 7)	1.4	(8, 7)	0.8
cv = 1.0																	
	$N = 10$			$N = 20$			$N = 30$		
$p = 0.0$		$p = 0.1$		$p = 0.2$	$p = 0.0$		$p = 0.1$		$p = 0.2$	$p = 0.0$		$p = 0.1$		$p = 0.2$
(3, 5)	20.5	(3, 5)	11.2	(3, 5)	6.6	(3, 5)	23.8	(3, 5)	12.9	(3, 5)	7.4	(3, 5)	27.7	(3, 5)	14.4	(3, 5)	7.9
(4, 3)	10.3	(4, 3)	6.4	(4, 3)	4.0	(2, 3)	14.3	(2, 3)	8.4	(2, 3)	4.9	(2, 3)	17.6	(2, 3)	9.7	(2, 3)	5.4
(2, 4)	9.7	(2, 4)	6.1	(2, 4)	3.8	(6, 2)	9.3	(6, 2)	5.5	(6, 2)	3.3	(6, 2)	11.9	(6, 2)	6.6	(6, 2)	3.8
(6, 2)	5.8	(6, 2)	3.8	(6, 2)	2.4	(7, 6)	5.4	(7, 6)	3.3	(7, 6)	2.0	(7, 6)	7.0	(7, 6)	4.0	(7, 6)	2.3
(7, 6)	3.6	(7, 6)	2.4	(7, 6)	1.6	(8, 7)	2.1	(8, 7)	1.4	(8, 7)	0.9	(8, 7)	2.4	(8, 7)	1.5	(8, 7)	1.0
(8, 7)	1.4	(8, 7)	1.0	(8, 7)	0.7												

Effect of Using an "Incorrect Rule" for a Given Environment

Assume $\omega/\pi = 6$, and without loss of generality, assign $\pi = 1$, hence $\omega = 6$. For the environment $\{cv = 0.5, N = 20, \rho = 0\}$, we showed earlier that the cost-minimizing rule is Rule 7. With this rule, the expected total cost of idleness (using equation 5 and the entries in Table 2) is

$$E(C|\text{Rule } 7) = 9.9 + 6(2.81) = 26.76.$$

Instead of following the preceding procedure that considers the values of cv, N, ρ and ω/π in choosing a rule, if one of the eight good rules is simply chosen arbitrarily, then the worst possibility is when Rule 5 is chosen, with

$$E(C|\text{Rule } 5) = 54.85 + 6(0.12) = 55.57.$$

The cost penalty for this arbitrary choice is $(55.57 - 26.76)/26.76$, or 108%.

Using a "Universal Rule" for All Environments

Among the eight good rules, the most easily implementable are Rules 1 and 5 (the Bailey-Welch rules with two and four initial patients, respectively). Also, list (18) shows that Rule 1 appears to be roughly the "central" rule along any frontier. Therefore, a natural question is: "what is the cost implication of applying Rule 1 universally, regardless of the environment?"

Assume $\omega/\pi = 10$. For the environment $\{cv = 0.5, N = 20, \rho = 0\}$, Table 3 shows that Rule 6 is best, and from Table 2, $E(C|\text{Rule } 6) = 21.4 + 10(1.08) = 32.3$. Using Rule 1, however, gives $E(C|\text{Rule } 1) = 26.3 + 10(0.76) = 33.9$ (from Table 2 entries). The cost penalty for using Rule 1 is $(33.9 - 32.3)/32.3 = 5.0\%$.

Repeating these calculations for the 26 other environments gives the results shown in Table 4. These results indicate that the cost penalty of using Rule 1 universally may be substantial at times; the worst-case penalty is 70.6%. Assuming a more extreme ω/π-

TABLE 4

Cost Penalty Percentage for Using Rule 1 Universally Instead of the Minimum-Cost Rule When $\omega/\pi = 10$

	$N = 10$			$N = 20$			$N = 30$		
$cv = 0.2$									
ρ	0.0	0.1	0.2	0.0	0.1	0.2	0.0	0.1	0.2
Best Rule	5	1	3	6	1	3	6	1	5
Penalty %	70.6	—	17.8	58.0	—	13.3	55.3	—	8.8
$cv = 0.5$									
ρ	0.0	0.1	0.2	0.0	0.1	0.2	0.0	0.1	0.2
Best Rule	6	4	3	6	2	5	6	2	5
Penalty %	2.6	2.3	22.7	5.0	1.0	12.1	9.3	0.4	8.1
$cv = 1.0$									
ρ	0.0	0.1	0.2	0.0	0.1	0.2	0.0	0.1	0.2
Best Rule	4	3	5*	2	3	5	6	3	5*
Penalty %	2.1	12.7	33.2	0.2	3.3	12.8	2.6	0.8	7.3

* At upper left end of frontier. Rules with higher $E(P)/E(M)$ ratio desired.

value and/or a universal rule further from the center in list (18) will lead to a similar conclusion, but with the worst-case penalty even larger.

Effects of ρ, $cv(t)$ and N on the Rules' Performance

Scanning the entries in Table 3 reveals the following:

(1) For ρ's effect, consider the three adjacent cells in Table 3 for $cv = 0.2$, $N = 10$ and $\rho = 0$, 0.1 and 0.2. The slopes of the line segment (8, 7) decrease from 2.1 to 0.7 as ρ increases from 0 to 0.2. It can be easily verified that the slopes of all frontier line segments decrease with increasing ρ. Therefore, a clinic with given values of N, cv and ω/π should adopt rules further towards the upper left-hand corner of the frontier (or the left of list 18) when ρ increases. This effect can be explained by noting that on moving left along (18), one encounters rules that incur more patients' idle time and less m.p.'s idle time; this counteracts the increase in m.p.'s idle time when ρ increases.

(2) By, say, comparing the three cells in the first column of Table 3 for $N = 10$, $\rho = 0$ and $cv = 0.2$, 0.5 and 1, one can similarly verify that the slopes of the frontier's segments usually decrease with increasing cv. Hence rules further towards the left of list 18 become more advantageous when cv increases.

(3) The effect of the environmental factor N depends on the portion of the frontier and the values of ρ and $cv(t)$, and cannot be simply generalized.

(4) The rules' performance is most affected by the factor ρ, then by cv, and least by the factor N. Note, however, that ρ was not considered in most earlier studies.

9. Managerial Implementations

The Ideal Procedure

Ideally, the systems analyst should emphasize to the clinic administration that there is no such thing as a "universally good" rule, and an appropriate rule can only be chosen after the clinic's values of $cv(t)$, N, ρ and ω/π are ascertained. The first three values are obtainable from simple objective statistical data. Specifically, $cv(t)$ is obtainable from the same data required to estimate μ_t (the most fundamental and indispensable parameter recognized in all earlier studies); while N should be readily available from standard clinic records. ρ may not be commonly available from standard records, but its estimation is also very simple; furthermore, after ρ is estimated, it is opportune to stress to the administration that reducing ρ is the most important factor for reducing total idle cost.

The determination of ω/π is more subjective and difficult. Note, however, that only the ratio ω/π is needed, but not the absolute values of ω and π. See Fries and Marathe (1981) for additional comments and suggestions on estimating ω/π.

After the values of $cv(t)$, N, ρ and ω/π are estimated, the simple procedure explained in §7 can be used to identify the cost-minimizing rule. For environments with values of $cv(t)$, N and ρ differing from the ones tabulated in Table 3, the effects of these parameters stated at the end of §8 can be used for interpolation.

A major purpose of this study is to suggest a framework for evaluating different scheduling rules in different operating environments. If a clinic wants to implement other scheduling rules not considered here, or if the environment's parameters (e.g., the cv, N and/or ρ) are very different from the ones considered here, the clinic can easily perform their own simulations and organize their results following the approach explained in this paper.

A "Rule-of Thumb" Approach

Assume that the clinic administration is unwilling to obtain quantitative estimates of $cv(t)$, N, ρ and ω/π, but a rule has to be recommended anyway. Given the studies of Bailey (1952) and O'Keefe (1985), one may then assume $cv(t) \simeq 0.5$. Also, without

TABLE 5
Rule-of-Thumb Recommendation with Qualitative Estimates of ω/π and ρ

Estimated ω/π	Low		Medium		High	
Estimated ρ	Low	High	Low	High	Low	High
Recommended Rule	8	1	1	1	5	5

accurate parameter estimates, there is no point in differentiating among rules that are relatively close to one another on the efficient frontiers. Therefore, Figures 1 to 3 suggest that the efficient frontier's range can be covered by three rules: Rules 5, 1 and 8. The administration should now estimate whether ω/π is "low," "medium" or "high;" and whether ρ is "low" or "high." Assuming that:

(i) for ω/π, "low" means $\{\omega/\pi < 4\}$, "medium" means $\{4 \leq \omega/\pi < 20\}$, and "high" means $\{\omega/\pi \geq 20\}$; and

(ii) for ρ, "low" means $\{\rho \simeq 0\}$ and "high" means $\{\rho \simeq 0.1\}$, then the entries for $cv = 0.5$ in Table 3 lead to the following rule-of-thumb recommendations:

Considered above are two extreme conditions of availability of quantitative data. For intermediate circumstances, one can derive similar adaptations of the results presented in §6 to 8.

10. Summary and Conclusion

This study shows that, for evaluating appointment scheduling rules, the major characterizing factors of a clinical session ("environment") are (in decreasing order of importance): (i) ρ, the probability of no-show; (ii) cv of service times; and (iii) N, number of patients per session. Implementing a given scheduling rule in different environments will lead to very different combinations of $E(P)$ and $E(M)$ (expected patients' and m.p.'s idle times). Therefore, for the objective of minimizing the weighted sum $\pi E(P) + \omega E(M)$, there is no "one best" scheduling rule; a rule superior for one environment can be very inferior for others.

For environments characterized by 27 different combinations of the factors ρ, cv and N, simulated performance results of nine scheduling rules are presented. These results are presented in the form of "efficient frontiers," together with a procedure to identify the best scheduling rule for a session with given values of ρ, cv, N and ω/π. The procedure can be adapted easily if no quantitative estimates are available for one or more of ρ, cv, N and ω/π.

Together, the eight rules we considered appear to offer an adequate range of slopes to handle most realistic combinations of ρ, cv, N and ω/π. The remaining challenge is to devise new scheduling rules to push the entire frontier further leftward. At this stage, it is amazing to us that, after having tested more than 50 rules (many quite complicated), we are still unable to dislodge the two very simple Bailey-Welch rules (Rules 1 and 5) from the frontier.

References

BAILEY, N., "A Study of Queues and Appointment Systems in Hospital Outpatient Departments, with Special Reference to Waiting-Times," *Journal of the Royal Statistical Society*, A14 (1952), 185–199.

———, "Queuing for Medical Care," *Applied Statistics*, 3 (1954), 137–145.

CHARNETSKI, J., "Scheduling Operating Room Surgical Procedure with Early and Late Completion Penalty Costs," *Journal of Operations Management*, 5 (1984), 91–102.

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

FRIES, B. AND V. MARATHE, "Determination of Optimal Variables-Sized Multiple-Block Appointment Systems," *Operations Research*, 29 (1981), 324–345.

HO, C. AND H. LAU, "Minimizing Total Cost in Scheduling Outpatient Appointments," Working Paper, Oklahoma State University, 1987.

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

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

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

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

WELCH, J., "Appointment Systems in Hospital Outpatient Departments," *Operational Research Quarterly*, 15 (1964), 224–237.

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

WHITE, M. AND M. PIKE, "Appointment Systems in Outpatient's Clinics and the Effect of Patients' Unpunctuality," *Medical Care*, 2 (1964), 133–145.