CrossMark

Scheduling rules to achieve lead-time targets in outpatient appointment systems

Thu-Ba T. Nguyen[1] · Appa Iyer Sivakumar[1,2] · Stephen C. Graves[1,3]

Abstract This paper considers how to schedule appointments for outpatients, for a clinic that is subject to appointment lead-time targets for both new and returning patients. We develop heuristic rules, which are the exact and relaxed appointment scheduling rules, to schedule each new patient appointment (only) in light of uncertainty about future arrivals. The scheduling rules entail two decisions. First, the rules need to determine whether or not a patient's request can be accepted; then, if the request is not rejected, the rules prescribe how to assign the patient to an available slot. The intent of the scheduling rules is to maximize the utilization of the planned resource (i.e., the physician staff), or equivalently to maximize the number of patients that are admitted, while maintaining the service targets on the median, the 95th percentile, and the maximum appointment lead-times. We test the proposed scheduling rules with numerical experiments using real data from the chosen clinic of Tan Tock Seng hospital in Singapore. The results show the efficiency and the efficacy of the scheduling rules, in terms of the service-target satisfaction and the resource utilization. From the sensitivity analysis, we find that the performance of the proposed scheduling rules is fairly robust to the specification of the established lead-time targets.

Keywords Admission policy · Appointment lead-time · Outpatient clinics · Scheduling

1 Introduction

For a medical service or clinic, the appointment lead-time is the time from a patient's request date to his/her actual appointment date. In these settings a long lead-time can be a concern, as it necessarily delays the diagnosis and/or treatment of a patient's medical condition. Long lead-times can also result in increased patient no-show's, which can affect the efficiency with which the clinic operates. Hence, healthcare systems are increasingly monitoring the appointment lead-times of patients for each clinic, and then using this as one measure of performance. The context for this research, namely hospital outpatient services in Singapore, provides one example. The Singapore Ministry of Health (MOH) has set performance guidelines for its hospitals in terms of median, 95th percentile, and 100th percentile targets of appointment lead-times for its subsidized new patients. Tan Tock Seng Hospital (TTSH), which is governed by the MOH, is one hospital that is working to achieve the appointment lead-time targets.

TTSH has separate appointment blocks in its clinics for new patients (only) and for returning patients (only). A new patient is one who requests an initial appointment with a clinic and is not a returning patient; that is, any prior treatments or medical services for the patient have been completed or discontinued. One reason for scheduling new patients

✉ Thu-Ba T. Nguyen
ng0003ba@e.ntu.edu.sg

Appa Iyer Sivakumar
MSIVA@ntu.edu.sg

Stephen C. Graves
sgraves@MIT.EDU

[1] Manufacturing Systems and Technology Program, Singapore-MIT Alliance, 65 Nanyang Drive, Singapore 637460, Singapore

[2] School of Mechanical and Aerospace Engineering, Nanyang Technological University, 65 Nanyang Drive, Singapore 637460, Singapore

[3] Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Springer

separately from the returning patients is to manage the capacity that is allotted for new patients, so as to keep the right balance between new and returning patients. Moreover, the appointment scheduling for the returning patients is usually done immediately after each consultation and its timing is decided by the physician. Therefore, we focus on achieving the lead-time targets for new patients in this paper.

New (subsidized) patients are sent to a hospital based on a referral from a poly-clinic under the Singapore MOH. If the hospital is not able to provide a reasonable appointment lead-time to a referred patient, then it is permissible for the hospital to turn away the patient. In these cases, the patient will be referred to another hospital in the system, or will be directed to go to an open session (open sessions are unscheduled clinics that are not included within the appointment lead-time performance measures).

Nguyen et al. [23, 24] proposed a mathematical model to plan capacity and satisfy the service targets for an outpatient service at TTSH with returning patients (i.e., the treatment of each patient typically entails multiple visits). MOH can use the model in [23, 24] to determine a capacity plan, namely the number of physicians that a hospital would need to achieve the appointment lead-time targets for both new and returning patients. The structure of this capacity plan infers guidelines for how to schedule these patients so as to increase the likelihood of satisfying the service targets for appointment lead-time. The key observation is that patients should generally be scheduled with lead-times that match one of the targets, i.e., matching the median or 95th percentile or 100th percentile target. However, this observation does not translate immediately into implementable scheduling rules, due to the need to account for real-time constraints and considerations. At the operational level, for each patient request, decision have to be made as to whether a patient should be given an appointment or not, and if so when each patient's appointment needs to be scheduled without knowledge of future arrival demand. The scheduler needs to decide on how many arrivals to admit to utilize the planned resource but not to violate the lead-time targets. The scheduler also needs to decide which appointment lead-time (median, 95th percentile or 100th percentile target) to apply to each admitted patient. Therefore, the scheduler needs stopping constraints, which address different established lead-time targets, without knowing the complete set of arrival demands.

This paper considers how to schedule new patients' request, for a given physician plan. We do not consider the appointment scheduling for returning patients. For each returning patient, the physician will decide the timing of the next appointment based on the patient's health condition. Thus, we develop appointment scheduling rules only for new patients with an objective to maximize the utilization of the planned resource subject to assuring the achievement of the lead-time targets. The scheduling rules account for the satisfaction of the median, 95th percentile, and 100th percentile appointment lead-time targets. We also provide quantitative rules for making decision on an individual patient admission, assuming no knowledge of future arrivals. We minimize the number of patients that are turned away for appointments, while still assuring that the appointment lead-time targets are met. These rejections are allowed by the Singapore MOH, as these patients will be able to get medical treatment by other means. Some will be sent to another MOH-governed hospital where they will get a shorter appointment lead-time. Others may be served at the hospital in open clinic sessions for unscheduled patients. However, the specific consideration of how non-admitted patients are handled is beyond the scope of this paper. The problem formulation is shown in Section 3, which follows the literature review in Section 2. Section 4 reports on numerical experiments. We then discuss our findings in Section 5, and finally provide a conclusion in Section 6.

2 Literature review

There have been extensive literatures investigating the scheduling problem in the manufacturing field, known as machine scheduling problem. We categorize the machine scheduling problem into classical scheduling, online scheduling, and bin packing problems. Recent examples of research on classical scheduling problems include multiple machine scheduling to minimize the completion time and/or flow-time in Yenisey and Yagmahan [30], and/or idle time in Kacem and Kellerer [15], due-dates in Janiak et al. [13] and Yenisey and Yagmahan [30]. Similarly, the online scheduling problems consider the completion time, flow-time, and/or due-dates as its objective/s; however, the distinguishing feature is that the scheduling decisions are made in an online fashion, e.g., Liu et al. [21]. The bin packing problem aims to allocate a given set of objects with known sizes to minimize the number of used bins in Kaaouache and Bouamama [14]. Our study seems a variant of the machine scheduling problem with due-dates; however, the decisions need to be made in an online fashion for a known set of tasks and the future arrivals of patients are uncertain. Furthermore, the appointments need to be scheduled so as to keep within the appointment lead-time target. We are not aware of any machine scheduling literature that could apply to this problem setup.

Many studies have addressed the appointment scheduling problem to improve patient access as well as the provider's productivity. These studies measure patients' waiting time, physicians' idle time, and/or physicians' overtime as the objectives. Cayiril and Veral [5], and Lakshmi and Sivakumar [17] summarized the large number of investigations of appointment scheduling. Bailey [1] is one of the pioneers who investigated the relationship between patients' waiting time and physicians' idle time to optimize the number of patients per session and the appointment interval. Many subsequent studies have addressed

the appointment scheduling problem to derive an appointment scheduling rule or to evaluate various proposed appointment policies. An evaluation of the different appointment rules under various scenarios has been conducted in [3, 12, 16, 18, 25, 28, 29]. Furthermore, the development of an appointment scheduling policy to minimize some objective function was examined in [4, 6, 10, 11, 20, 22, 26, 27, 31].

There are several studies that evaluate appointment scheduling under scenarios-based weighted objectives of either patients' waiting time or physicians' idle time or physicians' overtime. Su and Shih [25] paid extra attention to patients' throughput time and a utilization rate of the service provider. Ho and Lau [12], Klassen and Rohleder [16], Su and Shih [25], and LaGanga and Lawrence [18] assume that patients show up on time. Blanco White and Pike [3], and Vissers [28] considered the lateness of patients while Vissers and Wijngaard [29] allow for early arrivals of patients. The effect of physicians' lateness was examined in [3, 28, 29]. Klassen and Rohleder [16] and Su and Shih [25] include walk-in patients in their models.

Brahimi and Worthington [4], Liu and Liu [20], and Hassin and Mendel [11] utilize queueing theory to develop appointment scheduling policies. These models trade off patients' waiting time and physicians' idle time to optimize the service's performance. Brahimi and Worthington [4], and Liu and Liu [20] investigated the impact of no-show rate to the performance. The authors assumed a deterministic appointment interval and a general distribution for service times. Brahimi and Worthington [4] assumed that doctors are punctual, whereas Liu and Liu [20] did not. Hassin and Mendel [11] assumed an exponential service time and no-show rate. The patient is assumed to be punctual in the above studies.

Mathematical programming methods are used to develop an appointment scheduling policy in [6, 10, 22, 26, 27, 31]. Fries and Marathe [10], Vanden Bosch and Dietz [26, 27] schedule a given number of patients into equal slot intervals, with objectives to minimize patients' waiting time, physicians' idle time, and/or physicians' overtime. Fries and Marathe [10] derived a dynamic programming for the case with exponential service time. Vanden Bosch and Dietz [26, 27] develop two heuristic algorithms for general service times, with an assumption of different no-show rates amongst types of patients.

Muthuraman and Lawley [22], Chakraborty et al. [6], and Zeng et al. [31] determine the maximum number of patients that can be scheduled into a given set of slots. These papers propose heuristic algorithms with stopping criteria to schedule patients' appointments into the provided slots. The length of the provided slots may be different. Physicians' idle time and overtime, and patients' waiting time are the performance measurements. Muthuraman and Lawley [22] and Chakraborty et al. [6] consider a patient scheduling problem with a unimodal objective function where the expected profit for a schedule is non-decreasing with the addition of some patient

and then monotone decreasing. The former study assumed an exponential service time, while the latter study assumed a general distributed service time. Zeng et al. [31] solves a multi-modular problem with an exponential distribution of service times. All of the above studies account for no-show rates but exclude the lateness of patients in their models. Fixed and equal slot intervals are assumed in [6, 22], while fixed but different slot lengths are explored in [31].

The above studies primarily try to obtain a balance between the patients' waiting time and some measure of physician utility that accounts for idle time and overtime. Various policies are considered in light of various factors, such as the patients' no-show rate, patients' lateness, and uncertainty of the consultation time. The general intent is to determine how to achieve high resource utilization, along with a high level of patient service quality, and do so in way that can be repeated day after day. However, the long-term appointment lead-time performance is not a major consideration.

Several studies addressed the concern of admission management scheme for a better performance. The literature on admission management develops criteria for making admissions decisions on patients, accounting for the patients' appointment lead-time. Culyer and Cullis [7], Leithch et al. [19], and Goddard and Tavakoli [9] developed priority schemes for admission. Bibi et al. [2] explored the impacts of the centralization and over-booking policies on the number of scheduled patients and on the mean appointment lead-time, while Dixon and Siciliani [8] focused on deriving the relationship between the distribution of scheduling patients and that of waiting time of treated patients.

In general, we did not find any literature that addresses how to set appointments in light of a set of given lead-time targets. Most of the research investigated how to schedule appointments, trading off patients' waiting with the impact on resource utilization. The research on admission policies, such as in [7, 9], did not account for appointment lead-time targets nor the appointment scheduling rule. Hence, this research attempts to fill this gap by considering both appointment scheduling and admissions decisions, in light of lead-time targets.

We derive an admission policy for new patients in light of the service targets on median, 95th percentile, and 100th percentile appointment lead-times, and propose heuristic algorithms to schedule these patients. The rules aim to schedule individual patients' appointment without knowledge of future arrivals. The detailed constraints as well as the appointment scheduling rules are developed in the next section.

3 Appointment scheduling problem

The outpatient appointment system requires new or first-visit (FV) patients to book an appointment in advance. For these requests, the operation staff uses the first-in-first-out (FIFO)

rule to provide an appointment to each FV patient until none of the designated capacity is available. This way of assigning patient appointments intuitively aims to minimize the unused resources in terms of empty slots or unscheduled slots; however, this approach does not consider the satisfaction of the appointment lead-time targets. As a result, it might lead to not meeting the MOH's appointment lead-time targets.

Extending the capacity planning models in [23], we describe a capacity planning model for setting the capacity level based on information of the estimated future demands. However, the actual patient demand will differ from the estimated data and the operation staff does not have full information of the future demands. To help the staff handle each FV patient's request, we develop an appointment scheduling rule in light of the MOH's median, 95th percentile, and 100th percentile appointment lead-time targets.

To develop the appointment scheduling rules, we need to assume that the given capacity for a specific type of patients cannot be used by the other type of patients. In effect, we assume that the given capacity for re-visit (RV) patients of the re-entry system is sufficient to sustain the continuity of care (i.e., there are enough available slots to schedule the next appointment for each RV patient) for all admitted FV patients regardless of the accuracy of the estimated FV demands. The assumption implies that the violation of RV continuity of care (or the violation of a RV patient's appointment lead-time) should not depend on the accuracy of the estimated FV demands. Furthermore, we do not consider the scheduling of RV patients in this study, since the RV patients' appointment relies on the patients' health condition and on the decision of physicians. Therefore, we will only develop appointment scheduling rules for determining the FV patients' appointment. The scheduling rules should determine the patients' appointment from the list of available appointment slots. We assume that the number and timing of the appointment slots have been predetermined and are given; and we assume that this determination has been done in a way that accounts for patient behavior, e.g., no-shows and lateness. Thus, the scheduling rules can be limited to just assigning patients to slots. The assumptions are summarized in Table 1. We define the median, 95th percentile, and 100th percentile time-windows at time "i" as $[i, i + u]$, $(i + u, i + v]$, and $(i + v, i + w]$, respectively, where u, v, and w are the given median, p^{th} percentile, and 100th percentile appointment lead-time targets.

3.1 Patients' admission constraints

The objective of the rules is to minimize the number of unscheduled slots while assuring the achievement of MOH's appointment lead-time targets. Since the rules aim to maintain the MOH's appointment lead-time targets throughout the arrival horizon, the percent of appointments whose appointment lead-time are within the median time-window must be greater than or equal to 50 % of the total number of appointments at

every time-unit. Similarly, the percent of appointments whose appointment lead-time are within the pth percentile time-window must be at least p% (with $50 < p < 100$) of the total number of appointments at every time-unit. In other words, at any time, at most 50 % and $(100 - p)$% of the patients have appointment lead-times greater than the median and p^{th} percentile appointment lead-time targets, respectively. In addition, none of the appointments has a lead-time beyond the 100th percentile lead-time target.

We define the constraints associated with the lead-time targets as the median-constraint, the pth-percentile-constraint, and the 100th-percentile-constraint. These constraints are dynamically executed at each time-unit i. The notation for development of the admission constraints is shown in Table 2.

a.   Exact-admission constraints set

Nguyen et al. [23, 24] found that in the capacity plan, the appointment lead-times for the FV patients were primarily set to one of the MOH targets. That is, the capacity plan tends to set the appointment lead-times of the FV patients to be on the median, p^{th} percentile, or 100th percentile appointment date. Hence, our scheduling policy will first try to schedule a patient's appointment within the median time-window; if this is not possible, then we schedule the patient so that the appointment lead-time is equal to the pth percentile or the 100th percentile lead-time targets.

We specify the admission constraints to assure that we will not violate the MOH lead-time targets. We note that whenever we admit a patient and can schedule the patient within the median time-window, then this increases the number of appointments satisfying the median target and hence, raises the total number of patients that can be admitted.

We state in Equations (1a) to (6a) the "*exact-admission constraints set*" for one time period at time-unit i to assure that each of the lead-time targets is not violated.

[Exact-admission constraints set]

$$c_i^m = \sum_{j=i}^{j=i+u} b_{ij} \tag{1a}$$

$$c_i^p = Min_{j=i+v}\left\{ b_{ij}, \frac{p}{50} A_i^m - \left(A_i^m + A_{i-1}^p \right) \right\} \tag{2a}$$

$$c_i^{100} = Min_{j=i+w}\left\{ b_{ij}, \frac{100}{p}\left(A_i^m + A_i^p \right) - \left(\left(A_i^m + A_i^p + A_{i-1}^{100} \right) \right) \right\} \tag{3a}$$

where:

$$A_i^m = A_{i-1}^m + a_i^m \tag{4a}$$

$$A_i^p = A_{i-1}^p + a_i^p \tag{5a}$$

$$A_i^{100} = A_{i-1}^{100} + a_i^{100} \tag{6a}$$

Table 1 List of assumptions for the appointment scheduling rules

No.	Assumptions
1	The planned capacity for each type of patient cannot be used by the other type of patient.
2	The rejection of patient-requests is permissible.
3	The appointment scheduling rules are only applied for the FV patient-requests.
4	Patients' preference on either a physician or an appointment time-unit is not considered.
5	A patient's appointment can be scheduled at the same time as the patient's request.
6	All scheduled patients show up.
7	The appointment lead-time targets are never violated.
8	Future demand is unknown.
9	One slot can be assigned to at most one patient.
10	The planned capacity for re-visit (RV) patients of the re-entry system is sufficient to sustain the continuity of care for all admitted FV patients.

Equation (1a) determines an upper bound on the number of patients that arrive in time-unit i and that can be scheduled within the median time-window ($[i, i + u]$). Similarly, Equation (2a) gives an upper bound as it relates to the pth-percentile-constraint in time-unit i. It determines an upper bound on the number of patients that arrives in time-unit i and that can be scheduled with an appointment lead-time set to the pth percentile lead-time target (v). On the right hand side of Equation (2a), the first term is the number of available slots for which the appointment lead-times are at the pth percentile lead-time target. The second term assures that the pth percentile lead-time target is not violated; namely it is the maximum number of patients that can be scheduled with an appointment lead-time between the median and the pth percentile lead-time target and still assure that p % of the patients have a lead-time within v time units. Equation (3a) applies to the 100th percentile constraint in time-unit i. It determines the maximum number of patients that arrives in time-unit i and that can be scheduled with an appointment lead-time equal to the maximum lead-time target (w). The structure and explanation are the same as for Equation (2a).

Equations (4a), (5a), and (6a) specify the cumulative number of scheduled patients in time-unit i, whose lead-times are within the targets of median time-window, at the pth percentile appointment lead-time, and at the 100th percentile appointment lead-time, respectively. The number of each type of the scheduled patients in each time-unit i (a_i^m, a_i^p, or a_i^{100}) is determined at the end of each time-unit. To assure no violation of the lead-time targets, the scheduling must abide by the upper bounds computed in (1a) – (3a); that is, $a_i^m \le c_i^m$, $a_i^p \le c_i^p$, and $a_i^{100} \le c_i^{100}$.

b. Relaxed-admission constraints set

The above "*exact-admission constraints set*" is restrictive in that for patients that have appointment lead-times greater than the median lead-time target, it restricts the lead time to being either v or w (the p % or 100 % target). To assess the impact of these restrictions, we develop an alternative set of constraints, named the "*relaxed–admission constraints set*".

[Relaxed-admission constraints set]

$$c_i^m = \sum_{j=i}^{j=i+u} b_{ij} \tag{1b}$$

$$c_i^p = Min\left\{ \sum_{j=i+u+1}^{i+v} b_{ij}, \frac{p}{50} A_i^m - (A_i^m + A_{i-1}^p) \right\} \tag{2b}$$

$$c_i^{100} = Min\left\{ \sum_{j=i+v+1}^{i+w} b_{ij}, \frac{100}{p}(A_i^m + A_i^p) - ((A_i^m + A_i^p + A_{i-1}^{100})) \right\} \tag{3b}$$

where:

$$A_i^m = A_{i-1}^m + a_i^m \tag{4b}$$

$$A_i^p = A_{i-1}^p + a_i^p \tag{5b}$$

$$A_i^{100} = A_{i-1}^{100} + a_i^{100} \tag{6b}$$

The *relaxed–admission constraints set* again determines upper bounds for the number of patients that can be scheduled with appointment lead-times within the median, within the pth percentile, and within the maximum lead-time target. The upper bounds again assure that the MOH's targets are maintained. However, now, the patients are allowed to have appointments within the pth percentile ($(i + u, i + v]$) or within the 100th percentile ($(i + v, i + w]$) time-window if they cannot get the appointments within the median ($[i, i + u]$) or p^{th} percentile ($(i + u, i + v]$) time-window, respectively. The development of the relaxed-admission constraints set is similar to that of the *exact-admission constraints set*.

3.2 Appointment scheduling rule

The appointment scheduling rule determines for each patient an appointment time from the calculated admission

Table 2 List of notation for development of the admission constraints

No.	Notation	Description
1	A_i^m	The cumulative number of admitted FV patients whose appointment date is scheduled within the median lead-time target, as of time-unit i.
2	A_i^p	The cumulative number of admitted FV patients whose appointment date is scheduled beyond the median lead-time targets but within the pth percentile lead-time target, as of time-unit i.
3	A_i^{100}	The cumulative number of admitted FV patients whose appointment date is scheduled beyond the pth percentile appointment lead-time but within the 100th percentile lead-time target, as of time-unit i.
4	a_i^m	The number of admitted FV patients who arrive at time-unit i and are scheduled so that their appointment lead-times are equal to or less than the median lead-time target.
5	a_i^p	The number of admitted FV patients who arrive at time-unit i and are scheduled so that their appointment lead-times are beyond the median appointment lead-time but within the pth percentile lead-time target.
6	a_i^{100}	The number of admitted FV patients who arrive at time-unit i and are scheduled so that their appointment lead-times are beyond the pth percentile appointment lead-time but within the maximum lead-time target.
7	b_{ij}	The number of available FV slots in time-unit j, measured at time-unit i.
8	c_i^m	A median upper bound at time-unit i.
9	c_i^p	A pth-percentile upper bound at time-unit i.
10	c_i^{100}	A 100th-percentile upper bound at time-unit i.
11	u, v, w	Lead-time targets for median, for p^{th} percentile and for 100th percentile

constraints (median, p^{th} percentile, and 100th percentile constraints) to maintain the satisfaction of the MOH's appointment lead-time targets at all times. The rule is myopic in that it admits and schedules the patients without considering future arrival information and with the assumption that the appointments cannot be changed. When scheduling a patient, the scheduling rule prioritizes the constraints in the order of median, p^{th} percentile, and 100th percentile constraints. We formulate the appointment scheduling rule (R1) as follows:

[R1]:
Step 1:
Given i. Set: $a_i^m = 0$; $a_i^p = 0$; $a_i^{100} = 0$.
Step 2:
Determine the latest appointment date $j = i + u$ so that the patient's appointment lead-time satisfies the median lead-time target (u).
Determine the median constraint c_i^m from (1a) or (1b).
Step 3:
Consider the patient requests one by one. Schedule each patient between period i and period j, inclusive until the constraint c_i^m is binding or until all patients are scheduled. If all the requests are scheduled, go to step 8. Otherwise, update a_i^m and go to step 4.
Step 4:
Determine the latest appointment date $j = i + v$ at which the patient's appointment lead-time satisfies the pth percentile lead-time target (v).

Determine the pth percentile constraint c_i^p from (2a) or (2b).
Step 5:
Consider the patient requests and schedule them one by one. Schedule each patient either in time period j or within time period j until the constraint c_i^p is binding or until all patients are scheduled. If all the requests are scheduled, go to step 8. Otherwise, update a_i^p and go to step 6.
Step 6:
Determine the latest appointment date $j = i + w$ at which the patient's appointment lead-time does not exceed the maximum lead-time target (w).
Determine the 100th percentile constraint c_i^{100} from (3a) or (3b).
Step 7:
Consider the patient requests and schedule them one by one. Schedule each patient either in time period j or within time period j until the constraint c_i^{100} is binding or until all patients are scheduled. Go to step 8.
Step 8:
Update: a_i^m, a_i^p, and a_i^{100}.

The scheduling rule is executed in each period i (or day i) for some scheduling horizon; for steps 5 and 7, either the *exact* or *relaxed* admission constraints can be used. Any patients that cannot be scheduled in a period are not admitted; these patients may be transferred to another hospital or assigned to

an open session. The rule assures that the appointment lead-time targets are maintained at all times.

Figure 1 illustrates the flow of the proposed appointment scheduling rules. The second column describes the scheduling conditions subject to median constraint (steps 3 and 4 of the rule R1). Similarly, the third and fourth columns present the scheduling conditions subject to p^{th} percentile and 100th percentile constraints, respectively. The last column is to reject all the remaining patients' requests when no available slot remains. The appointment scheduling rule R1 is called the "*exact rule*" if the *exact-admission constraints set* is used to admit the patients. Otherwise, we have the "*relaxed rule*" if the *relaxed-admission constraints set* are employed. We consider two variants of the *relaxed rule*, which are described below.

With the *exact rule*, we assume that the first-in-first-out (FIFO) sub-rule is used for the patients who get scheduled within the median time-window. In other words, each patient's appointment date is scheduled in the earliest available slot within the median time-window. The FIFO rule aims to minimize the number of the unused slots. If patients cannot get the appointments within the median time-window, the pth percentile appointment date will be considered before considering the 100th percentile appointment date. The procedure continues until the schedule horizon S is exceeded.

With the *relaxed rule*, if a patient cannot get the appointment within the median time-window, he/she will get an available slot within the pth percentile time-window. Admitting the patient to an available slot subject to the 100th percentile time-window is the last choice. For the patients that are scheduled within the median time-window, they are scheduled with a FIFO sub-rule. However, for appointments assigned to either the pth time-window or the 100th time-window, we use the first-in-last-out (FILO) sub-rule; that is, each patient is scheduled as late as possible but within the assigned time-window, namely either the pth-percentile or 100th percentile time-window. The intent of the FILO rule is to allow more patients to be scheduled in the future, which leads to more admitted patients and fewer unused slots. The FILO rule attempts to protect appointment slots that can be used in the future to schedule patients within the median lead-time target; as more patients are scheduled within the median target; then more patients can be admitted to within the pth-percentile-constraint. For example, in time-unit i, the usage of the FILO for p^{th} percentile time-window allows more patients to be scheduled within the median time-window in the next time-unit $(i + 1)$. This then allows for more patients to be scheduled within the pth percentile time-window in time-unit $(i + 1)$, which leads to the higher total number of patients that can be admitted in time-unit $(i + 1)$ and so on.

The first variant of the *relaxed rule* is the policy "*F2L1*". The patient's appointment is assigned to the median, 95th percentile, or 100th percentile time-windows, as done for the *relaxed rule* and as shown in Fig. 1. For the patients that are
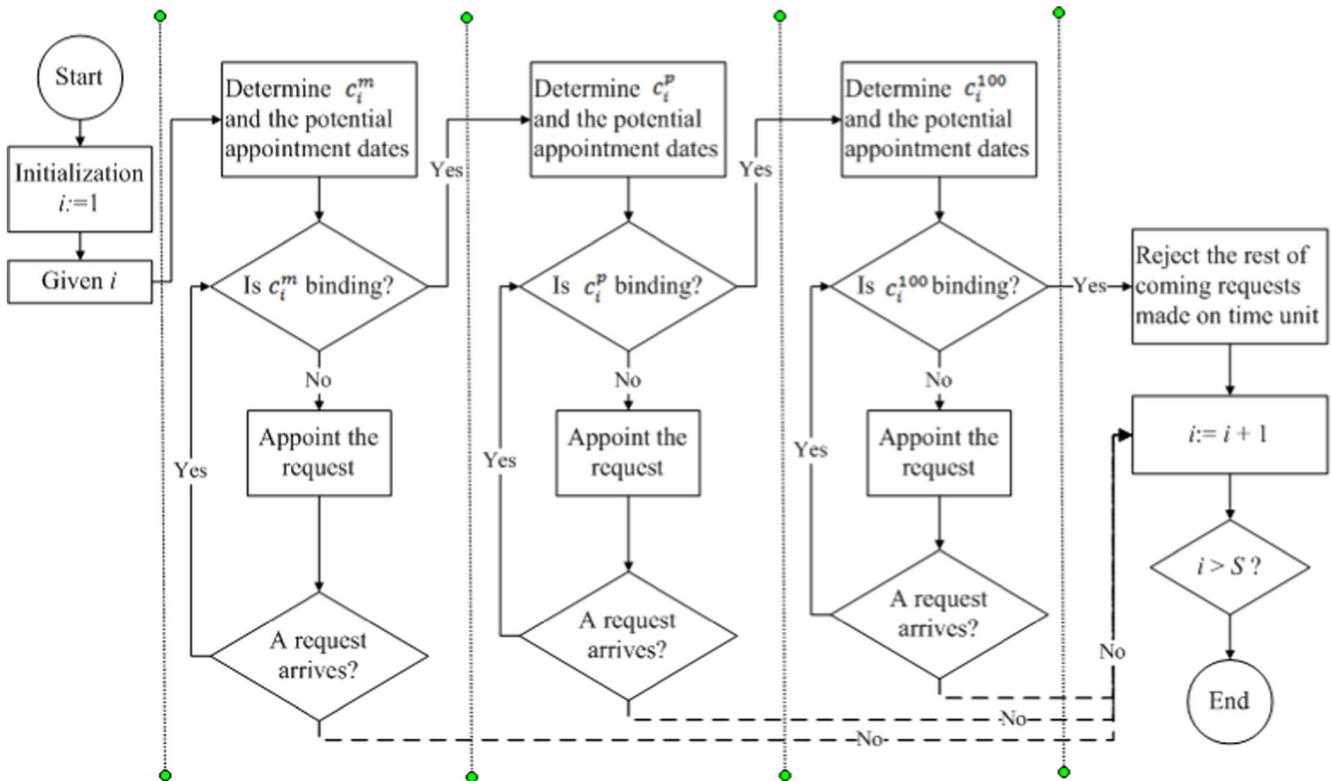

Fig. 1 Flowchart of the appointment scheduling rules

scheduled within the median and p^{th} percentile time-windows, this policy schedules the patient with the FIFO sub-rule. The policy uses the FILO sub-rule to schedule any patient's appointment within the 100th percentile time-window.

The second variant of the *relaxed rule* is the policy "*F3*". This policy uses the FIFO sub-rule for all patients assigned to each time-window. Table 3 summarizes the sub-rules for each time-window for the *exact rule*, for the *relaxed rule* and for the two policy variants.

4 Numerical experiments

We have conducted numerical experiments to test the effectiveness of the proposed appointment scheduling rules. The rules are evaluated on either the percentage of the unused slots or the percentage of the rejected requests.

4.1 Experimental design

We use data from years 2009 and 2010 from the Urology specialty at TTSH. We use this data to determine the inputs on the arrival demands as well as the actual provided capacity. The median, 95th percentile, and 100th percentile appointment lead-time targets are 2 weeks, 6 weeks, and 9 weeks, respectively. We generate test problems by varying the patient discharge rate, the mean appointment lead-time for the RV patients, and the appointment lead-time targets for the FV patients. The optimal planned capacity for the test scenarios is obtained from the model that was developed in Nguyen et al. [23]. We consider 5 different levels for the discharge rate, 6 levels of the RV's mean appointment lead-time, and 15 sets of the FV's appointment lead-time targets. For each test scenario, we determine the optimal capacity plan, using the model described in [23]. Consequently we have $1 + 5 \times 6 \times 15 = 451$ test cases, which correspond to the base case with the actual capacity plan, plus 450 test scenarios, each with a different optimal planned capacity for FV patients.

For each test scenario, the inputs are the planned capacity for serving the FV patients, the FV demands in each time-unit i, the targets for median, 95th percentile, and 100th percentile appointment lead-times. The experiments are done for the *exact rule*, the *relaxed rule*, and the policy variants "*F2L1*" and "*F3*". We simulate one year, corresponding to the patient arrivals in 2009. We start the simulation with the existing appointments being given, as of Jan. 1st, 2009; that is, an input is the appointments that were scheduled in 2008 for 2009. The computation time to simulate the rules, which are built in spreadsheets with Visual Basic Application, is less than 30 seconds for each test case.

4.2 Experimental results

For the base case with the actual demand, actual capacity, and actual lead-time targets, the MOH targets could not be met; the actual patients' appointment lead-times for median, 95th percentile, and 100th percentile were 2.1, 2.0, and 1.7 times as much as those from the MOH's guideline, respectively. When we simulate the base case, we find that for both the *exact rule* and *relaxed rule* we need to reject 8.1 % of the demand to achieve the three appointment lead-time targets. The policy variants "*F2L1*" and "*F3*" can obtain better lead-time performances but result in an even higher rejection rate. The policy variant "*F2L1*" obtains 3 weeks versus the 95th percentile lead-time target of 6 weeks, while the policy variant "*F3*" achieves 3 and 7 weeks versus the 95th percentile and 100th percentile lead-time targets of 6 and 9 weeks, respectively. However, the policy variants "*F2L1*" and "*F3*" need to reject an additional 1.1 % and 1.4 % of the demand, respectively (Table 4). The rejection happens to the arrivals in the first 8 weeks of the year. During the first 8 weeks of 2009, the total number of arrivals was 1382 patients; however, the total number of available slots was only 487 slots during the first 10 weeks of 2009. After the first 10 weeks, though, a very high number of slots were planned. Due to the limited capacity in the first 10 weeks, a large percent of patients had to be rejected in order to not violate the median appointment target of 2 weeks. Beyond this time interval there was sufficient capacity, but this could not be used. The results show the

Table 3 Summary of the sub-rules of the exact and relaxed rules

| Constraint for	Decision on the patients' appointment			
	The exact rule	The relaxed rule	Policy variant	
			F2L1	F3
Median time-window	The FIFO sub-rule	The FIFO sub-rule	The FIFO sub-rule	The FIFO sub-rule
p^{th}-percentile time-window	Schedule exactly at the pth-percentile appointment lead-time v.	The FILO sub-rule		
100th-percentile-time-window	Schedule exactly at the 100th-percentile appointment lead-time w.		The FILO sub-rule	

Table 4 Performance of the proposed rules for the base case

Policy		Rejected demand		Lead-time performance (time-unit)		
		Number of patients	%	Median	95th percentile	100th percentile
Exact rule		795	8.1	2	6	9
Relaxed rule		795	8.1	2	6	9
Policy variant	*F2L1*	902	9.2	2	3	9
	F3	922	9.5	2	3	7

potential value for an implementation of the proposed *exact rule* and *relaxed rule*.

For the remaining test problems, we report the percentage of the unused slots. For each test case, more unused slots means that there were more patients were rejected; hence, fewer unused slots is better. In Fig. 2, we compare the cumulative frequency of the percentage of the unused slots for the optimal capacity plan for these test cases. The *exact* and *relaxed* rules dominate the policy variants *F2L1* and *F3* in terms of the capacity utilization (Table 5). The average unused capacity from the policy variant "*F2L1*" (4 %) is 2.6 and 4 times as much as that from the *exact rule* (1.5 %) and *relaxed rules* (1 %), respectively. For the policy variant "*F3*", the average percentage of the unused capacity (5 %) is 3.3 and 5 times as high as that from the *exact* (1.5 %) and *relaxed* (1 %) rules, respectively. The maximum percentage of the unused capacity is 13.3 % for policy variant "*F2L1*" and 17.5 % for "*F3*". In comparison, the maximum percent of unused capacity is only 7 % for the *exact* and *relaxed* rules.

Moreover, the exact rule does the same or better than the policy variants "*F2L1*" and "*F3*" in 87.3 % and 91.3 % of the test cases, respectively (Table 6). From Table 6, we see that the *relaxed rule* has the same or fewer unused slots in 97.7 % of cases, compared to "*F2L1*", and in 89.6 % of cases, compared
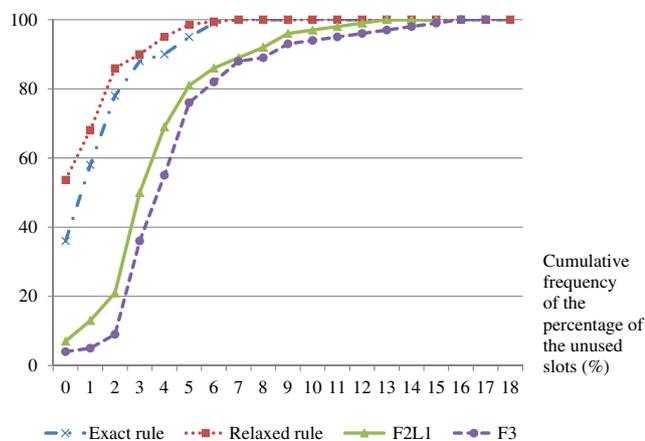
to "*F3*". These results suggest that the *exact* and *relaxed* rules should be used for better resource utilization under the restriction of achieving the established lead-time targets.

In comparing the *relaxed rule* to the *exact rule*, we see from Fig. 2 and Table 5 that the *relaxed rule* performs better across the population of test problems. From Table 6, we see that the *exact rule* performs better than the relaxed rule in 12.7 % of the test cases, and that the *relaxed rule* does better than the *exact rule* in 37.9 % of the test cases. We did not discern any specific pattern for differentiating the scenarios in which the *exact rule* outperforms the *relaxed rule*, or vice-versa.

The findings show that although the *relaxed* scheduling rule might (on average) improve the utilization of the planned resource, the *exact rule* performs nearly as well or better in many instances. Therefore, to help the scheduling achieve the lead-time targets, the *exact rule* might be preferred due to its simplicity. The *relaxed rule* can be considered whenever a high-priority patient cannot be scheduled using the *exact rule*; for example, this might be a severe case that gets an approval from a physician but no slot identified from the *exact rule* is available.

The numerical experiments show that the proposed *exact* and *relaxed* appointment scheduling rules are effective and efficient in guiding the assignments of the patients' requests. The proposed appointment *exact* and *relaxed* scheduling rules help to utilize the resource in which the future arrivals of FV patients are unknown but still maintaining the MOH's appointment lead-time targets at all times. The analysis signifies that patients should have the pth percentile or the 100th percentile appointment date if they cannot get appointments within the median time-window. This result is coincident with the findings in [23]. However, the relaxation of the patients' appointment dates restricted by the pth percentile or the 100th percentile target may be considered only as it is necessary (i.e., exceptional cases that get approval from physician/s). This relaxation has a negligible impact onto the performance.

We examine the sensitivity of the performance of the *exact* and *relaxed* rules to changes of the appointment lead-time targets. Table 7 summarizes the analysis of 450 test cases. We find that the percentage of the unused

Fig. 2 Cumulative frequency of the percentage of the unused slots

Table 5 Performance of the proposed appointment scheduling rules

Policy		Percentage of the unused slots (%)					
		Min	Median	90th percentile	95th percentile	Max	Mean
Exact rule		0	1	4	4.5	7	1.5
Relaxed rule		0	0	3	3.5	7	1
Policy variant	*F2L1*	0	3.6	7.6	9.3	13.3	4
	F3	0	4.3	8.2	10.1	17.5	5

slots does not depend on the length of the established appointment lead-time targets; there is no significant correlation to the changes of the lead-time targets (Table 7). This finding implies that any change of the guideline would just impact the patients' lead-time, but not the utilization of the systems. The patients' appointment lead-time for median, 95th percentile, and 100th percentile will be equal to the established service targets due to the structure of the scheduling rule; hence, any change of the lead-time targets changes the patients' lead-time in practice. In other words, a commitment of guided appointment scheduling rules to admit patients will lead to the achievements of the established lead-time targets, while the planned resource promises to be highly utilized.

5 Discussions

We have proposed the *exact* and *relaxed* appointment scheduling rules and two policy variants "*F2L1*" and "*F3*" for making decisions on the patients' appointments. Firstly, the exact appointment scheduling rule (or *exact rule*) aims to schedule the patients onto the pth percentile or the 100th percentile appointment date if the patients cannot get an appointment within the median time-window. Secondly, the relaxed appointment scheduling rule

(or *relaxed rule*) allows the appointment dates of patients to be more flexible. If patients have to wait longer than the median appointment lead-time target, and none of the potential slots is available on the pth percentile or the 100th percentile appointment date, then the patients are scheduled within the pth percentile or the 100th percentile time-window, respectively. Thirdly, the policy variant "*F3*" uses FIFO sub-rule to schedule the patients' appointment within the three time-windows. Finally, the policy variant "*F2L1*" uses FIFO to schedule the patients' appointment within the median and 95th percentile time-windows restricted by the achievement of the median and 95th percentile lead-time targets, respectively; the FILO sub-rule is used to schedule patients' appointment within the 100th percentile time-window. We have developed admission constraints (median, 95th percentile, and 100th percentile constraints) to schedule each patient's appointment so that the established appointment lead-time targets are maintained in any time-unit. The numerical investigation for the effectiveness of the proposed rules shows that the exact appointment scheduling rule should be considered, since its mean percentage of unused slots is similar to that of the relaxed rule and dominates those from the policy variants "*F2L1*" and "*F3*". The *relaxed rules* should be used as a guideline to physicians for deciding first whether to accept exceptional cases and

Table 6 Performance comparison between the rules

Policy			The number of cases that (A) performs better than or equal to (B)					
			(B)					
			Relaxed rule	Policy variant		Relaxed rule	Policy variant	
				F2L1	*F3*		*F2L1*	*F3*
Exact rule	(A)	Better	57	346	401	12.7 %	76.7 %	88.9 %
		Equal to	223	48	11	49.4 %	10.6 %	2.4 %
		Total	280	394	412	62.1 %	87.3 %	91.3 %
Relaxed rule		Better	-	370	245	-	82.0 %	54.3 %
		Equal to	-	71	159	-	15.7 %	35.3 %
		Total	-	441	404	-	97.7 %	89.6 %
Total test cases			451			100 %		

Table 7 Summary of the sensitivity analysis of the *exact* and *relaxed* rules to the changes of the appointment lead-time targets

RV's mean lead-time		Percentage of the unused slots										Notes:	
	FV discharge rate	0.38	0.38	0.38	0.5	0.7	0.38	0.38	0.38	0.5	0.7		
	RV discharge rate	0.1	0.2	0.32	0.32	0.32	0.1	0.2	0.32	0.32	0.32		
	FV lead-time	Exact rule					Relaxed rule						
5	Median	1	0	1	1	1	1	0	1	1	1	Sign	Description
	95th percentile	1	−	1	1	1	1	1	1	1	1	−	A negative correlation
	100th percentile	1	+	1	+	−	1	+	1	1	1	+	A positive correlation
10	Median	1	1	1	1	1	1	1	1	1	1	1	A similar performance
	95th percentile	−	1	1	1	1	1	1	1	1	1	0	An indeterminate correlation
	100th percentile	1	1	+	1	1	1	1	+	1	1		
16	Median	1	0	1	1	1	1	0	1	1	1		
	95th percentile	1	1	1	1	1	1	1	1	1	1		
	100th percentile	1	0	1	1	1	1	0	1	1	1		
20	Median	1	1	1	−	−	1	1	1	1	1		
	95th percentile	1	1	1	1	1	1	1	1	1	1		
	100th percentile	1	−	1	1	1	1	1	1	1	1		
25	Median	1	−	1	1	−	1	−	1	1	−		
	95th percentile	1	1	1	1	1	1	−	−	1	1		
	100th percentile	1	1	1	1	1	1	0	0	1	1		
30	Median	1	1	−	−	−	1	−	1	−	−		
	95th percentile	1	1	1	1	1	1	0	−	1	1		
	100th percentile	+	1	1	1	1	+	1	0	1	1		

then to set the patient appointment date, possibly hurting the hospital appointment lead-time performance.

We believe that the study contributes to the research literature on appointment scheduling in which the appointment lead-time targets must be satisfied. The proposed *exact* and *relaxed* rules are simple to implement and have been shown to be effective. However, the scheduling rules are proposed to only address a single type of patients for a single set of lead-time targets. For a system with different categories of patients, we would need to develop an extension to the proposed rule. Hence, future research should examine how to address simultaneously multiple set of appointment lead-time targets.

6 Conclusions

In this study, we propose practical appointment scheduling rules (*exact* and *relaxed* rules) with admission constraints to maintain the MOH's lead-time targets and to achieve high utilization of the resource as well. If patients cannot get an appointment within the median lead-time target, they should be scheduled at the pth percentile or maximum lead-time target, in that priority. In addition, the achievement of the service targets and an optimal utilization of the provided resource can be obtained with the commitment of guided scheduling rules.

We compare the proposed rules with the policy variants "F2L1" and "F3", from which we establish the efficacy of the scheduling sub-rules.

There are limiting assumptions that underlie the research: we assume that appointment blocks exist for new patients that allow the new patients to be scheduled separately from returning patients. As a consequence, the research just focuses on the scheduling of the new patients. Additionally, we assume that it is permissible to not admit new patients, if this would violate one of the appointment lead-time targets. These assumptions are specific to the research context, and thus, the findings herein might not apply to other systems. Future work is warranted to examine how to adapt the findings when these assumptions are relaxed. In addition, the assumptions of a single set of the established service targets could limit the applicability of the proposed scheduling rules. Hence, future work should consider multiple appointment lead-time targets to facilitate the staff in scheduling multiple types of patients.

References

1. Bailey NTJ (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. J R Stat Soc 14(2):185–199

2. Bibi Y, Cohen AD, Goldfarb D, Rubinshtein E, Vardy DA (2007) Intervention program to reduce waiting time of a dermatological visit: managed overbooking and service centralization as effective management tools. Int J Dermatol 46(8):830–834

3. Blanco White MJ, Pike MC (1964) Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. Med Care 2(3): 133–141 + 144–145.

4. Brahimi M, Worthington DJ (1991) Queueing models for out-patient appointment systems - a case study. J Oper Res Soc 42(9):733–746

5. Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Prod Oper Manag 12(4):519–549

6. Chakraborty S, Muthuraman K, Lawley M (2010) Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Trans 42(5):354–366

7. Culyer AJ, Cullis JG (1976) Some economics of hospital waiting lists in the NHS. J. Soc. Policy 5(3):239–264

8. Dixon H, Siciliani L (2009) Waiting-time targets in the healthcare sector: how long are we waiting? J Health Econ 28(6):1081–1098

9. Goddard J, Tavakoli M (2008) Efficiency and welfare implications of managed public sector hospital waiting lists. Eur J Oper Res 184(2):778–792

10. Fries BE, Marathe VP (1981) Determination of optimal variable-sized multiple-block appointment systems. Oper Res 29(2):324–345

11. Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. Manag Sci 54(3):565–572

12. Ho CJ, Lau HS (1992) Minimizing total cost in scheduling outpatient appointments. Manag Sci 38(12):1750–1764

13. Janiak A, Janiak WA, Krysiak T, Kwiatkowski (2015) A survey on scheduling problem with due windows. Eur J Oper Res 242:347–357

14. Kaaouache MA, Bouamama S (2015) Solving bin packing problem with a hybrid genetic algorithm for VM placement in cloud. Procedia Comput Sci 60:1061–1069

15. Kacem I, Kellerer H (2014) Approximation algorithms for no idle time scheduling on a single machine with release times and delivery times. Discret Appl Math 164:154–160

16. Klassen KJ, Rohleder TR (1996) Scheduling outpatient appointments in dynamic environment. J Oper Manag 14(2):83–101

17. Lakshmi C, Sivakumar AI (2013) Application of queueing theory in health care: A literature review. Oper Res Health Care 2(1–2):25–39

18. LaGanga LR, Lawrence SR (2007) Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 38(2):251–276

19. Leitch AG, Parker S, Currie A, King T, McHardy GJR (1990) Evaluation of the need for follow-up in an out-patient clinic. Respir Med 84(2):119–122

20. Liu L, Liu X (1998) Block appointment systems for outpatient clinics with multiple doctors. J Oper Res Soc 49(12):1254–1259

21. Liu M, Xu Y, Chu C, Zheng F (2009) Online scheduling to minimize modified total tardiness with an availability constraint. Theor Comput Sci 410:5039–5046

22. Muthuraman K, Lawley M (2008) A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Trans 40(9):820–837

23. Nguyen TBT, Sivakumar AI, Graves SC (2015) A network flow approach for tactical resource planning in outpatient clinics. Health Care Manag Sci 18(2):124–136

24. Nguyen TBT (2015) Modelling, analysis, and optimization in resource planning for outpatient clinics. PhD thesis, Singapore – Massachusetts Institute of Technology Alliance Programme

25. Su S, Shih CL (2003) Managing a mixed-registration-type appointment system in outpatient clinics. Int J Med Inform 70(1):31–40

26. Vanden Bosch PM, Dietz DC (2000) Minimizing expected waiting in a medical appointment system. IIE Trans 32(9):841–848

27. Vanden Bosch PM, Dietz DC (2001) Scheduling and sequencing arrivals to an appointment system. J Serv Res 4(1):15–25

28. Vissers J (1979) Selecting a suitable appointment system in an outpatient setting. Med Care 17(12):1207–1220

29. Vissers J, Wijngaard J (1979) The outpatient appointment system: Design of a simulation study. Eur J Oper Res 3(6):459–463

30. Yenisey MM, Yagmahan B (2014) Multi-objective permutation flow shop scheduling problem: literature review, classification and current trends. Omega 45:119–135

31. Zeng B, Turkcan A, Lin J, Lawley M (2010) Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Oper Res 178(1):1030–1047