3rd National Conference on Innovative Research Trends in Computer Science and Technology (NCIRCST 2018) ISSN: 2454-4248
Volume: 4 Issue: 3 106 – 114

# A Real-Time Access Control of Patient Service in the Outpatient Clinic

A.Prabhakaran
UG Scholar,IT
S.A.Engineeering College

S.Mohan
UG Scholar,IT
S.A.Engineering College

P.J.Sathish Kumar
Assistant Professor,IT
S.A.Engineering College

***Abstract***—With the increasing demand of patients for limited healthcare services, e.g., expert physicians in general hospitals and the long on-site waiting time of patients, a real-time admission control (AC) policy was developed considering both distinction and fairness among heterogeneous patients. The AC policy chooses the customer for the next service from the waiting area to minimize the expected total disutility according to the current system state. A continuous-time Markov decision process in a finite horizon was developed, and a myopic optimal policy was derived using a switching curve to discriminate patient admission based on patient types and waiting time. The effect of each parameter on the switching curve is analyzed, and managerial insights are discussed. In the numerical experiment, an empirical case in the Pediatric Department of Peking University

*Note to Practitioners*—Healthcare facilities with limited resources, for example, emergency departments in the U.S. and expert physicians in China, are always inundated with waiting patients. To guarantee the real-time property of the control policy, a discriminating inequality with a simple switching curve, which was derived from the myopic optimal situation, is proposed. With the implementation of an automatic call-in system based on this admission control policy, outpatient clinics can immediately decide which patient to prioritize, and a tradeoff between patient distinction and fairness can be warranted.

***Index Terms***—*Admission control (AC), healthcare management, Markov decision process (MDP), myopic optimal policy, patient utility.*

*****

## I. INTRODUCTION

The most distinct challenge in Chinese healthcare service is that the supply cannot meet the demand, particularly for specialist (expert) medical resources. The no-show rate of appointments is relatively low because of the difficulty in registering in a specialist department. Patients in China can access a specialist health care service in two ways. One method is through an online appointment system. The other is by directly visiting a hospital. These approaches have their own strengths and weaknesses. Patients adopting advance appointment are guaranteed admission, but they may have to endure longer waiting time. Walk-in patients may not be able to secure an appointment, because a specialist only offers limited consultant capacities in one day. However, indirect waiting time may be shortened if their requests can be accepted. Thus, two classes of patients are considered: *advance appointment patients* who reserve appointments in advance and *walk-ins* who come to a clinic for the same-day service. The service process in an outpatient clinic is presented in Fig. 1.

When a patient's request is accepted by a hospital, he or she is given a registration number that indicates the patient's turn in the waiting area. Advance appointments close when a walk-in window opens at the start of office hours. Thus, advance appointment patients obtain a sequence index prior to receiving service in a hospital, and walk-ins confirm their sequence on site. For a physician with a workload of 30 patients daily, numbers 1–12 are for appointment patients and numbers 13–30 are for walk-in patients. Each patient has a unique sequence index, and no group visit is allowed, which is constricted by the regulation "one registration for one person." Hospitals in Beijing only provide each patient an appointment block, e.g., morning or afternoon, rather than an individual time slot because of the probability of increased traffic around the hospital and patients' tardiness. All patients need to wait in the waiting area until they are called in by a doctor. From the hospitals perspective, this type of appointment system scheme is more suitable for Beijing hospitals than the "time-slot" design that is widely used in developed countries, because specialist healthcare resources in China are more limited considering the large population. The "nonslot" scheme guarantees high resource utilization in a heavy-arrival stochastic service system. However, this configuration also results in a critical issue. Most hospitals assign advance appointment patients with low sequence indexes, whereas they provide walk-in patients larger sequence indexes. To encourage advance appointments to effectively prepare for staff allocation, hospitals apply the "sequence-index-based" policy (SIBP) to the call-in process, where doctors serve patients according to the sequence index

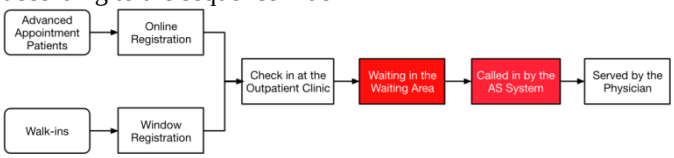



Fig. 1. Healthcare service process in China's outpatient clinics.

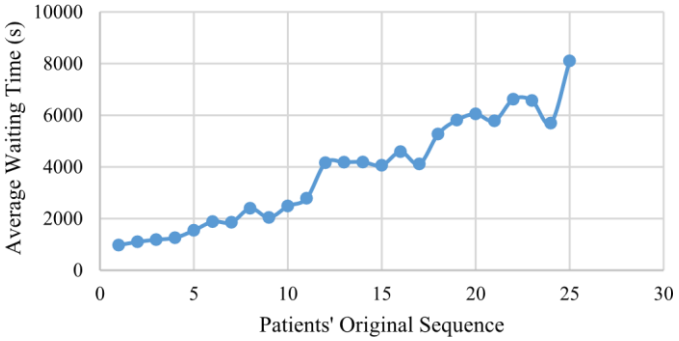



Fig. 2. Patients' average waiting time in the outpatient clinic of the PUTH,

Beijing.

in ascending order. However, advance appointment patients arrive unexpectedly. Patients with low sequence indexes are allowed to "cut in line," which leads to an extremely long direct waiting time of walk-ins holding large sequence indexes. Thus, unfairness occurs among patients in the same block.

The multiple choices problem commonly occurs when the staff determines which patient should be called-in next, since the backlog in the waiting area always exists in reality. According to our literature review, most of the papers consider a system where patients have a specific appointment time slot with the physician [2]–[9]. However, we consider a heavy demand healthcare system where patients' arrival begins anterior to the start point of physician's work, inevitably causing an original backlog in the waiting area. Therefore, when a new diagnosis begins, the scheduler always needs to select a patient from multiple alternatives, and generally, different patients do not have the same priority. They possess dynamic priorities that depend on both the patient class and the duration they have been waiting.

Furthermore, the conference version of this paper is extended for the following reasons, which are all included in this paper [1].

- A practical and theoretical proof of the existence of backlog in the waiting area is needed to support the necessity of solving the multiple-choice call-in issue in outpatient clinics.
- In the conference version, the myopic policy (MP) is constricted to at most two patient classes [1]. In this paper, the MP is extended to the scenario with multiple patient classes (three or more) by introducing a standard patient class, which provides a benchmark, and proposing a more general myopic optimal selection policy (MOSP).
- The proposed MOSP is highly related to some predetermined parameters. Thus, a sensitivity analysis is conducted on MOSP, and managerial insights are determined.

The rest of this paper is organized as follows. The literature review is provided in Section II. In Section III, a finite-horizon MDP model is established. An MP is proposed in Section IV to describe the dynamic patient scheduling within a clinic. In Section V, the sensitivity analysis on each parameter is presented. An empirical study, which is conducted to observe the performance of each policy, is indicated in Section VI. The simulation results confirm the applicability of the proposed model to a more general case.

## II. Literature Review

Most of the previous studies focus on the static healthcare appointment and admission system [2]–[10], where all decisions, including the determination of the appointment sequence, are made before the service session begins. However, works under dynamic settings, where the scheduler needs to make decisions on the basis of current state and the anticipation to the ulterior system performances, have been published recently. A known research investigated a dynamic system in healthcare facility [11]. Some previous studies focus on the dynamic priority rules for admitting patient into service [12], [13], whereas other works attempt to solve a dynamic resource allocation or demand assignment problem [14]–[17]. This paper focuses on the dynamic real-time admission policy based on the state of the sequential healthcare service system. Though the sequence has been determined in advance, we still review works on sequential appointment or admission system as follows.

averse in the domain of time, the negative exponential utility function should be used. Perceived waiting time negatively affects customer service satisfaction [26].

This paper develops the dynamic admission control (AC) policy in a sequential healthcare service system with tardy patients. The most significant distinction is that the patient dissatisfaction, which is measured by utility theory model, is proposed as the objective function.

## III. Model Description

The AC system in an outpatient clinic was formulated by using a finite-horizon CTMDP model. In this section, the decision epochs, state space, action space, state transition, and the utility-based cost function are provided. For simplification, the system only includes one physician without any interruption before all patients were served. Patients with different sequence indexes arrive according to heterogeneous distributions, but the service time is homogeneous. The allocation of capacity is not included in this paper. In other words, the numbers of advance appointment patients and walk-ins are acknowledged at the beginning of office hours. All the bold notations in the equations and expressions represent the vector form of the variables.

### A. Scheduling Horizon and Decision Epochs

The service system within a day was considered as the start of patient arrival anterior to the start of physician service. The system releases a total of $n$ appointments in a day, and the length of scheduling horizon $T \in \Re^*$ depends on the end time of a physician's work instead of being constant. Registrations in the scheduling horizon were allocated to advance appointment patients and walk-ins in advance such that $n_a$ out of $n$ are for advance appointment patients. The remaining $n-n_a$ are for walk-in patients. Appointment access is terminated before the walk-in registration window opens. Thus, the sequence indexes of advance appointment patients must be smaller than that of walk-in patients. Each patient obtains a sequence index $i \in \{1,2,\ldots,n\}$ once the patient finishes the registration process and is nominated as "patient $i$" for convenience. The actual scheduled numbers of registrations for two types of patients defined as $n_a$ and $n_w$ are calculated by $n_a = \min\{d_a, n_a\}$, $n_w = \min\{d_w, n - n_a\}$

$$\boldsymbol{S}^m = (t^m, \boldsymbol{AT}^m, \boldsymbol{b}^m) \mid 0 \le t^m < T;\ AT_i^m \le t^m \text{ or } AT_{im} = M; b_{im} = 0, 1;\ 1 \le i \le n\}.$$

The starting point of arrival is anterior to that of physician service. Thus, the definition for the original space is the system state when the entire scheduling horizon begins, which can be expressed as

$$\boldsymbol{s}^0 = (0, \boldsymbol{M}, \boldsymbol{0}) \text{ where } \boldsymbol{M} \in {}^n \text{ with all}$$

elements equivalent to $M$, and $\boldsymbol{0} \in \Re^n$ with all elements equivalent to 0.

### *C*. Action Space

The scheduler's task in this model is to decide the patient to call for the next service. The action space at the $m$th decision epoch is

$$A^m = a^m = i \mid AT_i^m = M, b_i^m = 0$$

where $a^m$ is the sequence index of the patient called-in at the $m$th decision epoch. This expression illustrates that alternatives are available for those who have arrived ($AT_i^m = M$) but have not been serviced ($b_i^m = 0$). Patient classes are determined by sequence indexes. The waiting time of the arbitrary patient $i$ in the waiting pool whose $b_i^m = 0$ at the $m$th decision epoch defined by $WT_i^m$ is calculated as

$$WT_{im} = t^m - AT_i^m, \quad AT_i^m = M \text{ and } b_i^m = 0 \quad 0, \text{ otherwise.}$$

When decision $a^m = i$ is made, the "postdecision state" is expressed as follows:

$$\boldsymbol{s}^{m+} = (t^{m+}, \boldsymbol{AT}^{m+}, \boldsymbol{b}^{m+}) = t^m, AT_i^m, \boldsymbol{b}^m + \boldsymbol{e}_i$$

where $\boldsymbol{e}_i \in {}^n$ that only the $i$th element is one, whereas that of others are zero.

### *D. State Transition*

The transition from the last postdecision state to the next predecision state is investigated in this section. The stochastic elements in the state transition include the time of decision epoch $t^{m+1}$ and the arrival time vector $\boldsymbol{AT}^{m+1}$. The newly arrived patients during $t^m$ to $t^{m+1}$ should be added to the waiting pool, and their waiting time should be updated. The state transition from $\boldsymbol{s}^{m+}$ to $\boldsymbol{s}^{m+1}$ is shown as follows:

$$t_m, AT_{im}, \boldsymbol{b}_m + \boldsymbol{e}_i \rightarrow t_{m+1}, AT_{im+1}, \boldsymbol{b}_{m+1}$$

where $\boldsymbol{b}^{m+1} = \boldsymbol{b}^m + \boldsymbol{e}_i$.

The transition probability of epoch time from $t^m$ to $t^{m+1}$ is equivalent to the probability that the sum of the $m$th service time and the $m$th physician's idle time is equal to $t^{m+1} - t^m$. $\tau_0$ is defined as the start time of service process, $\tau_i, 1 \le i \le n$, as the service time for patient $i$, and $\tau_j, 1 \le j \le n$, as the physician's idle time before the $j$th service. Thus, the time for the $m$th decision where the patient in the $m$th service is sequenced as $i$. In the real-world system, the service process starts after the arrival process starts, and the demand of walk-in patients is high. These two characteristics of the outpatient clinic lead to the existence of backlogs in the waiting area. Lemma 1 theoretically provides the upper bound of the probability that no backlog exists at the $m$th decision epoch $t^m$, given the arrival rate $\lambda$, the service rate $\mu$, and the start time of service process $\tau_0$. Let $X_t$ denote the number of patients in the system at time $t$, and Fig. 3 shows the upper bound of the nonbacklog probability based on realistic data.

*Lemma 1:* Given the traffic density $\rho = \lambda/\mu$, the upper bound of the probability that no backlog exists at the $m$th decision epoch $t^m$ is

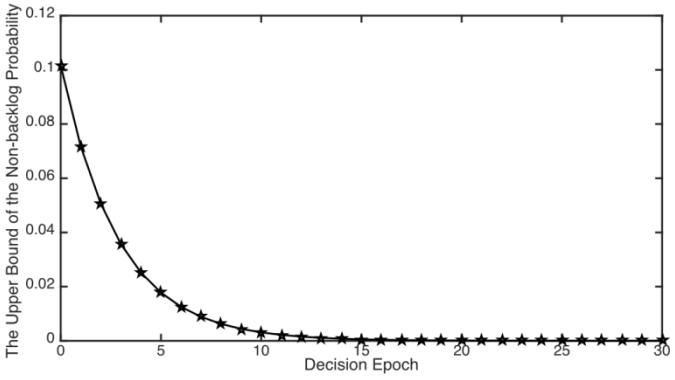
$$\Pr\{X_{tm} = 0 \mid X_0 = 0\} = (e^{-\lambda\tau_0} + \lambda\tau_0 e^{-\lambda\tau_0})[(1+\rho)e^{-\rho}]^{m-1} \quad (1)$$

which can be approximated to zero when $\lambda\tau_0 \ge 3$ and $\rho \ge 1$.

Thus, the transition probability at decision epoch $m$ can be approximated as

$$\Pr\{t_{m+1} = t_m + t \mid t_m\} \approx \Pr\{\tau_i = t \mid t_m, \tau_i > 0\} = p_{st}(t). \quad (2)$$

*Proof:* See Appendix A.




Lemma 1, which reveals that the upper bound of probability that $X_{t^m} = 0$ is small enough, also enables us to make the following assumptions for model simplification.

- The backlog is always nonempty before all the patients arrive at the clinic, so that the physician's idle time $\tau_m$ before an arbitrary decision epoch $m$ is zero.
- Time intervals between two adjacent decision epochs can be approximated to the service time that is homogeneously distributed.
- Patient backlog always occurs during the scheduling horizon, and the selection policy is necessary when multiple alternatives are in the waiting list.

### E. Objective Function

One of the highlights of this paper is that we measure the waiting costs of patients on the basis of the utility theory instead of simply expressed by linear function. Patients in different classes have heterogeneous disutility function to convey their dissatisfaction of waiting in the queue, and the scheduler should work to minimize the total dissatisfaction throughout the scheduling horizon. We assume that the current state is $\boldsymbol{s}^m = (t^m, AT_i^m, b_i^m)$, so that the waiting time for patient $i$ in the waiting area is

$$_m t_j - AT_{im}, \text{ being served at period } j, j \le m \; WT_i = t_m - AT_{im+},$$

otherwise

$$\mathbf{E}V\boldsymbol{s}^{n}{}_{+}{}^{1},a^{n}{}_{+}{}^{1}= V\boldsymbol{s}^{n}{}^{+1}, a^{n+1}) = f_i WT_i^{n}{}_{+}{}^{1} \quad (6)$$

where $a_*{}^{m}{}_{+}{}^{1} \in A\boldsymbol{s}^{mm}{}_{++}{}^{11}$, and is the optimal action at the $f_i(\cdot)$ is the waiting cost functions of $m$+1th period

given the state
patient $i$, which is convex and increasing under the waiting aversion assumption. Let $T_0$ denote the patient's waiting time expectation [25]. Taking account of the different sensitivities of walk-in $n_a + 1 \le i \le n$ and advance appointment patient $0 \le j \le n_a$, the relations between $f_i(t)$ and $f_j(t)$ with the same waiting time $t$ should be

$$f_j(t) < f_i(t),\ 0 \le t < T_0\ f_j(t) \ge f_i(t),\ t \ge T_0$$

since advance appointment patients must be more sensitive to the waiting time [25]. The objective function explicitly balances the effects of patient classes and waiting time on the dynamic scheduling. Complete fair policies, such as first-come first-serve (FCFS), attenuate the satisfaction of more sensitive patients, while complete class-distinguished policies make some walk-in patients extremely unhappy. In Section IV, we develop a dynamic scheduling policy based on Bellman's equation, which finds a tradeoff between distinction and fairness.

## IV. Model Integration of Heterogeneous-Arrival and Homogeneous-Exponential Service System With Negative Exponential Utility Function

### A. Objective Function Derivation

According to [25], the negative exponential utility function is suitable to express the disutility of customers' experience in a wait-averse queue. The expected utility for patient $i$ to wait in a queue is

$$\mathbf{E}[u_i(t)] = -A_1(T_0)^{\gamma} \exp(-c_i(T_0 - t))$$

where $A_1$, $\gamma$, and $c_i$ are positive constants. $A_1$ can be interpreted as a measure of a customer's value of time. $\gamma$ captures the direct impact of the initial expectation of the total waiting time $T_0$, and $c_i$ is a measure of risk aversion for patient $i$. The cost function is equal to the opposite value of the utility as $f_i(t) = -\mathbf{E}[u_i(t)] = A_1(T_0)^{\gamma} \exp(-c_i(T_0 - t))$. (7)

The first- and second-order derivatives of expression (7) can be calculated as $f_i(t) = c_i A_1(T_0)^{\gamma} \exp(-c_i T_0)\exp(c_i t) > 0$

$$f_i(t) = {}^{c}{}_{i}{}^{2} A_1(T_0)^{\gamma} \exp(-c_i T_0)\exp(c_i t) > 0.$$

Therefore, the expression (7) matches the requirement to the objective function with increasing monotonicity and convexity. $c_i$ is the crucial parameter in such a model reflecting a patient's tolerance to long waits ($t > T_0$) such that the larger $c_i$ is, the lower tolerance the patient has, so that the cost incurred by waiting is higher. The value of $c_i$ can be estimated by the demographic statistics, i.e., the age and the gender, combined with the health status of the patient. However, in this paper, we categorize patients based on registration. Coherent to Section III-E, the sensitivity parameter of advance appointment patients must be larger, and thus

$$c_j > c_i, \qquad 1 \le j \le n_a, \qquad n_a + 1 \le i \le n.$$

For notation convenience, let $\alpha$ denote the constant coefficient of $A_1(T_0)^{\gamma}$. Thus, the entire objective function originally formed as expression (5) is rewritten as when $1 \le m \le$

$$\min_{m \in A^m} \mathbf{E}[V \qquad (\boldsymbol{s}^m, a^m)]$$

$$= \qquad \Pr\{\boldsymbol{s}_{m+1} \mid \boldsymbol{s}_m\}$$

$$\times \; - \qquad -c_i\left(T_0 - WT_i^{m+1} \qquad -c_i\left(T_0 - \qquad e^{WT_{im}\alpha}\right.\right.$$

$$e$$

$$i=1$$

$$+ \mathbf{E}V\boldsymbol{s}^{m+1}, a_*^{m+1} \quad (8)$$

where the final overall cost is

$$n \min \mathbf{E}[V(\boldsymbol{s}_{n+1}, a_{n+1})] = \alpha\, e^{-caT_0 - WT_{in+1}}. \quad (9)$$

$$i=0$$

### B. Myopic Optimal Selection Policy

Since the state space of the MDP model is continuous and uncountable, the closed-form global optimal policy is intractable. We figure out the optimal action minimizing the expectation augmentation of the total cost function at the next decision epoch instead, which is called the MOSP. Assume the state space at the $m$th decision epoch is $\boldsymbol{s}^m = (t^m, \boldsymbol{AT}^m, \boldsymbol{b}^m)$, where there are more than one patients $i \in \{1,2,...,n\}$ that $AT_i^m = M$ (patient $i$ has arrived) and $b_i^m = 0$ (but has not been served). New notations for the calculation in this section include the following.

- $c_i$, the risk-aversion index for patient $i$.
- $\mathbf{E}[\triangle f_i(t \mid WT_i^m)]$, the expected cost augmentation if the waiting time expansion is $t$ and the current waiting time is $WT_i^m$ with the expression $\mathbf{E}[\, f_i(WT_i^m + t) - f_i(WT_i^m)]$.

Thus, the objective function of the one-step-forward myopic optimization problem takes the following form as follows when $1 \le m \le n$:

$$_m\min_m \mathbf{E}[V(\boldsymbol{s}^m, a^m)] - \mathbf{E}V\boldsymbol{s}^{m+1}, a_*^{m+1})]\} =$$

$$a \in A \Pr\{\boldsymbol{s}m+1 \mid \boldsymbol{s}m\}\, \boldsymbol{s}_{m+1} \in \boldsymbol{S}^{m+1} \cdot {}_n{}^{m+1} - WT_i^m \mid WT_i^m.$$

$$\mathbf{E}f_i WT_i$$

$$i=1 \quad (10)$$

The *priori* knowledge of patients in the backlog contains two parts: 1) the patient class and 2) the current waiting time. The MOSP is intended to find a tradeoff between the two parts above, by which the patient with maximized $\mathbf{E}[\, f_i(t \mid WT_i^m)]$ should be selected. Therefore, the formulation of the MOSP is, for $1 \le m \le n$ max $\mathbf{E}f_i WT_i^{m+1} - WT_i^m \mid WT_i^m$

$$1 \text{s.t.} \le i \le n\, AT_i m < M$$

$$b_i^m = 0. \quad (11)$$

According to the model assumption, the interepoch time is identically and independently distributed in accordance with the density function $p_{st}(t)$ that is exponential. By analyzing the cost augmentation function, the structure of the MOSP for multiple alternatives within two classes is shown as Theorem 1.

The crucial metric evaluating the priority of patients in the MOSP is the revised waiting time $\mathrm{RWT}^m{}_i$, which practically represents the standard waiting time of patient $i$ equivalent to that of another patient with risk-aversion index $c_0$. Specifically, when $c_i = c_0$, the slope $\vartheta_i = 1$, and the intercept $\beta_i = 0$, which alludes to the revised waiting time $\mathrm{RWT}^m{}_i$ equal to its original waiting time $WT_i^m$. After the transformation using (14), Corollary 1 releases an MOSP on the basis of the revised waiting time, and Theorem 1 is a special case with two types of patients.

## V. Sensitivity Analysis on MOSP

The scheduler's decision to call in a proper patient next is directly influenced by switching-curve parameters $\vartheta$ and $\beta$. Furthermore, we discuss the influence of basic parameters $T_0$, $\mu$, $c_i$, and $c_j$ on slope $\vartheta$ and intercept $\beta$. We assume $c_i < c_j$, and let $c$ denote a constant with the expression of the parameter $\vartheta$ represents the *equivalent marginal increasing rate* of two types of patients, which means that one unit waiting time augmentation for patient $i$ is equivalent to $\vartheta$ units augmentation for patient $j$, whereas the parameter $\beta$ represents the *basic equivalent-priority waiting time* of patient $i$, which means that the priority of the newly arrived patient $i$ is equal to that of patient $j$'s who has waited for $\beta$ units of time. Propositions 1–5 explicate how parameters in cost functions affect the switching curve, and additionally, influence the decision of the scheduler. Basic relationships between $T_0$, $\mu$, $c_i$, $c_j$, and $\beta$ are exhibited in Figs. 5–7 as follows. Proofs of Propositions 4 and 5 are in the appendixes.

*Proposition 1:* For any $c_i < c_j < \mu$, $\vartheta$ is proportional to $c_i$, but in inverse proportion to $c_j$.

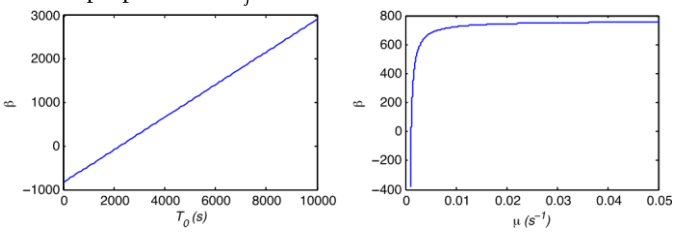



Fig. 5. Curve of parameter $\beta$ within the feasible domain of $T0$ and $\mu$.

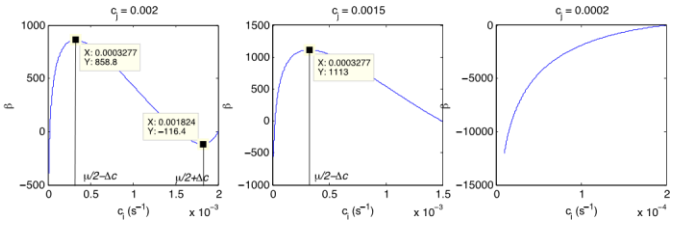
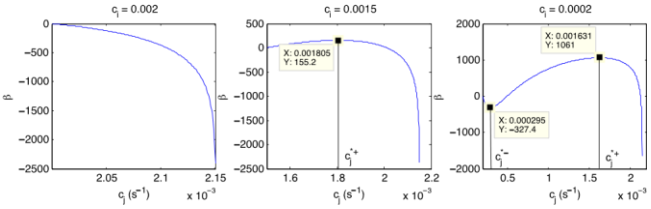


Fig. 6. Curve of parameter $\beta$ within the feasible domain of $c_i$ under different cases.



1) If $c_i < \mu/2 - c$, in the interval $(c_i, \mu/2)$, one and only one root of the equation $\phi(c_j) = 0$ exists, denoted by $c^{*-}_j$. Similarly, in the interval $(\mu/2, \mu)$, one and only one root of the equation $\phi(c_j) = 0$ exists, denoted by $c^{*+}_j$. Thus, there are three monotone intervals of $c_j$: $\beta$ decreases monotonically as $c_j$ increases in the interval $(c_i, c^{*-}_j)$ and $[c^{*+}_j, \mu)$, and increases in $[c^{*-}_j, c^{*+}_j)$.
2) Else, if $\mu/2 - \Delta c \le c_i < \mu/2 + c$, in the interval $(c_i, \mu)$, one and only one root of the equation $\phi(c_j) = 0$ exists, denoted by $c^{*-}_j$. Thus, there are two monotone intervals of $c_j$: $\beta$ increases monotonically as $c_j$ increases in the interval $(c_i, c^{*+}_j)$, and decreases in $[c^{*+}_j, \mu)$.
3) Otherwise, there is only one monotone interval of $c_j$: $\beta$ decreases monotonically as $c_j$ increases in the interval $(c_i, \mu)$.

*Proof:* See Appendix B-B.

Through further analysis, we conclude the following insights when we assume that the patient $i$ is a walk-in and patient $j$ is an advance appointment patient where $c_i < c_j$.

1) According to Proposition 1, the *equivalent marginal increasing rate* is directly determined by the ratio of different types of patients' risk-aversion indexes. The managerial insight is that the more averse a patient to the waiting, the more waiting time the other type of patient should have to obtain the same priority with him.
2) Proposition 2 indicates that the longer expected waiting time leads to a lower *basic equivalent-priority waiting time* for just-arrived walk-in patients. The managerial insight is that the increase of the overall expected waiting time extenuates the basic priority of walk-ins.
3) Proposition 3 indicates that the higher service rate leads to a higher *basic equivalent-priority waiting time* for just-arrived walk-ins. The managerial insight is that the increase of the service speed promotes the basic priority of walk-in, which implies the opposite effect that the waiting time expectation and the service rate have because of the negative coefficient between them.
4) Propositions 4 and 5 show that patients' risk-aversion indexes do not have monotone property to the parameter $\beta$. However, they expound the extreme points of $c_w$ and $c_a$ in different scenarios. Exceptionally, if both $c_w$ and $c_a$ are relatively small, that is $c_w, c_a \mu$, they have opposite monotone to the *basic equivalentpriority waiting time* of walk-ins. This finding implies that the higher $c_w$ is, the higher basic priority the walk-in possesses, while $c_a$ has the negative effect on it.

## VI. Empirical Study: The Pediatric Department in the PUTH

The pediatric department is one of the most well-known departments at PUTH. Patients from Beijing need high-quality treatment, which is scarce, leading to a crowded service environment. According to field research, the waiting area is uncomfortable and full of screaming of sick children. The current call-in policy called SIBP is completely based on the sequence of the registration. In other words, the nurse calls the patient in the waiting area with the smallest sequence index for the next service. To encourage patients to make the appointment in advance, the number of advance appointment

TABLE I Input Parameters of the Numerical Example

| Sequence Index | Arrival Time (s) | Service Time (s) |
|---|---|---|
| 1 | 4100 | 1000 |
| 2 | 2900 | 1000 |
| 3 | 1000 | 1000 |
| 4 | 3900 | 1000 |
| 5 | 1900 | 1000 |

patients is smaller than the number of walk-in patients, which in the real-world case, numbers 1–12 are allocated to advance appointment patients and numbers 13–30 are reserved for walk-ins. In fact, walk-ins usually arrive at the registration window on the first floor of the PUTH early to compete for the registration, and then, directly go to the pediatric department on the second floor. Since the opening time of registration is half an hour ahead of that of the pediatric department, the walk-ins have to wait until the physicians start to work. In a worse case, the patients with smaller sequence index have higher priority in the waiting area. Thus, advance appointment patients with small sequence index always jump the queue. The main problem of SIBP is the poor fairness among patients, wherein some patients with large sequential index who arrive at the system early may receive the service quite late.

In this paper, we compare frequently used scheduling policies to our proposed MOSP.

- *Waiting-Time-Based Policy:* The priority is completely based on the waiting time that patients have spent in the queue, which means that the patient with the longest waiting time is selected next. This policy is identical to the FCFS policy.

- *SIBP:* The policy of current system that has been defined.
- *Time Benchmark Policy:* The scheduler sets up a time benchmark for the waiting patients such that they are classified into two groups: underbenchmark patients and overbenchmark patients. Overbenchmark patients have a strongly higher priority than others. Patients in a group are ordered by the SIBP.
- *MOSP:* The scheduler chooses patients in accordance with the discriminant (13).

### A. Simple Example

We consider a problem with five patients $n = 5$, where two are appointment patients $n_a = 2$ to show four policies more clearly. We assume that the horizon begins at time 0, and that the service process begins at time $t_0 = 2000$. The time benchmark for time benchmark policy (TBP) is 2000. The easily computed input parameters, including the arrival time and service time, are demonstrated in Table I. We set parameters in the objective function as follows: $A_1 = 1$, $\gamma = 1$, $T_0 = 1000$, and $c_i = 0.0008$ for $i = 1,2$, and $c_i = 0.0005$ for $i = 3, 4, 5$, and thus, we acquire the call-in decision at each decision epoch and the service start time of each patient under four policies in Table II.

TABLE II

SERVICE START TIME OF EACH PATIENT

| Sequence Index | WTBP | SIBP | TBP | MOSP |
|---|---|---|---|---|
| 1 | 6000 | 5000 | 5000 | 5000 |
| 2 | 4000 | 3000 | 3000 | 4000 |
| 3 | 2000 | 2000 | 2000 | 2000 |
| 4 | 5000 | 4000 | 6000 | 6000 |
| 5 | 3000 | 6000 | 4000 | 3000 |

TABLE III

EXPERIMENTAL RESULTS OF FOUR POLICIES

| Sequence Index | WTBP | SIBP | TBP | MOSP |
|---|---|---|---|---|
| WT of 1 | 1900 | 900 | 900 | 900 |
| WT of 2 | 1100 | 100 | 100 | 1100 |
| WT of 3 | 1000 | 1000 | 1000 | 1000 |
| WT of 4 | 1100 | 100 | 2100 | 2100 |
| WT of 5 | 1100 | 4100 | 2100 | 1100 |
| Avg. WT | 1240 | 1240 | 1240 | 1240 |
| Var. of WT | 138,000 | 2,738,000 | 738,000 | 238,000 |
| Total disutility | 6240 | 7759 | 6576 | 5791 |

The actual service sequence for each policy is different. For waiting time-based policy (WTBP), it is 3, 5, 2, 4, 1; for SIBP, it is 3, 2, 4, 1, 5; for TBP, it is 3, 2, 5, 1, 4; and for MOSP, it is 3, 5, 2, 1, 4. We demonstrate the results, including the waiting time of each patient, the average waiting time, the variance of waiting time, and the total disutility in Table III.

Table III illustrates that the MOSP has the best performance (lowest disutility), followed by WTBP, and then the TBP, and the SIBP performs worst. Given that the average waiting time for each policy is identical, one of the reasons for the difference of objective function is the variance of waiting time. The variances of WTBP and MOSP are much lower than those of SIBP and TBP, resulting in the improved performance of these two policies. When comparing the WTBP and MOSP, we find that the only difference is the decision at the fourth service ($t = 5000$ s). The physician should choose one from patient 1, whose waiting time is $WT_1^4 = 900$ s, and patient 4, whose waiting time is $WT_4^4 = 1100$ s. Although the current waiting time for patient 4 is larger, patient 1 is the appointment patient more sensitive to long waiting. According to expression (13) in Proposition 1, the MOSP chooses patient 1, and the experimental results show that it is the right choice.

The TBP performs poorly as well because of the unfairness even though it is not as severe as the SIBP. Both the WTBP and the MOSP provide high fairness to patients in the same class. The WTBP provides complete fairness to all of the patients. Similar to the first scenario, patients under the WTBP all wait for around an hour, which is roughly equal to the expected time $T_0$. For the MOSP, most of the advance appointment patients wait for less than 50 min, but the walkins waiting time ranges from an hour to one-and-a-half hours. The fluctuation of curves in Fig. 9 is mainly due to the heterogeneous arrival time of patients with different sequence indexes. We use the rough mean-variance model to handle the arrival data, and intrinsically, these random arrival times are regarded as a Gaussian distribution with irrelevant means and variances. Given that neither the arrival time nor the interarrival time is homogeneous but the service time is, the waiting time of each patient cannot eventually converge to a certain number or distribution. Therefore, curves in Fig. 9 cannot be smooth.

## VII. CONCLUSION AND FUTURE RESEARCH

The relationship between patients and physicians has become vulnerable in China recently and has caught public attention. Apparently, the uncomfortable service environment in the healthcare system is a major cause of patient exasperation. Among the enormous factors that affect patient's service experience, the waiting time is the most direct factor according to field research, as anxiety accumulates over time. This paper studied the issue from the patient's perspective to minimize the total dissatisfaction produced by waiting at the outpatient clinic. We studied the following special properties: 1) scarce medical resources; 2) impatient patients; and 3) unpredictable arrivals. The waiting area is always full of patients and the waiting time must be long because of property 1). In addition, patients are very eager to obtain scarce resources; thus, they are willing to wait for diagnosis, which indicates low no-show and cancelation rates and rare service abandonment. Property 2) implies patients' displeasure with the experience of excessive waiting. Patients lack patience and panic more easily. In this case, using a risk-averse model to formulate patients' utility is rational and indispensable. As for property 3), diseases are usually unpredictable such that the fraction of reserved

registrations cannot be high. According to our interviews with patients, plenty of them obtain the sameday demand for treatment of acute diseases. Consequently, the arrival of patients is difficult to predict; thus, a complete advanced appointment system is impractical. These conditions make reservations for walk-in patients essential.

This paper formulated the dynamic AC in outpatient clinics by the MDP method. We transferred the continuous-time model into a tractable discrete time one by introducing a timecounting dimension in the state space. An applicable MP that minimizes the overall disutility at the next decision epoch was proposed along with a switching curve that identifies the expected performance of possible alternatives. We also conducted a sensitivity analysis on parameters that influence the intercept and the slope of the switching curve. Collaborating with the PUTH, we conducted an empirical study based on the pediatric department at this hospital. Although the main part of model formulation in this paper was based on the Markovian conditions of Poisson arrival and exponential service time, the simulation results confirm the rationality of the prerequisite, which exhibited a consistent trend among all four policies.

The call-in policy MOSP can be implemented in clinics wherever the information system has been constructed. An automatic admission system based on the MOSP can be developed to replace the current SIBP and improve selection and calling-in tasks. Once a patient arrives and checks-in, he/she waits in the waiting area until the system automatically calls him/her. The main challenge of the automatic call-in system is determining the sensitivity parameter $c_i$ of each class of patients, which is highly related to the objective function value and the MOSP. Some experimental research revealing the actual feeling (utility) of patients in the waiting area should be proposed based on scene imitation in the future to handle this issue.

Potential research directions are as follows. First, we can extend our work from outpatient clinics to other service facilities with similar but not identical environments. The homogenous service duration assumption can be relaxed with different types of patients in the system. Second, the disutility used to measure waiting can also be extended to other metrics, e.g., the morality of patients with more severe diseases. Finally, several approximate dynamic programming methods, e.g., reinforcement learning, can be used to evaluate the scaling parameters to change the myopic optimal policy to a global optimal one.

## APPENDIX A
## PROOF OF LEMMA 1

*Proof:* The historical data explicate that the total arrival rate $\lambda$, including both advance appointment patients and walkins, is higher than the service rate $\mu$, which means that the traffic density $\rho = \lambda/\mu > 1$, and further indicates that the backlog accumulates over time. We use the $M/M/1$ queue as an approximation to verify the existence of the backlog.

Let $X_t$ denote the number of backlogs at time $t$. Then, the probability that $X_{\tau_0}$ at time $\tau_0$ equals to 0 with the prior knowledge of an empty original queue is calculated as follows: $\Pr\{X_{\tau_0} = 0 \mid X_0 = 0\} = \exp(-\lambda\tau_0)$.

We can easily find that when $\tau_0 > 3/\lambda$, $\Pr\{X_{\tau_0} = 0 \mid X_0 = 0\} < 0.05$. We assume that the time of the $m$th decision epoch is $t^m$. Then, the upper bound of the probability that no backlog exists at this time is as follows:

$$\begin{aligned}
&\Pr\{X_{t^m} = 0 \mid X_0 = 0\} \\
&= \left(\Pr\{X_{\tau_0} = 0 \mid X_0 = 0\} \Pr\{X_{t^1} = 1 \mid X_{\tau_0} = 0\}\right. \\
&\quad \left. + \Pr\{X_{\tau_0} = 1 \mid X_0 = 0\}\right) \prod_{i=1}^{m-1} \Pr\{X_{ti+1} = 1 \mid X_{ti} = 0\} \\
&< \Pr\{X_{\tau 0} \le 1 \mid X_0 = 0\} \Pr\{X_{\tau i} \le 1 \mid X_{ti} = 0\} \\
&\quad \prod_{i=1} \\
&= (e^{-\lambda\tau_0} + \lambda\tau_0 e^{-\lambda\tau_0})[(1+\rho)e^{-\rho}]^{m-1}. \qquad (15)
\end{aligned}$$

## APPENDIX B
## PROOF OF THE SENSITIVITY ANALYSIS

1) If $c_j > c_{i,\min}$, both of the extreme points lie in the domain of $c_i \in (0, c_j)$; thus, three monotone intervals exist.
2) Else, if $c_{i,\max} < c_j \le c_{i,\min}$, only $c_{i,\max}$ lies in the domain of $c_i \in (0, c_j)$, thus two monotone intervals exist.
3) Else, if $c_j \le c_{i,\max}$, no extreme point lies in the domain of $c_i \in (0, c_j)$, such that $\phi$ increases throughout the domain of $c_i$.

$$\phi(\mu/2) = T_0 c_i - 2 - \ln \frac{\mu - \phi(c^i)}{c_i} - \mu$$

$$\lim_{c_j \to \mu} \phi(c_j) \to -\infty.$$

By analyzing the function $\phi$, we elucidate the monotone intervals of $c_j$ in the following three cases.

1) If $c_i < \mu/2 - c$, we obtain $\phi(c_i) < 0$ and $\phi(\mu/2) > 0$. Thus, in the interval $(c_i, \mu/2)$, one and only one root of the equation $\phi(c_j) = 0$ exists, denoted by $c^{*-}_j$. Similarly in the interval $(\mu/2, \mu)$, one and only one root of the equation $\phi(c_j) = 0$ exists, denoted by $c^{*+}_j$. Thus, three monotone intervals of $c_j$ exist: $\phi$ decreases monotonically as $c_j$ increases in the interval $(c_i, c^{*-}_j)$ and $[c^{*+}_j, \mu)$, and increases in $[c^{*-}_j, c^{*+}_j)$.
2) If $\mu/2 - c \le c_i < \mu/2 + c$, we get $\phi(c_i) > 0$ and $\phi(\mu/2) > 0$. Thus, in the interval $(c_i, \mu)$, one and only one root of the equation $\phi(c_j) = 0$ exists, denoted by $c^{*-}_j$. Thus, two monotone intervals of $c_j$ exist: $\phi$ increases monotonically as $c_j$ increases in the interval $(c_i, c^{*+}_j)$, while decreases in $[c^{*+}_j, \mu)$.
3) If $c_i \ge \mu/2 + c$, we obtain $\phi(c_i) < 0$ and $\phi(\mu/2)$ is irrational. Thus, in the interval $(c_i, \mu)$, $\phi(c_j) < 0$ is always satisfied. Hence, only one monotone interval of $c_j$ exists: $\phi$ decreases monotonically as $c_j$ increases in the interval $(c_i, \mu)$.

## REFERENCES

[1] Y. Qiu, J. Song, and Z. Liu, "Real time access control of patient service in the pediatrics department," in *Proc. IEEE Int. Conf. Autom. Sci. Eng.*, Aug. 2015, pp. 734–739.

[2] P. M. V. Bosch and D. C. Dietz, "Minimizing expected waiting in a medical appointment system," *IIE Trans.*, vol. 32, no. 9, pp. 841–848, 2000.

[3] L. W. Robinson and R. R. Chen, "Scheduling doctors' appointments: Optimal and empirically-based heuristic policies," *IIE Trans.*, vol. 35, no. 3, pp. 295–307, 2003.

[4] K. Muthuraman and M. Lawley, "A stochastic overbooking model for outpatient clinical scheduling with no-shows," *IIE Trans.*, vol. 40, no. 9, pp. 820–837, Jul. 2008.

[5] B. T. Denton, A. J. Miller, H. J. Balasubramanian, and T. R. Huschka, "Optimal allocation of surgery blocks to operating rooms under uncertainty," *Oper. Res.*, vol. 58, no. 4, pp. 802–816, 2010.

[6] L. W. Robinson and R. R. Chen, "A comparison of traditional and open-access policies for appointment scheduling," *Manuf. Service Oper. Manage.*, vol. 12, no. 2, pp. 330–346, 2010.

[7] J. Luo, V. G. Kulkarni, and S. Ziya, "Appointment scheduling under patient no-shows and service interruptions," *Manuf. Service Oper. Manage.*, vol. 14, no. 4, pp. 670–684, 2012.

[8] C. Zacharias and M. Armony, "Joint panel sizing and appointment scheduling in outpatient care," Working paper 5.6.1, 2013.

[9] H.-Y. Mak, Y. Rong, and J. Zhang, "Appointment scheduling with limited distributional information," *Manage. Sci.*, vol. 61, no. 2, pp. 316–334, 2014.

[10] M. A. Begen and M. Queyranne, "Appointment scheduling with discrete random durations," *Math. Oper. Res.*, vol. 36, no. 2, pp. 240–257, May 2011.

[11] L. V. Green, S. Savin, and B. Wang, "Managing patient service in a diagnostic medical facility," *Oper. Res.*, vol. 54, no. 1, pp. 11–25, Feb. 2006.

[12] N. Liu, S. Ziya, and V. G. Kulkarni, "Dynamic scheduling of outpatient appointments under patient no-shows and cancellations," *Manuf. Service Oper. Manage.*, vol. 12, no. 2, pp. 347–364, 2010.

[13] W. T. Huh, N. Liu, and V.-A. Truong, "Multiresource allocation scheduling in dynamic environments," *Manuf. Service Oper. Manage.*, vol. 15, no. 2, pp. 280–291, 2013.

[14] J. Patrick, M. L. Puterman, and M. Queyranne, "Dynamic multipriority patient scheduling for a diagnostic resource," *Oper. Res.*, vol. 56, no. 6, pp. 1507–1525, Dec. 2008.

[15] A. Sauré, J. Patrick, S. Tyldesley, and M. L. Puterman, "Dynamic multiappointment patient scheduling for radiation therapy," *Eur. J. Oper. Res.*, vol. 223, no. 2, pp. 573–584, Dec. 2012.

[16] S. A. Erdogan and B. Denton, "Dynamic appointment scheduling of a stochastic server with uncertain demand," *INFORMS J. Comput.*, vol. 25, no. 1, pp. 116–132, 2013.

[17] H. Balasubramanian, S. Biehl, L. Dai, and A. Muriel, "Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments," *Health Care Manage. Sci.*, vol. 17, no. 1, pp. 31–48, 2014.

[18] R. Hassin and S. Mendel, "Scheduling arrivals to queues: A singleserver model with no-shows," *Manage. Sci.*, vol. 54, no. 3, pp. 565–572, 2008.

[19] L. W. Robinson and R. R. Chen, "Estimating the implied value of the customer's waiting time," *Manuf. Service Oper. Manage.*, vol. 13, no. 1, pp. 53–57, 2011.

[20] L. R. LaGanga and S. R. Lawrence, "Appointment overbooking in health care clinics to improve patient service and clinic performance," *Prod. Oper.* Manage., vol. 21, no. 5, pp. 874–888, 2012.

[21] D. Ge, G. Wan, Z. Wang, and J. Zhang, "A note on appointment scheduling with piecewise linear cost functions," Math. Oper. Res., vol. 39, no. 4, pp. 1244–1251, 2013.

[22] H.-Y. Mak, Y. Rong, and J. Zhang, "Sequencing appointments for service systems using inventory approximations," Manuf. Service Oper.Manage., vol. 16, no. 2, pp. 251–262, 2014.

[23] P. Guo and P. Zipkin, "Analysis and comparison of queues with different indulevels of delay information," Manage. Sci., vol. 53, no. 6, pp. 962–970, Beijing, Jun. 2007.

[24] M. Baker and J. Wurgler, "Market timing and capital J. Finance, vol. 57, no. 1, pp. 1–32, 2002.sity, Beijing, in 2016.

[25] P. Kumar, M. U. Kalwani, and M. Dada, "The impact of waiting timethe guarantees on customers' waiting experiences," Marketing Sci., vol. no. 4, pp. 295–314, 1997.MO, USA.

[26] F. Biélen and N. Demoulin, "Waiting time influence on the satisfaction Manag. Service Quality, Int. J., vol. 17,simulation and optimization, stochastic and dynamic no. 2, pp. 174–193, Mar. 2007.programming, and their applications in scheduling, resource allocation, and health care fields.