Integrated Optimization of Appointment Allocation and Access Prioritization in Patient-Centred Outpatient Scheduling

Abstract

Walk-in patients can take a large portion of outpatient visits in comprehensive hospitals. In current common practice, an even number of appointments is assigned to each slot and patients with appointment are given a strictly higher priority than walk-in patients. As a result, walk-in patients usually wait longer and incur noticeable dissatisfaction. To improve their satisfaction, we propose an integrated optimization approach for patient-centred outpatient scheduling. This approach aims to allocate appointments in line with the temporal variation of patient walk-ins, and further develop a time-invariant threshold-based prioritization scheme to balance the service priorities between the two types of patients in real time. To substantiate this approach, we develop a prospect theory-based model to quantify each patient's real-time satisfaction. We apply discrete-event simulation-based optimization to maximize the average patient satisfaction over the two types of patients. To verify our integrated optimization approach, we conduct a pilot in the endocrine department of Shanghai Sixth People's Hospital. Our results suggest that our approach substantially decreases the waiting time of walk-in patients without much compromise on the satisfaction of patients who made advance appointment.

Keywords: application in healthcare system; resource allocation; real-time scheduling; customer behaviour

1. Introduction

Due to growing demand on outpatient appointment and severe shortage on medical personnel, outpatient scheduling has received great attention in many countries, including China. Large state-owned comprehensive hospitals (typically in large cities) accept walk-in patient visits for same-day consultation in addition to patients with consultation appointments made in advance. To many Chinese comprehensive hospitals, walk-in patients constitute as high as three quarters of the outpatient visitation volume (Jiang, 2010). There are two sources of walk-in patients. One is local residents. While advance appointment scheduling has been implemented for decades (Chinese Ministry of Health, 2009), the majority of local patients still prefer to visit hospital directly without an appointment. The other source of walk-in patients are non-local patients. Given significant geographic imbalance of medical providers, it is common for patients residing in suburban and rural areas to seek quality medical care in large cities. Many such patients have their first visit. Note that only those who have visited the hospital previously and have their personal information verified have the right to make appointment. Thus, non-local patients also become a source of walk-ins. The phenomenon of many walk-ins is not unique to China. For example, in a state-owned hospital in Turkey, walk-in visits take up more than 70% of total outpatient visits (Cayirli and Gunes, 2014). Hence, the resultant patient arrival process, while ensuring some level of equity among all patients, is much less manageable and presents a significant burden on the use of already deprived medical resource. It becomes important to incorporate the need from sizable walk-in visits and improve the satisfaction of the overall patient population. For exposition simplicity, we refer to outpatient visits of the former category as appointment patient (AP) visits and the latter as walk-in patient (WP) visits.

We next describe the common practice of outpatient scheduling as follows. An outpatient consultation session is divided into a number of slots with equal length (e.g. an hour); multiple AP visits are scheduled in each slot in advance; and WPs are seen only if all AP visits scheduled in that slot have been seen or there are no AP visits scheduled in that slot at the first place. As a result, a WP may have to wait long to be seen. Further, to ensure fairness, patients from the same category are scheduled with a first-come-first-service (FCFS) principle.

Now that WPs can account for a large proportion of outpatient visits, it is important to ensure their satisfaction without making much compromise on APs' satisfaction. For example, for a WP who has waited long, right before seen by a doctor, a batch of AP visits scheduled in the same slot are inserted before him. Thus, the WP may feel extremely unsatisfied. We find in Shanghai Sixth People's Hospital, our partnering hospital in China, arguments often take place at the beginning of each slot when WPs complain they have waited long but are not sure how much longer they need to wait. On the other hand, they witness lots of APs who have appointments at the slot just arrived and can be seen right away. Further, it is evident that the frustration becomes more noticeable when some WP waits longer than his/her expected waiting time. In this paper, our objective is to maximize average waiting-time-dependent satisfaction[1]. From the phenomenon above, we conclude that waiting-time-dependent satisfaction of a patient (a more meaningful objective) does not necessarily diminish linearly with respect to the increase in his/her waiting time, which is commonly modelled in the literature (e.g.,

[1] We can extend our objective to a more general notion of utility. In the outpatient scheduling literature, utility is often defined as the gain generated from service completion and the loss incurred by waiting. For same-day consultation scheduling, the gain is insensitive when the service is completed during the day. This is why in many papers, the objective to optimize is the waiting time. However, in the following, we argue maximization of waiting-time-dependent satisfaction is not equivalent to minimization of waiting time.

Robinson and Chen, 2003; Koeleman and Koole, 2012; Chen and Robinson, 2014). Hence, we use so-called prospect theory (Kahneman and Tversky, 1979) to characterize the waiting-time-dependent satisfaction. An S-shaped function is constructed, which captures reference point, loss aversion and nonlinear relationship between waiting time and satisfaction. For simplicity, we use utility to refer to the waiting-time-dependent patient satisfaction in the reminder of the paper.

To maximize the average utility, the existing literature has studied real-time access control strategies that adjust awaiting patient priorities to balance WPs and APs (e.g., Song et al., 2017). Meanwhile, studies on optimizing the number of appointments have appeared in the literature (e.g., Wang, Liu, and Wan, 2018) to improve WPs' satisfaction in congested slots and streamline the overall outpatient service process. In this paper, we propose a strategy to integrate the two aspects above. That is, the strategy is not only intended to control patient access at each single slot to balance AP and WP visits in real time, but also allocate different numbers of appointments at different slots to alleviate the impact of slot-wise variation of WP visits (clearly seen at our partnering hospital; see Fig. 1) on the access congestion and patient utility.

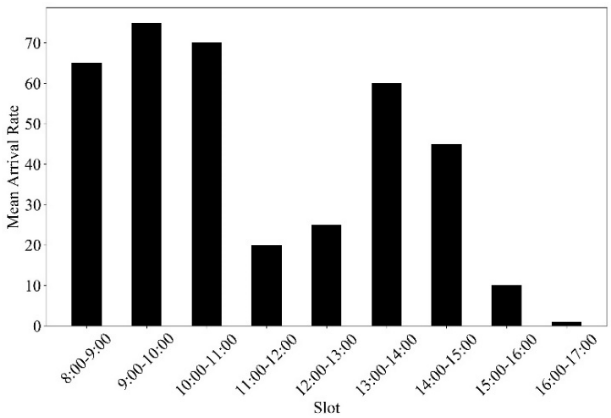

Fig. 1. An illustration of the temporal pattern of WP visits[2]

[2] Data comes from Shanghai Sixth People's Hospital. The figure reports the average hourly arrival rate of WP visits at the hospital's endocrine department on Mondays from 2013 - 2017.

More specifically, we determine the number of appointments to be assigned in each slot a priori. Taking an integrated approach, we further apply real-time access prioritization to balance AP and WP visits. Different from previous literature on making real-time access decisions based on patient's utility (e.g., Green, Savin, and Wang, 2006; Qu et al., 2015), we consider easing the difficulty of practical implementation through a time-invariant threshold-based prioritization scheme. Based on our prioritization scheme, whenever a patient has been seen and leaves the process, the system updates the status of the awaiting AP and WP at the beginning of their respective queue based on their utility at that moment. Based on respective thresholds (one determining whether the AP is satisfied or not; the other one determining whether the WP is satisfied or not), the two patients will be labelled as either unsatisfied AP or satisfied AP, and either unsatisfied WP or satisfied WP. Then the access decision following a time-invariant rule that assigns priorities to unsatisfied APs, then unsatisfied WPs, then satisfied APs and finally satisfied WPs. To efficiently solve the resultant integrated optimization problem on the number of appointments in each slot and time-invariant utility thresholds, we apply a simulation optimization approach that embeds the Optimal Computing Budget Allocation (OCBA) algorithm into a Particle Swarm Optimization (PSO) engine.

We conduct a proof-of-concept case study based on the real data collected at the endocrine department of Shanghai Sixth People's Hospital. In addition to numerically justifying the viability of our integrated optimization approach in the baseline case, our ensuring sensitivity analysis identifies the proportions of AP and WP arrival rates under which the proposed approach remains to be effective. We also find that our approach remains effective even if the waiting behaviour of patients varies noticeably. These results suggest our approach can be promoted in different real-world settings.

Further, we conduct a pilot study with an implementation based on the integrated optimization approach. Encouragingly, fewer patient complaints occur during the pilot study than before. In addition, we observe that the integrated approach decreases WPs' average waiting time and variance significantly, whereas APs' average waiting time almost stays the same and their variance is only slightly increased. With these promising results, the partnering hospital has considered promoting our integrated optimization approach to other departments.

This paper makes the following contributions. First, we construct a utility function based on the prospect theory to quantify patient's waiting-time-dependent satisfaction. In addition, we conduct a field experiment to validate the utility function and extract the model parameters. We justify that the prospect theory can provide a viable supplement to patient-centred outpatient scheduling. Second, we propose an integrated optimization approach that combines slot-based appointment allocation and real-time priority-based access control. The former makes resource commitment more balanced between the two types of patients and the later further helps schedule patients based on real-time utility assessment. Third, we pilot our scheduling strategy at a top-ranked Chinese hospital in Shanghai. The pilot confirms that our approach improves patients' average satisfaction effectively.

The remainder of this paper is organized as follows. We review relevant literature in Section 2. In Section 3, we overview the integrated optimization approach, specify the prioritization scheme, and formulate the optimization problem. In section 4, we provide details of our methodology, including the utility function modelling and simulation optimization algorithm design. In section 5, we describe our pilot study and report results. Finally, we draw conclusions and outline future work in Section 6.

2. Literature review

Outpatient scheduling has been a research area for a long time with numerous studies in the literature. Cayirli and Veral (2003), Gupta and Denton (2008), Ahmadi-Javid, Jalali, and Klassen (2017) have provided comprehensive reviews. This paper focuses on appointment scheduling and patients' access control.

There are two types of appointment scheduling problems: individual appointment scheduling and block appointment scheduling (Ho and Lau, 1992). With the former type, the scheduler divides each consultation session into small slots and schedules each patient at an individual slot (Denton and Gupta, 2003; Kong et al., 2013; Chen and Robinson, 2014). With the latter type, the scheduler divides a consultation session into several relative long slots and each slot allows multiple patients to make appointments (Yan, Tang, and Jiang, 2014; Song, Qiu, and Liu, 2017; Wang, Liu, and Wan, 2018). For outpatient scheduling of the latter type, almost all work studies a system that only involves APs. The common objective is to allocate appointments optimally to reduce the negative effect of APs' no-show or unpunctuality (LaGanga and Lawrence, 2012; Kong et al., 2013; Zacharias and Pinedo, 2014). Few studies have considered a system involving both APs and WPs. The objective is usually to decrease two types of patient's average waiting time. Some literature (Koeleman and Koole, 2012; Cayirli and Gunes, 2014) assumes that only one type of patients will be served in a single slot, and determines which particular slots open to WPs. Some other work considers the situation where two types of patients can be served in one slot (usually with a relatively large slot length) and determines the number of appointments for APs in each slot. For example, Yan, Tang, and Jiang (2014) studied a sequential appointment scheduling problem to determine the optimal capacity of APs and the optimal appointment time for each AP; Wang, Liu, and Wan (2018) optimized the total number of appointments for each day as well as allocation in each slot of a day considering WPs. Our paper studied the scenario where two types of patients can be served in a same slot, and except considering how to allocate appointments in each slot, we considered integrating a real-time access control to balance AP and WP visits.

In the outpatient access control literature, most studies focus on the real-time decision of admitting a sequence of patients of multiple types for services. For example, Gocgun et al. (2011) dynamically scheduled four types of patients with different degrees of urgency to maximize the total expected net revenue. Qu et al. (2015) proposed an admission policy depending on the number of remaining slots and the expected total APs to decide whether to admit a WP and when a WP should be seen. Geng and Xie (2016) considered two types of patients with different waiting time targets and different rewards by switching curves to maximize the expected total number of APs seen at a diagnostic facility. Like above studies, we also considered a system with two types of patients and that APs have higher priority than WPs. However, we directly investigated the use of a time-invariant threshold-based prioritization scheme, which is believed to be easier to implement in practice. This differs from the existing literature, much of which, as described earlier, formulated stochastic dynamic programming models with intention of deriving optimal admission policies based on the current system state, including waiting time and service utilization. Nevertheless, those stochastic dynamic programming models face serious computational challenges due to the large state space associated with the complex service system. So often, the authors sought heuristic policies with proven performance guarantees.

Another significant difference of our work with the previous literature is that we developed a new type of utility function for estimating patients' utility. Among literature considers the relationship of utility with waiting time, most of the work assume the utility changes linearly with time. For example, Gocgun et al. (2011), Koeleman and Koole (2012) and Qu et al. (2015) all treated waiting time cost per unit as a constant. Few other work study the nonlinear relationship of patient's utility and waiting time. Robinson and Chen (2003) considered summing several linear functions for patient waiting cost with respect to the average queue length and occupation rate. Ge et al. (2014) used piecewise linear functions to model the waiting cost with the assumption that the appointment time and processing time are both integers. Kemper, Klaassen, and Mandjes (2014) modeled the waiting cost with quadratic functions based on the Von Neumann-Morgenstern expected utility. Song et al. (2017) constructed an exponential disutility function to express the waiting cost. While the latest work has attempted to incorporate some realism in the modelling of patient utility, the stochastic dynamic programming models presented in the above papers and the desire of seeking implementable policies based on the models may have hindered the authors from taking a data-driven approach to develop accurate waiting-time-dependent utility (or cost) functions. For example, patients usually have an expected waiting time as a reference point, then their perception on the waiting is usually different before and after reaching the expectation. After waiting time exceed the expectation, the patients will typically be more impatient; implying the loss aversion property inherent to patients. We argue that the key phenomena reflecting the relationship between patient waiting and utility can be better captured by the so-called prospect theory (Kahneman and Tversky, 1979). Therefore, in our study, we elected to develop a prospect-theory-based utility function.

The prospect theory is an economics theory proposed by Kahneman and Tversky in 1979. The theory earned Daniel Kahneman the Nobel Memorial Prize in Economics in 2002 for it describing the utility perception behavior along with reference point, loss aversion and risk aversion. Several studies have applied the theory to estimate awaiting people's utility in other areas. For example, Avineri (2004) used it to describe how passengers perceived their utility while waiting for a bus. However, to the best of our knowledge, there is no study applying the prospect study in the area of outpatient waiting.

3. Research methodology framework

For this research project, we went through three phases: (1) field observation; (2) approach development; (3) pilot study. We summarize the project procedure in Fig. 2.

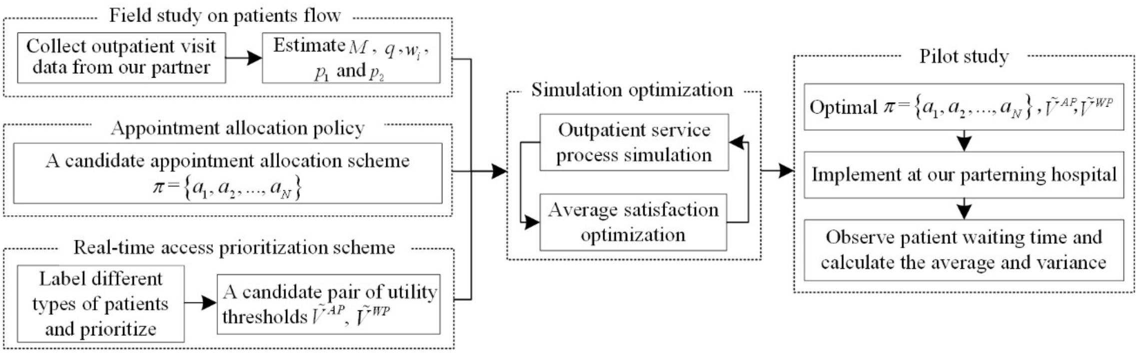

Fig. 2. An illustration of the project procedure

With a field study at our partnering hospital in Shanghai, we acquired information for modelling the patient flows of the hospital. Let M be the total number of patients (both APs and WPs) over the entire consultation session and q be the proportion of APs. The entire consultation session is divided into N slots, each of equal time length. Based on our observations, we assume that WPs' arrival for each slot follow a time-varying Poisson process with w_i being the mean arrival rate in the i^{th} slot. We also assume that the proportions of early arrival APs and late arrival APs are given, denoted by p_1 and p_2, respectively. We next assume that the waiting of early arrival APs (i.e., arriving before the start of the booked slot) only occurs at the beginning of each slot. On the other hand, late APs (i.e., arriving after the end of booked slot) are punished by being treated as WPs. As a result, a late AP enters the WP queue and takes the last spot of the queue.

Our integrated optimization approach consists of two aspects. First, we determine the number of appointments to each slot based on the average temporal pattern of WP arrivals over the consultant session. We use $\pi = \{a_1, a_2, \ldots, a_N\}$ to denote the number of appointments scheduled in each of the N slots. The intent is to complement temporal pattern of appointments made to anticipate temporal pattern of WP arrivals so that we get similar numbers of patient visits (both AP and WP) at different slots. As a result, such streamlining of the service process can lead to reduced waiting. Note that the daily arrival pattern may differ in different days of the week and different months of the year. We keep that in mind when deriving the WP daily patterns based on real-world data.

In the second aspect of our integrated optimization approach, we intend to improve the balance of AP and WP visits further, through making real-time control on their access by comparing the waiting-time-dependent utility, whenever a physician becomes available. To ensure successful implementation, we consider a time-invariant threshold-based prioritization scheme to identify the patient to be seen immediately next.

We describe the prioritization scheme in detail (see Fig. 3). Whenever a patient has been seen and leaves the process, we make real-time evaluation on the first AP and the first WP in the respective queue based on the prospect-theory-based utility function that characterizes either patient's utility (details in section 4.1).We first label the selected AP (WP) with two options by comparing the his/her utility with the threshold $\tilde{V}^{AP}$ ($\tilde{V}^{WP}$). An AP whose utility is below $\tilde{V}^{AP}$ is labelled as a UAP (an unsatisfied AP), otherwise labelled as an SAP (a satisfied AP). A WP whose utility is below $\tilde{V}^{WP}$ is labelled as a UWP (an unsatisfied WP), otherwise labelled as an SWP (a satisfied WP). We then prioritize the two patients based on the following agreement UAP > UWP > SAP > SWP, i.e., the AP always is always seen first unless s/he is labelled as an SAP and the corresponding WP is labelled as an UWP.

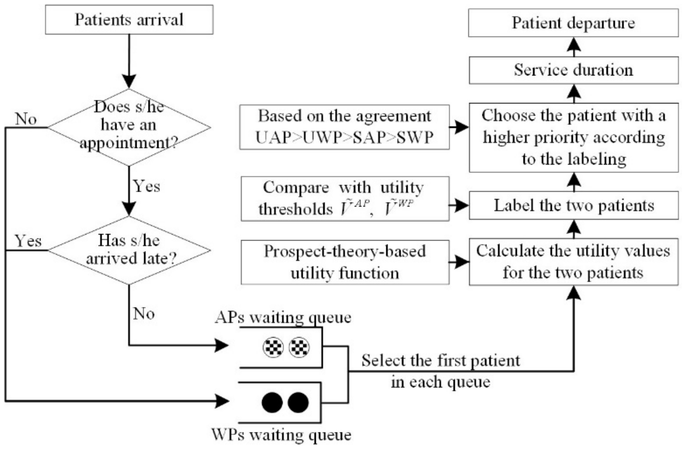

Fig. 3. Illustration of the outpatient scheduling process.

Based on the above descriptions, it is clear that $\pi=\{a_1, a_2, ..., a_N\}$ and $\tilde{\Gamma}^{AP}, \tilde{\Gamma}^{WP}$ all influence our objective, i.e., the average patient utility. As $\pi=\{a_1, a_2, ..., a_N\}$ affect the number of APs to be seen in each slot, $\tilde{\Gamma}^{AP}, \tilde{\Gamma}^{WP}$ should be adjusted accordingly to ensure the overall patient utility. We next formalize these implications as

$$\max_{\pi, \tilde{\Gamma}^{AP}, \tilde{\Gamma}^{WP}} k \times AV_{AP}(\pi, \tilde{\Gamma}^{AP}, \tilde{\Gamma}^{WP}) + AV_{WP}(\pi, \tilde{\Gamma}^{AP}, \tilde{\Gamma}^{WP}) \tag{1}$$

$$s.t. \quad \sum_{i=1}^{N} a_i = M \times q, \tag{2}$$

$$0 \le a_i \le A, \text{integer} \tag{3}$$

where k is a weighting coefficient of AP waiting-related utility as opposed to that for the WPs. We use AV_{AP} and AV_{WP} to denote the average utilities of APs and WPs, respectively. We use A to specify the maximum number of appointments that can be made in each slot. To solve the above problem, we resort to simulation optimization. Section 4.2 provides more information about the algorithm design.

Finally, in the third phase of the research project, we ran a pilot in our partnering hospital to assess the performance of the integrated optimization approach. The pilot results suggest slightly increased waiting time average and variability for APs in return for much reduced waiting time average and variability for WPs. The pilot demonstrates the viability of our approach.

4. Methodology specifics

We first describe the process of developing the utility function to model either patient's utility. Then we present the simulation optimization algorithm that jointly searches the optimal number of appointments in each slot and the pair of access control thresholds.

4.1. The prospect theory-based utility function

The utility function describes the relationship between patients' utility V and actual waiting time WT in real time. We apply prospect theory so-called S-shape value function, which captures reference point, loss aversion and nonlinear relationship between waiting time and satisfaction. Reference point means the utility is affected by the comparison with current state and reference point, e.g., the expectation. The utility function has a kink at the reference point. Loss aversion means people are more sensitive to loss than gain, where loss and gain are defined by the reference point. Nonlinear relationship between waiting time and satisfaction means along with the waiting time departure from the reference point, the utility increases or decreases in a nonlinear way. The utility function could be expressed as

$$V(WT) = V_{ET} + \mathrm{c}\left\{[(ET - WT, 0)^{+}]^{\alpha} - \lambda[(WT - ET, 0)^{+}]^{\beta}\right\} \tag{4}$$

where ET, V_{ET}, c, α, β and λ are positive constants. The patient's expected waiting time ET and its corresponding utility V_{ET} is the reference point. The gain region is where $0 \le WT < ET$ and $V > V_{ET}$, while loss region is where $WT \ge ET$ and $V \le V_{ET}$. c can be interpreted as the measurement of a patient's value of waiting time. α, β describe the nonlinear relationship between waiting time and satisfaction for gain and loss, respectively. λ describes the degree of loss aversion.

Assuming that the reference point of APs and WPs are equal, we could construct utility function of m th AP and n th WP as following to satisfy the assumptions.

$$V_{AP}(WT_m^{AP}) = V_{ET} + \mathrm{c}\left\{[\left(ET^{AP} - WT_m^{AP}, 0\right)^+]^{\alpha^{AP}} - \lambda^{AP}[\left(WT_m^{AP} - ET^{AP}, 0\right)^+]^{\beta^{AP}}\right\} \quad (5)$$

$$V_{WP}(WT_n^{WP}) = V_{ET} + \mathrm{c}\left\{[\left(ET^{WP} - WT_n^{WP}, 0\right)^+]^{\alpha^{WP}} - \lambda^{WP}[\left(WT_n^{WP} - ET^{WP}, 0\right)^+]^{\beta^{WP}}\right\} \quad (6)$$

where ET^{AP} and ET^{WP} denote the expected waiting time of APs and WPs which are predetermined by patients.

Then, we conduct an experiment to provide field evidence and evaluate fitness of the utility function. We first design a questionnaire and recruit a cohort of respondents. With the questionnaire, we collect individual basic information from each respondent to specify his/her patient type to be either AP or WP. Then we randomly assigned to the respondent two hypothetical scenarios of waiting in the outpatient clinic. In the first scenario, each respondent is assumed to wait several minutes longer than the expected waiting time. Inversely, in another scenario, each respondent is assumed to wait shorter than expected. The patients' expected waiting time is given. The specific over-time and reduce-time are randomly assigned to the respondent within a feasible range. Each respondent is asked to evaluate his/her wait time satisfaction with a score from 0 to 100. The details of the survey is presented in the appendix.

As a result, we collected 141 questionnaires after filtering 16 invalid samples, obtaining 67 and 74 questionnaires of APs and WPs, respectively. We obtain the relationship between patients' utility and waiting time by gathering each respondent's score under different scenarios. Then we perform a nonlinear regression, which indicates the R-square of Eqs. (5) and (6) achieve 0.7039 and 0.8078, respectively. We also try to apply exponential function (Song et al., 2017) to express utility by using the same data of the questionnaires. The nonlinear regression shows R-square achieve 0.5533 and 0.6003 for APs' and WPs' utility function, respectively, which do not match the practical data as well as our model. Therefore Eqs. (5) and (6) are better to express patients' utility.

Remark. The time-invariant threshold-based prioritization scheme decreases the variance of WPs' waiting time. For a verification of the remark, please see Appendix A.

As stated that some WPs who encounter the plugging in of batch of APs may have quite longer waiting time than others, even if they arrive hospital at similar time. Decreasing the variability of waiting time of WPs indicates that time-invariant threshold-based prioritization scheme could improve the system's fairness and decrease WPs' dissatisfaction.

4.2. The simulation optimization algorithm

We build a model to simulate the scheduling process in outpatient clinic in Microsoft Visual Studio 2012. Then, a heuristic simulation optimization algorithm is proposed to help us search the best results. The algorithm combines PSO and OCBA, where PSO provides a general framework to search optimal solutions and OCBA considers eliminating the effect of simulation randomness.

Following the PSO framework, the feasible search area is divided into R same size regions and each region is allocated several particles as initial solutions. Then, the algorithm will search recursively to find a better solution. In this paper, the fitness value of each particle is the patients' weighed average utility, which can be achieved from Eq. (1). The particle has N+2 dimensions including the number of appointments in each slot and a pair of control thresholds, defining as $X_i = (a_1, a_2, ..., a_N, \tilde{V}^{AP}, \tilde{V}^{WP})$. We propose the method to update the particles in each iteration through an integration. In each iteration, the particle updates itself twice. The first update is based on the OCBA (Chen et al., 2000), which considers the allocation of limited simulation budget based on stochastic simulation output to optimize the probability of correct selection. It determines optimal numbers of particles in each region. The second update is based on the PSO (Zhou et al., 2018), which mimics the predation behavior of a flock of birds to search for the optimal solution. It guides particles towards the best particle of their own region at a given speed.

In the first updating process, we calculate the mean $\overline{f_r}$ and standard deviation σ_r of the fitness value of the particles which belong to their own region r according to Eq. (7) and Eq. (8).

$$\overline{f_r} = \sum_{j=1}^{c_{r,t}} f_{rj} \Big/ c_{r,t} \tag{7}$$

$$\sigma_r = \sqrt{\sum_{j=1}^{c_{r,t}} (f_{rj} - \overline{f_r})^2 \Big/ (c_{r,t} - 1)}, r = 1, 2, ..., R; find(k) = \arg\min_r \overline{f_r} \tag{8}$$

where t is the current iteration step. $c_{r,t}$ is the number of particles in region r .Then the new numbers of particles in each region, $c_{1,t+1}, c_{2,t+1}, \ldots, c_{R,t+1}$, are calculated according to Eq. (9) and Eq. (10), which helps to get more particles from the region that may contains better results.

$$\frac{c_{l,t+1}}{c_{j,t+1}} = \left(\frac{\sigma_l / d_{l,k}}{\sigma_j / d_{j,k}} \right)^2, for all\, l \neq j \neq k, \text{and} d_{l,k} = \overline{f_l} - \overline{f_k} \tag{9}$$

$$c_{k,t+1} = \sigma_k \sqrt{\sum_{l=1, l \neq k}^{R} \frac{c^2_{l,t+1}}{\sigma_l^2}} \tag{10}$$

According to the new numbers of particles allocation, the specific-number locations of particles in each region are randomly created. These current locations are randomly assigned to each particle as its first update $X_{io}(\mathrm{t}+1)$.

In the second updating process, particles adjust their speed $V_i(\mathrm{t+1})$ and position $X_i(\mathrm{t+1})$ by the following Eq. (11) and (12).

$$V_i(\mathrm{t+1}) = cm(X_{rBest} - X_{io}(\mathrm{t+1})) \tag{11}$$

$$X_i(\mathrm{t+1}) = X_{io}(\mathrm{t+1}) + V_i(\mathrm{t+1}) \tag{12}$$

where c is the cognitive parameter and m is a random number. X_{rBest} is the position of the best particle in region r ,which have a highest fitness f_{rBest}. Those formulas guide a particle to search a better solution.

The best solution is obtained by repeatedly updating the solution according to the method above. If the iteration of the optimization algorithm meets the maximum number of iterations, this algorithm will end, obtaining the optimal solution, $X_i^* = (a_1, a_2, ..., a_N, V^{AP}, V^{WP})$ and return the corresponding objective value f_i^* . Otherwise, it will continue.

The steps of the algorithm are summarized as follows.

Step 1. Divide feasible solution space into R regions and original particles X_i are generated uniformly and randomly in the R regions.

Step 2. Execute the simulation model and calculate the weighted average utility of patients' satisfaction f_i .

Step 3. Partition particle swarm according to the positions of R regions

If $X_i \in region\ \mathrm{r}$, $f_{rj} = f_i$ and $c_{r,t}$=$c_{r,t}$+1,$j = 1, 2, ..., c_{r,t}$

Step 4. Update the region-best in region r according to:

If $f_{rj} < f_{rBest}$, $f_{rBest} = f_{rj}$, X_{rBest}=X_{rj} .

Step 5. Calculate the mean of particles $\overline{f}_r$, and standard deviation σ_r according to Eq. (7) and Eq. (8)

Step 6. Calculate the new number allocation of particles in each region, $c_{1,t+1}, c_{2,t+1}, \ldots, c_{R,t+1}$ according to Eq. (9) and Eq. (10). Then, generate certain numbers of particles uniformly and randomly in the R regions, and assign current locations randomly to each particle.

Step 7. Adjust the location and speed for each particle according to Eq. (11) and (12) .

Step 8. Terminating condition.

If t < max iteration, return step 2, else go to step 9.

Step 9. Obtain the $X^* = (a_1, a_2, ..., a_N, V^{AP}, V^{WP})$ and $f^* = \max\{f_1, f_2, ..., f_r\}$. The algorithm ends.

5. Research results

In this section, we first present proof-of-concept computational experiments based on real hospital data to verify the efficiency of our integrated optimization-based scheduling approach. We then examine how several system performance measures with the optimal strategy behave with respect to varied environmental and behavioural factors. Finally, we report a pilot study we conducted at our partnering hospital to demonstrate the practical benefit of our integrated optimization approach.

5.1. Computational experiments at the baseline

We conduct a case study based on the outpatient scheduling operations of Shanghai Sixth People's Hospital, a comprehensive hospital, which has 33 clinical departments and 9 technical departments. The endocrine department, as a key department to the hospital, allows both APs and WPs to be seen outpatient. From our field investigation, the hospital uses a myopic appointment scheduling strategy in their practice. That is, allocating an even number of appointments in each slot and assigning APs higher priority than WPs when both types of patients wait in the queue. We refer to this strategy as EA-APF (i.e., Even appointment Allocation with Appointment Patient First). The hospital practitioners voice their concern on low satisfaction of the WPs which are the majority of the outpatients. Motivated by the practical concern, we evaluate our optimization-based strategy (Optimal appointment Allocation with Real-Time Prioritization, OA-RTP) with a case study based on realistic hospital data. Through this study, we find that our approach could result in noticeable improvement on WP satisfaction while not compromising much on AP satisfaction, compared to EA-APF.

We generated numerous computational scenarios based on the historic data to examine how the proposed OA-RTP strategy performs in general. First, to model WP arrivals, we extracted records of outpatient visits from 2013 to 2017 and filtered out invalid and irrational records. The data showed that daily WP arrivals fluctuated on a weekly basis. We thus divided WP arrivals according to the time of the day and day of a week, and computed a point estimate for each time-dependent arrival rate and its 95% confidence interval. For each scenario, we sampled uniformly the arrival rate for each slot-day combination from the corresponding range. To model AP arrivals, we first specified the total number of APs based on the realistic percentage of AP volume to the total outpatients. This total number varied by the day of the week. Meanwhile, we constructed day-of-the-week specific ranges on the proportions of early and late arriving APs to the total APs. For each scenario, we sampled uniformly the numbers of early and late arriving APs from their respective ranges for each slot. For the service time of each outpatient, we found that a normal distribution would be a good fit to this random quantity and sampled the service time from the distribution. To sample the number of doctors, we first set the lower and upper bounds of the outpatient service system utilization to be 0.7 and 0.9, respectively. Then we specified a range on the number of doctors accordingly and drew a uniform sample from the range.

Patient satisfaction is set by the utility function. Given V_{ET} is a constant and has no influence on patient utility with respect to the waiting time, we let V_{ET}=2000, c=300 to avoid negative utility. We interviewed some patients and doctors about the expected waiting time in queue and determined ET^{AP} = 20 min and ET^{WP} = 40 min. For other parameters, we conduct a field experiment by the methodology we proposed in section 4.1 and through regression analysis, we obtain $\alpha^{AP}=\beta^{AP}=0.31$, $\alpha^{WP}=\beta^{WP}=0.28$, $\lambda^{AP}=1.81$, $\lambda^{WP}=1.51$.

We show in Table 1 the comparisons of average performance measures over the scenarios between our proposed strategy (i.e., OA-RTP) and the hospital implemented myopic strategy (i.e., EA-APF). In the table, the value of *k* in each column under OA-RTP represents the relative weight of AP satisfaction to WP satisfaction. We report the scenario-average of the coefficient of variation (CV) on the number of patients seen in each slot, i.e., CV = $\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}[a_i+w_i-\overline{(a+w)}]^2}}{\overline{(a+w)}}$, where $\overline{(a+w)}=\frac{1}{N}\sum_{i=1}^{N}(a_i+w_i)$, to quantify the effect of changing the appointment slot-assignment of APs (i.e., actionable decision in this paper) on the arrivals of all outpatients. In addition, we report the percentage of priority improved WPs (PIWP), i.e., the percentage of WPs in the queue are adjusted to see doctors earlier than awaiting APs over time. Finally, we repot the sample mean and variance of AP and WP waiting times over both patients and scenarios.

Table 1. Average performance comparison between OA-RTP and EA-APF over scenario

Average performance measures	EA-APF	OA-RTP							
		k=0.5	*k*=1	*k*=1.5	*k*=2	*k*=2.5	*k*=3	*k*=3.5	*k*=4

CV	0.63	0.52	0.55	0.57	0.58	0.59	0.60	0.61	0.61
PIWP (%)	0	26.76	21.62	19.19	17.57	10.81	5.95	4.59	4.05
Avg. WT of AP (min)	18.98	28.72	24.43	22.77	20.29	19.49	18.50	17.56	17.38
Avg. WT of WP (min)	54.66	25.95	27.64	29.02	31.25	39.63	44.60	48.92	51.61
Var. WT of AP (min^2)	19.09	55.42	49.78	44.25	30.59	27.05	24.51	22.54	21.26
Var. WT of WP (min^2)	656.62	60.73	65.86	67.12	70.4	129.51	171.08	203.83	232.96

With a large weighting coefficient, e.g., k = 4, the average CV only decreases slightly. Since WP satisfaction is not cared much, WP scheduling priority does not change much in real time (i.e., average PIWP is only about 4%). Thus, our OA-RTP based scheduling strategy does not lead to sufficient reduction in the WP waiting time. With a small coefficient, e.g., k = 0.5, the average CV result suggests that OA-RTP can noticeably improve the balance between AP and WP arrivals through improved AP appointment scheduling (i.e., the average CV decreases by 20%). However, since WP satisfaction is cared too much, more than a quarter of WPs are adjusted to be seen earlier than SAPs. It makes the AP waiting time increases significantly. As a result, we conclude that WP satisfaction is overly compensated. When we disseminated the research findings to the hospital administrators, they concluded that setting $k = 2$ can effectively reduce the WP waiting time without making much a compromise on AP satisfaction.

Next, we compare our OA-RTP strategy with two strategies that are not of two phases but not optimization based in both phases. The two strategies are Even appointment Allocation with Real-Time Prioritization (EA-RTP) and Optimal appointment Allocation with Appointment Patient First (OA-APF). EA-RTP implies assigning an even number of appointments to each slot and applying real-time prioritization similar to that in OA-RTP. OA-APF implies optimizing appointment allocation at each slot but always giving APs higher priority than WPs. We show the average performance results over scenarios in Table 2.

Table 2. The average performance comparison of the four strategies

Policy	EA-APF	EA-RTP	OA-APF	OA-RTP
Object Function	4465.39	7843.24	6149.88	8019.9
Avg. WT of AP (min)	18.98	20.83	21.03	20.29
Avg. WT of WP (min)	54.66	34.83	41.63	31.25
Var. WT of AP (min^2)	19.09	45.10	38.66	30.59
Var. WT of WP (min^2)	656.62	109.09	95.32	70.4

We find that according to the objective function, the average utility of patients, the OA-RTP performs best among most scenarios, followed by the EA-RTP, and then the OA-APF. The current policy EA-APF performs the worst. In other words, either applying optimal number of appointments in each slot or time-invariant threshold-based prioritization scheme is effective to improve patients' satisfaction, while the integrated optimization achieves the most improvement. Furthermore, it is noted that when the value of objective function is best, the average and variance of patient's waiting time also performs well. As shown in Table 2, three optimization polices decrease WPs' average waiting time significantly while increase AP's slightly. OA-RTP makes WPs wait the shortest and prolongs APs the least. The variance of patient's waiting time is the smallest under OA-RTP, which indicates it most effectively improves system's fairness and WPs' satisfaction. It justifies that the objective function based on the prospect theory can provide a viable supplement to patient-centred outpatient scheduling.

5.2 Sensitivity analysis

We next varying the environment factors and behaviour factors to assess the integrated scheduling optimization approach performs as opposed to comparative strategies.

5.2.1. With respect to environment factors

Our partnering hospital aims to promote appointment making among patients rather than walk-in with no appointment. This aim will lead to a reduced proportion of WPs. Hence, we are interested whether the proportion reduction will greatly influence the efficiency of our approach and under which scenarios the approach will be invalid. As shown in Fig. 4, with the AP proportion changing from 10% to 85%, the optimization based scheduling strategy we proposed has a up to 20% improvement in patient utility. However, when AP proportion is too low or too high, one of the two patient types will occupy the majority of outpatients, leaving little space for time-invariant threshold-based prioritization scheme to work. For the studied department, we estimate that the satisfaction improvement is most significant when APs take up around 70% of patients.

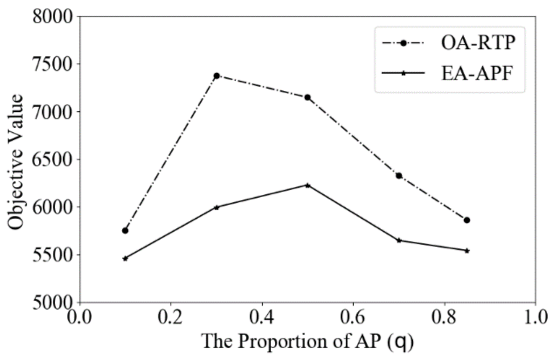

Fig. 4. The objective value comparison under different AP proportions

5.2.2. With respect to behaviour factors

The optimization results of our approach are also influenced by behaviour factors such as ET^{WP} ET^{AP}, α^{AP} α^{WP}, β^{AP} β^{WP} λ^{AP} λ^{WP}

study

find that when ET^{WP} is smaller, as shown in Fig. 5, the optimization effect of OA-RTP is more obvious. Because in this case, WPs are easier to become unsatisfied, then have more chance to be scheduled before APs and help the system works better. Similar conclusions can be obtained when

ET^{AP} increases. Additionally, we observe that OA-RTP could effectively improve patients' utility with different values of α^{AP}, α^{WP}, β^{AP}, β^{WP}, λ^{AP} and λ^{WP}. Figure 6 reports the objective function values when we vary the loss aversion parameters λ^{AP} and λ^{WP} simultaneously. From the figure, we observe that when patients were more sensitive to long waiting time (i.e., more likely to have aversion behavior), the patient satisfaction under EA-APF would decrease greatly. However, through adjusting the service priority of unsatisfied WPs by OA-RTP, we can achieve a similar system-wide satisfaction within a reasonable range of the loss aversion parameter. It suggests that our integrated optimization based scheduling approach could take more advantage of patients with increased loss aversion behaviour.

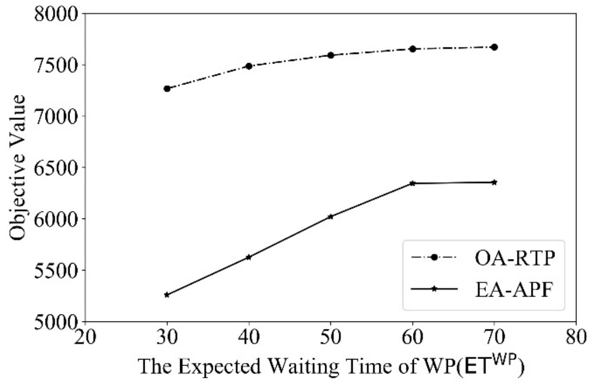

Fig. 5. The objective value comparison under different ET^{WP}

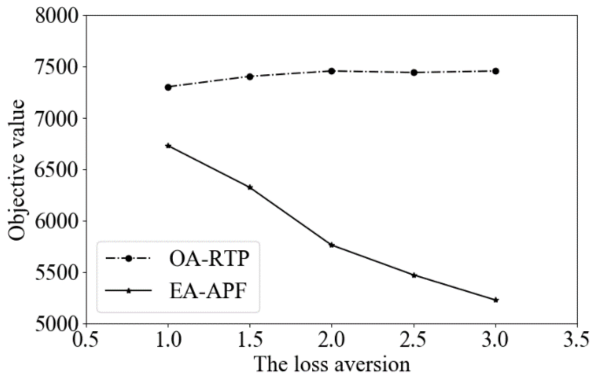

Fig. 6. The objective value comparison under different λ^{AP} and λ^{WP}

5.3 *Real-world pilot study*

We have presented our previous results of computational experiments to the Shanghai Sixth People's Hospital, and they agreed to conduct the pilot study in the hospital. We first analyze patients' daily visits and achieve the average arrival pattern for each weekday. Thus, according to the real-world case, we design different appointment allocation and corresponding thresholds for each weekday. We began our pilot first in the endocrine department in early 2018.

With one year's pilot, we evaluated the effects of the practice. As seen in Table 4, we compared the current results with the performance history of 2017 when the department was managed under EA-APF. We find that our proposed strategy decreases WP's average waiting time by 22.76% while causes no significant increase to AP's average waiting time despite the fact that the number of total patient visits increases by 1.57%. The variance of WPs' waiting time is decreased by 89.48% with APs' increases by 9.28%. The managers of the hospital also informed of us that patients reacted well to the new policy and complained less than before. The positive results show the method's efficiency and the hospital is considering extending our method into other departments.

Table 4. The performance comparison of Shanghai Sixth People's Hospital

Year	Number of patient visits	Avg. WT of AP (min)	Var. WT of AP (min^2)	Avg. WT of WP (min)	Var. WT of WP (min^2)
2017	201319	19.25	39.44	96.08	547.89
2018	204480	19.78	43.10	74.21	57.62

6. Conclusions and future research

This paper investigates outpatient scheduling in the situation where WPs account for a large proportion of patients. Considering slot-wise WP arrival pattern and incorporating patient waiting behaviour, we propose an integrated optimization method for the number of appointments in each slot and time-invariant threshold-based prioritization scheme to maximize patients' average utility. The utility function, based on the prospect theory, measures patient's real-time satisfaction. An experiment validates the expression and extracts practical parameters. The simulation model based on the algorithm, combining PSO and OCBA, helps us search the optimal solution. Finally, we conduct a pilot study in the endocrine department of Shanghai Sixth People's Hospital. The study demonstrates that our method has important practical value, decreasing patients' average waiting time as well as variability among WPs. Since our method was shown in the pilot study to work under different scenarios, it has been implemented in many departments of Shanghai Sixth People's Hospital.

Future research could be extended in several directions. One direction is to investigate the decision of optimal length of each slot to improve satisfaction of both APs and WPs. Another direction is to study how the length of slot affects AP's arrival behaviour. Though most literature assumes appointment patients' arrival is independent of appointment slot, we have data showing that for systems with both WPs and APs, the slot length does, indeed, influence the patient's arrival. The third direction is to consider heterogeneous patients, e.g., patients with different expected waiting time and waiting behaviour.

Acknowledgements

Our research would like to thank the domain experts in the Shanghai Sixth People's Hospital for helping with field investigation and data collection. This work was supported in part by the National Natural Science Foundation of China Grant 71871138, Grant 71432006 and Grant 71471114. This work was also supported by the U.S. National Science Foundation Grant 1405357.

References

Ahmadi-Javid, Amir, Zahra Jalali, and Kenneth J. Klassen. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research 258* (1):3-34.

Avineri, Erel. (2004). A Cumulative Prospect Theory Approach to Passengers Behavior Modeling: Waiting Time Paradox Revisited. *Journal of Intelligent Transportation Systems 8* (4):195-204.

Cayirli, Tugba, and Evrim Didem Gunes. (2014). Outpatient appointment scheduling in presence of seasonal walk-ins. *Journal of the Operational Research Society 65* (4):512-531.

Cayirli, Tugba, and Emre Veral. (2003). Outpatient scheduling in health care: A review of literature. *Production & Operations Management 12* (4):519-549.

Chen, Rachel R., and Lawrence W. Robinson. (2014). Sequencing and Scheduling Appointments with Potential Call-In Patients. *Production and Operations Management 23* (9):1522-1538.

Chen, Chun-Hung, Jianwu Lin, Enver Yücesan, and Stephen E. Chick. (2000). Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization. *Discrete Event Dynamic Systems 10* (3):251-270.

Chinese Ministry of Health. (2009). Opinions of the ministry of health on the implementation of appointment service in public hospitals. *Gazette of the National Health and Family Planning Commission of People's Republic of China 12*:8-10.

Denton, Brian, and Diwakar Gupta. (2003). A Sequential Bounding Approach for Optimal Appointment Scheduling. *IIE Transactions 35* (11):1003-1016.

Ge, D., G. Wan, Z. Wang, and J. Zhang. (2014). A note on appointment scheduling with piecewise linear cost functions. *Mathematics of Operations Research 39* (4):1244-1251.

Geng, N., and X. Xie. (2016). Optimal Dynamic Outpatient Scheduling for a Diagnostic Facility with Two Waiting Time Targets. *IEEE Transactions on Automatic Control 61* (12):3725-3739.

Gocgun, Y., B. W. Bresnahan, A. Ghate, and M. L. Gunn. (2011). A Markov decision process approach to multi-category patient scheduling in a diagnostic facility. *Artificial Intelligence in Medicine 53* (2):73-81.

Green Linda V., Sergei Savin, and Ben Wang. (2006). Managing Patient Service in a Diagnostic Medical Facility. *Operations Research 54* (1):11-25.

Gupta, Diwakar, and Brian Denton. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions 40* (9):800-819.

Ho Chrwan-Jyh, and Hon-Shiang Lau. (1992). Minimizing Total Cost in Scheduling Outpatient Appointments. *Manage. Sci. 38* (11):1750-1064.

Jiang Zhiyu. (2010). Study on the Status and Influencing Factors of Clinical Reservation Services. *Medicine and Society 23* (6):61-63.

Kahneman, Daniel, and Amos Tversky. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica 47* (2):263-291.

Kemper B., C. A. J. Klaassen, and M. Mandjes. (2014). Optimized appointment scheduling. *European Journal of Operational Research 239* (1):243-255.

Koeleman, Paulien M., and Ger M. Koole. (2012). Optimal outpatient appointment scheduling with emergency arrivals and general service times. *Iie Transactions on Healthcare Systems Engineering 2* (1):14-30.

Kong Qingxia, Chung Yee Lee, Chung Piaw Teo, and Zhichao Zheng. (2013). Scheduling Arrivals to a Stochastic Service Delivery System Using Copositive Cones. *Operations Research 61* (3):711-726.

LaGanga, Linda R., and Stephen R. Lawrence. (2012). Appointment Overbooking in Health Care Clinics to Improve Patient Service and Clinic Performance. *Production & Operations Management 21* (5):874-888.

Qu, Xiuli, Yidong Peng, Jing Shi, and Linda LaGanga. (2015). An MDP model for walk-in patient admission management in primary care clinics. *International Journal of Production Economics 168*:303-320.

Robinson, Lawrence W., and Rachel R. Chen. (2003). Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions 35* (3):295-307.

Song Jie, Yunzhe Qiu, and Zekun Liu. (2017). A Real-Time Access Control of Patient Service in the Outpatient Clinic. *IEEE Transactions on Automation Science & Engineering 14* (2):758-771.

Wang Shan, Nan Liu, and Guohua Wan. (2018). Managing Appointment-Based Services in the Presence of Walk-In Customers. *Management Science* forthcoming.

Yan C., J. Tang, and B. Jiang. (2014). Sequential appointment scheduling considering walk-in patients. *Mathematical Problems in Engineering 2014* (4):1-12.

Zacharias, Christos, and Michael Pinedo. (2014). Appointment scheduling with no-shows and overbooking. *Production & Operations Management 23* (5):788-801.

Zhou, Hongming, Jihong Pang, Ping-Kuo Chen, and Fuh-Der Chou. (2018). A modified particle swarm optimization algorithm for a batch-processing machine scheduling problem with arbitrary release times and non-identical job sizes. *Computers & Industrial Engineering 123*:67-81.

Appendix

A. A Verification of the Remark on Page 15

The real-time service prioritization in our integrated optimization based scheduling approach applies a prioritization rule that specifies UAP > UWP > SAP > SWP. For SWPs whose satisfaction is higher than the threshold, their priority will not change, so that their waiting time remains the same, compared to the system where APs always have a higher service priority. On the other hand, for UWPs, their waiting time will become long. In contrast, with a prioritization rule that specifies UWP > SAP, UWPs' waiting time is reduced as they are scheduled to be served earlier than SAP. Overall, for all WPs, the waiting time distribution tends to have lower average and reduced variance.

Let us take a numerical example to further illustrate our remark. We assume APs and WPs follow a Poisson process with mean arrival rate 0.08 and 0.07, respectively. Their service time is modelled identical with normal distribution $N(6,0.4)$. If APs always have a higher priority than WPs, the arrival and departure times of each patient in the first slot (8:00~9:00) are shown as follows.

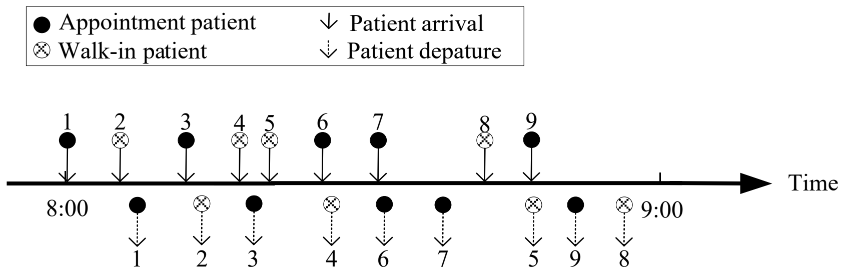

Fig.7 Arrivals and departures under the current prioritization rule

The sequence of patient service is: 1→2→3→4→6→7→5→9→8. The waiting time of each patient is reported in Table 5.

Table 5 Waiting time under the current prioritization rule

Patient index	1	2	3	4	5	6	7	8	9
Waiting time/min	0	3	3	2	18	2	1	10	0
Avg. of WPs' waiting time/min					8.25				
Var. of WPs' waiting time /min^2					41.19				

Let $\tilde{V}^{WP}=V_{WP}(8)$, $\tilde{V}^{AP}=V_{AP}(5)$. Then under the time-invariant threshold-based prioritization scheme, patients are scheduled as follows.

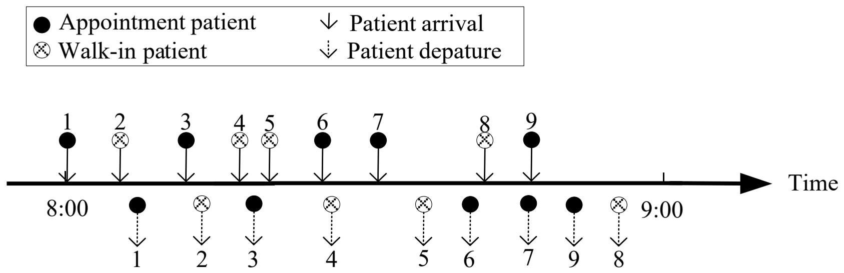

Fig.8 Arrivals and departures under our prioritization rule

The sequence of patient service is: 1→2→3→4→5→6→7→9→8. The waiting time of each patient is reported in Table 6.

Table 6 Waiting time under our prioritization rule

Patient index	1	2	3	4	5	6	7	8	9
Waiting time/min	0	3	3	2	10	12	12	10	0
Avg. of WPs' waiting time/min					6.25				
Var. of WPs' waiting time /min^2					23.6875				

We observe from Table 6 that the WP indexed 5 is scheduled to receive service before the AP indexed 6 because his/her utility is lower than the threshold but AP is not. The waiting time of this WP indexed is thus decreased from 18 minutes to 10 minutes. As for other WPs (indexed 2, 4 and 8), their waiting times are unchanged. As a result, the variance of WP waiting time is decreased by 42.49%.

B. The outpatients' waiting time sensitivity questionnaire

In this appendix, we provide the questionnaire we used to survey online respondents and measure parameters in utility function based on the prospect theory. The original questionnaire was written in Chinese. We provide a English translation here.

Introduction

You are being invited to take part in a research study about waiting time sensitivity. Please note that there are no right or wrong answers to any questions in this questionnaire.

We are only interested in your opinions and feedback. Your kind and valid response will help the Sixth People Hospital in Shanghai devise better appointment policy and will help make you feel more satisfied about waiting for medical service in the future. This questionnaire should take approximately 1-3 minutes to complete. The information will be kept confidential and will not be linked to any personal identifiable information. Thank you for your participation and cooperation!

1) Do you make an appointment in advance before going to the clinic?

a) Yes b) No

In the following, we will present two fictitious scenarios where we would like to measure your sensitivity about waiting time. Please note that there are no correct or incorrect responses, and your choice should be based on your own preferences, experiences, and specific needs.

Suppose you are currently queuing in a general outpatient clinic of an innovative medical institution. You have made an appointment before visiting hospital[3]. Assume that the institution has told you are expected to wait for 20 minutes[4], however you are now waiting for 5 minutes[5] more than expected. What would you score to reasonably express your satisfaction with this hospital?

2) Please enter the score.(Suppose 60 represent the score to express your satisfaction when you wait as long as expect.)

____.[6]

---

[3] The situation is consistent with respondent's actual patient type. If the respondent choose b) in question 1), then this sentence is replaced with "You have not made an appointment before visiting hospital"

[4] The expect waiting time is rely on respondent's actual patient type. Appointment patient's expect waiting time is 20 minutes and walk-in patient's is 40 minutes.

[5] The overtime is randomly assigned to the respondent with some specific range.

[6] It is a scroll-down selection box, which include the integer number from 0 to 100.

If you are now waiting for 10 minutes[7] less than expected. What would you score to reasonably express your satisfaction with this hospital now?

3) Please enter the score. Suppose 60 represent the score to express your satisfaction when you wait as long as expected.

____.[8]

[7] The reduce-time is randomly assigned to the respondent with some specific range.

[8] It is a scroll-down selection box, which include the integer number from 0 to 100.