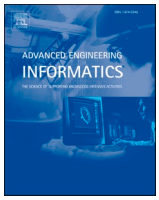
Advanced Engineering Informatics 59 (2024) 102252

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Full length article

Multiserver time window allowance schedules for virtual visits with uncertain time-dependent no-shows and service times

Xiaoxiao Shen [a], Na Li [a,*], Xiaoqing Xie [b]

[a] Department of Industrial Engineering and Management, School of Mechanical Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

[b] Department of Economics and Decision Sciences, China Europe International Business School, 699 Hongfeng Road, Shanghai 201206, China

ARTICLE INFO

Keywords:
Virtual visit
Appointment scheduling
Time window allowance schedule
Stochastic programming
Benders decomposition

ABSTRACT

Virtual care serves as a new mode that can divert non-urgent visits from traditional office visits. Whether virtual service can improve the access to medical treatment and reduce the burden of traditional office services, the key issue is to generate efficient appointment schedules with the lowest operation cost. In this paper, considering the uncertainty of time-dependent no-shows and service times, we investigate a multiserver time window allowance appointment scheduling problem, where time window constraints that restricts virtual visits to be served during the particular period are explicitly modeled. We formulate the problem as a stochastic mixed-integer program to optimize decisions of physician allocation and appointment time simultaneously. Based on the sample average approximation, a stabilized Benders decomposition algorithm is developed by incorporating acceleration techniques, such as cut aggregation and feasibility cuts. Numerical results based on real data indicate the effectiveness of the proposed multiserver time window allowance schedules (MTWAS) and algorithm. Comparing with the off-the-shelf solver Gurobi, the developed algorithm demonstrates high performance in terms of computation speed and solution quality. Under different time-dependent no-show patterns of virtual and office visits, the obtained MTWAS perform better than previous solutions in almost all test cases. In addition, we offer useful managerial insights to aid the virtual service provider in making better scheduling decisions.

1. Introduction

As an emerging type of digitalization medical service, Internet hospitals provide patients with virtual appointments through messages, phone calls, or videos. Virtual visits and telemedicine have emerged as essential options with the advantages of rapid response and no transportation. However, various operation management ambiguities occur in the combination process of virtual and office services [1]. Appointment systems assist in lowering the variability in patient arrival times, resulting in shorter wait times for patients and continued high system utilization. In order to make virtual services truly alleviate the burden on office service and improve medical accessibility, appointment scheduling systems are required further studies to control the patient demand for various services.

In the one of Internet hospitals we surveyed, virtual services are available in almost all departments such as internal medicine, surgery, gynecology, and traditional Chinese medicine. The appointment time available for virtual service of each department is generally in certain mornings and afternoons of certain days. For example, appointments for gastroenterology can be made on Monday afternoon, Wednesday afternoon and Saturday afternoon. The current practice is generally to centralize treating virtual visits at certain periods during the day, with few departments having virtual care available throughout the day. Each department has multiple doctors to choose, and each appointment is fixed for half an hour. Serious issues with long patient waiting and low physician utilization are emphasized in our interaction with the hospital. The hospital struggle with making efficient appointment scheduling decisions and seeking easy-to-implement guidelines regarding specified appointment periods for virtual visits with in-depth systems analysis. This motivated us to study the appointment scheduling problem for virtual care where a new type of schedule named multiserver time window allowance schedule is formally proposed and analyzed.

One new challenge is the uncertainty inherently observed in practical service systems. Virtual service greatly breaks through time and geographic space limitations, which may lead to changes in the patient behavior. We consider stochastic time-dependent no-shows and service

* Corresponding author.
E-mail addresses: sjtusxx98@sjtu.edu.cn (X. Shen), na-li03@sjtu.edu.cn (N. Li), xkristine@ceibs.edu (X. Xie).

https://doi.org/10.1016/j.aei.2023.102252
Received 4 May 2023; Received in revised form 17 September 2023; Accepted 6 November 2023

times. These uncertainties pose great difficulties in solving the problem. To tackle such difficulties, we deploy a Benders decomposition framework based on the sample average approximation (SAA) approach to formulate the problem as a stochastic mixed-integer program (SMIP). In addition, a scenario generation procedure is proposed to address the issue of a lack of no-show data in some appointment times. We develop a stabilized Benders decomposition algorithm with several accelerated measures to solve this problem. Numerical studies are implemented to illustrate the superior performance of proposed multiserver time window allowance schedules compared with previous single-server schedules and traditional schedules without time windows.

The main contributions of this study are threefold. First, we investigate the appointment scheduling in the presence of virtual medical service. A new type of schedule named multiserver time window allowance schedule is formally proposed, optimized and analyzed to manage virtual visits with uncertain time-dependent no-shows and service times. To promote the healthcare digital transformation, this issue is meaningful and interesting and deserves to be investigated. Second, we propose a stabilized Benders decomposition algorithm based on several acceleration techniques. Numerical results demonstrate that our method outperforms the benchmark in computation time and solution quality. The proposed algorithm can achieve ϵ-optimal solutions by a surprisingly small number of iterations even if the optimality tolerance ϵ is very small. Finally, through numerical studies, we conclude that the proposed MTWAS solutions almost always illustrate superiority over the schedules obtained in previous studies. In addition, we offer interesting insights as guidelines provided to hospitals, which aids practitioners in conducting virtual medical care. First, instead of separately scheduling appointments for each physician using a single-server model, we suggest jointly scheduling multiple servers when physicians treat virtual and office visits significant differently. Second, under different show-up patterns, adopting reasonable MTWAS generates superior performance than the traditional schedule without time windows.

We organize the remainder of this article as follows. In Section 2, we review and summarize the relevant literature. Section 3 describes the problem and explicitly formulates it as a stochastic mixed-integer program. Then, in Section 4, we provide the details of our proposed stabilized Benders decomposition algorithm. In Section 5, numerical studies are conducted, and some potential insights are explored. Section 6 summarizes the conclusions and presents directions for future work.

2. Literature review

In this section, to elaborate the contribution of our study, we review mainly healthcare operations management literature on various service requests and relevant works from the appointment scheduling literature that are either very close to this study or very recent.

2.1. Operations management literature on various service requests

Various service requests, such as prescheduled [15,16], walk-in [17–19,40,41], emergency [20,21], revisit [22,23], and ambulatory care services [52], are investigated and modeled in the healthcare operations management literature. Marynissen and Demeulemeester [48] focus on reviewing the literature on multi-stage appointment scheduling problems. Gupta and Denton [46] provide a thorough review on appointment scheduling systems applied in different settings, such as primary, specialty care clinics and elective surgeries. Operations management about virtual visits has gained growing research interest in recent years. Zhong et al. [10] use queuing theory to study the conditions for system parameters to improve service outcomes when introducing e-visits. Xiang et al. [11] investigate the performance of different scheduling policies for managing both office and virtual visits using a probabilistic analytical approach. The results reveal that the first-come-first-serve strategy achieves optimal system performance. Zhong [1] model the appointment capacity planning problem with office and virtual visits as a queueing model. They analyze the effect of different system parameters on the system queue length and conclude that not more virtual visits are better. In chronic care, Bayram et al. [12] formulate the problem of allocating capacity for office and virtual appointments as a stochastic dynamic programming model to maximize the aggregate health benefits. To fulfil the diverse needs of virtual visits and improve the efficiency of online physicians' resource allocation, Wen et al. [13] study the problem of recommending physicians for patients on an online platform. They use a heuristic algorithm to obtain the optimal physician information display order that maximizes the physician–patient matching degree. Pan et al. [14] study the matching problem for randomly arriving virtual visits with perishable resources. They propose an online heuristic algorithm for deciding whether to accept a patient and which physician to assign to the patient to maximize the total expected reward. Shen et al. [69] develop a modified progressive hedging algorithm to solve the advance schduling problem considering online or offline revisit uncertainty and continuity of care. Based on the above-reviewed literature, the research gaps are found out. To the best of our knowledge, research on operation management of the virtual service is still quite limited and is confined to physician recommendation strategies, capacity planning, and advance scheduling of appointment days. The goal of this study is to investigate the appointment scheduling problem to optimize the appointment time for virtual care. A new type of schedule named multiserver time window allowance schedule is formally developed that determines the schedule of virtual visits in a specified period, which has important practical implications for launching virtual care service.

2.2. Uncertainties in appointment scheduling

There is a vast body of literature on appointment scheduling addressing uncertainties in health care delivery systems. Studies [2,3] show that patients exhibit the significant phenomenon of no-shows and that failure to consider no-shows in appointment scheduling results in a waste of resources. Since no-shows can have harmful effects [26], methods such as overbooking [3,5,30,44,49,51], open access [15,31,39], and adjustable appointment intervals [4,29] have been developed to remediate the adverse impacts of no-shows in the literature. Dantas et al. [47] present a systematic review of no-shows in appointment scheduling. The optimal decision [50] and policy [54] of appointment scheduling are investigated with considerations of patient no-shows. Stochastic service time and patient no-shows are mainly addressed. Concerning the uncertainty of service time [42], Chen and Robinson [15] and Zhou and Yue [28] generate many equally likely scenarios to capture the uncertainty of service time. A few studies address this uncertainty by optimizing the worst-case performance to achieve robust solutions. For example, Jiang et al. [29] and Kong et al. [4] develop distributionally robust optimization models by assuming that the probability distribution of the uncertain service time is ambiguous. The difference is that the distribution set considered by Jiang et al. [29] is only characterized by first moment information, while Kong et al. [4] assume both first and second moment information is known. Mandelbaum et al. [7] use a data-driven approach considering uncertain service time in a cancer infusion unit. Abovementioned studies [4,15,28,29] use the same approach to address stochastic no-shows. To characterize the no-show, the heterogeneous no-show is considered in some recent studies. For instance, the no-show is job-dependent in papers [3,28,32], meaning that different patients have different show-up rates. Kong et al. [4] and Zhou and Yue [28] show that customer no-show behavior depends on the time of day. Liu and Ziya [5] further find that patient heterogeneity and no-show rates are correlated. Different no-shows and service times for patients are considered in [24–27]. Similar to Kong et al. [4] and Zhou and Yue [28], this study addresses stochastic service time and time-dependent no-shows. In our analysis, we focus on investigating how the proposed MTWAS performs

Table 1
Model notations.

	Symbol	Definition
Sets	K	Set of servers with $K = \{1, \cdots, m\}$
	I	Set of all appointment intervals with variable lengths $I = \{1, \cdots, n\}$
	S	Set of scenarios
Parameters	$r_o(r_e)$	Proportion of office (virtual) visits to be scheduled
	q_i^s	Show-up status of the patient assigned to appointment interval i in scenario s
	d_i^s	Service time of the patient assigned to appointment interval i in scenario s
	$c_w^o(c_w^e)$	Unit waiting time cost of the office (virtual) patient
	c_k^o	Unit overtime cost of server k
	c_k^z	Unit idle time cost of server k
	T	Regular working session
	M	A large number
	$[\underline{a}^o, \overline{a}^o]([\underline{a}^e, \overline{a}^e])$	Time window for the scheduled arrival time of the office (virtual) patient
Decision variables	$x_{ik} \in \{0, 1\}$	Binary variable, 1 if appointment interval i is set on server k, 0 otherwise
	$a_i \in \mathbb{R}$	Continuous variable, the scheduled arrival time of the patient assigned to appointment interval i
	$y_i^o(y_i^e)$	Binary variable, equal 1 if an office (virtual) patient is assigned to appointment interval i and zero otherwise
	$w_i^s \in \mathbb{R}$	Continuous variable, denoting the waiting time of the patient assigned to appointment interval i in scenario s
	$z_k^s \in \mathbb{R}$	Continuous variable, capturing the idle time of server k in scenario s
	$o_k^s \in \mathbb{R}$	Continuous variable, representing the server k's overtime in scenario s

and exploring potential insights for adopting MTWAS when virtual and office visits present different time-dependent show-up patterns.

2.3. Patient scheduling on single or multiple servers

Previous literature modelled single or multiple service providers for different appointment scheduling systems. Several papers formulate the single-server scheduling problem as stochastic programming model [4,15,29] or single-server queuing model [5,33]. Assigning appointment to multiserver has aroused interests in many application areas, such as surgery scheduling [6], cancer chemotherapy [7–9], specialty care [34], and outpatient clinic [35–38,45]. Abovementioned works model problems with the framework of stochastic optimization and solve them with exact or heuristic iteration optimization algorithms. Some optimization goals, such as maximizing net reward or minimizing the expected sum of the patient waiting time and physician idle and overtime cost, are achieved by optimizing the scheduled arrival time interval between adjacent appointments. A few works employ queueing theory to analyze problems [32]. Pan et al. [36] propose a stochastic approximation algorithm under unbiased gradient estimators to solve the appointment scheduling problem considering multiple servers, no-shows and unpunctuality. Wu and Zhou [37] address the joint optimization of sequencing and scheduling appointments with random service durations and unpunctual times on multiple servers by developing a L-shaped based heuristic method. Shnits et al. [38] adopt a multi-server numerical-based algorithm to schedule a given sequence of patients and indicate how server pooling improves the system performances. Yan et al. [53] develop a stochastic overbooking model for outpatient clinics considering multiple physicians and patient preference for physicians and their appointment times. Alvarez-Oh et al. [55] formulate a two-stage stochastic integer programming model to schedule patients who need to be seen by one of two available nurses first and then to be seen by her provider. In our analysis, we mainly concentrate on examining the value of multiserver modeling as well as how the multiserver model functions when doctors exhibit various working rhythms, such as mean and standard deviation of service times.

2.4. Solution approaches for appointment scheduling

For solving the stochastic optimization models in appointment scheduling literature, Sample Average Approximation (SAA) approach is popularly adopted as an efficient scenario-based method [30,34,36], which generates many random samples and computes the average objective function values for all scenarios to approximate the objective value of the stochastic program. Benders decomposition algorithm are often used along with SAA to solve large-scale stochastic mixed-integer programs. For the field of appointment scheduling, see, for example, Zhou et al. [34], Jiang et al. [56], Chen and Robinson [15], Xiao et al. [57], and Zhou and Yue [28]. In addition, exact algorithms such as branch-and-price-and-cut [59], and heuristics [58,60] are developed to solve optimization problems of health care service. For Benders decomposition, acceleration techniques are adopted in literature, including selecting the strongest cuts generated by dual solutions [61,62], applying combinatorial cuts to remove fractional solutions to the master problem [63–66], and generating strengthened Benders' cuts [67]. For more acceleration techniques, please see Rahmaniani et al. [68] for a review. We develop an efficient stabilized Benders decomposition (SBD) algorithm to solve the appointment scheduling problem addressed in this study. Specifically, compared with the existing solution methods like those already discuss in abovementioned studies, various acceleration methods are carefully proposed including cut aggregation, feasibility cuts, and ε- optimal strategy to enhance the algorithm. The proposed SBD algorithm shows its superiority both in terms of computation speed and solution quality, especially for large-scale problems.

3. Problem formulation

In this section, we present the model description for the multiserver time window allowance appointment scheduling problem under uncertainty with both virtual and office visits. Some notation definitions are developed. Considering the uncertainty in service times and no-shows inherently, we formulate the problem as a stochastic mixed-integer program (SMIP) model.

Before virtual visits were implemented, a server's workday was primarily made up of scheduled outpatient office visits. As an alternative, if virtual medical service is provided, physicians must set aside time to process texts or videos. In this study, we focus on how clinic decision-maker to schedule both office and virtual visits for multiple physicians. The problem is that a decision-maker needs to schedule daily appointments for both office and virtual visits who make their appointment requests far in advance. We define n patients to be the total number of visits with a ratio of r_o office patient and r_e virtual patient. Set $I = \{1, \cdots, n\}$ is the set of appointment intervals with variable lengths to be determined. The decision-maker must assign each patient to an appointment interval, and must determine the scheduled arrival time for the patient assigned to appointment interval i, captured by a_i. Once each a_i is decided, the length of each flexible appointment interval is fixed. Through our investigation and interaction with practitioners, we found that physicians set a time window for texting, calling, video before or after all office visits, which is generally applied in the medical practice with Internet hospitals. Motivated by the above reality and facts, we assume that a_i must meet time window constraints for both patient types. That is, if an office (virtual) patient is assigned to appointment interval i, a_i must be within the time window $[\underline{a}^o, \overline{a}^o]([\underline{a}^e, \overline{a}^e])$, where $\underline{a}^o(\underline{a}^e)$ and $\overline{a}^o(\overline{a}^e)$ are the allowances for earliness and lateness in the office (virtual) patient's appointment time window. We define the decision variable y_i^o (y_i^e) to be a binary variable that equals one if an office (virtual) patient is assigned to appointment interval i and zero otherwise.

We define the set of parallel servers to treat patients simultaneously to be $K = \{1, \cdots, m\}$. The terms "server" and "physician" are used interchangeably. The decision variable x_{ik} is defined to be equal to one if appointment interval i is set on server k and zero otherwise. The servers may generate overtime or idle time. We calculate the overtime (the time of service completion time exceeding the planned session length T) of each server separately. There are two reasons for server vacancies. First, treatment of a patient may be completed before the scheduled arrival time of the next patient. Second, the no-show of a patient may impede servers from starting the treatment at the scheduled appointment time. We denote unit overtime cost or idle cost on server k as c_k^o and c_k^z, respectively. We assume a finite set of scenarios representing service times' and no-shows' uncertainty. The show-up status and the service time of the patient assigned to appointment interval i in scenario s is denoted as q_i^s and d_i^s, respectively. Note that $q_i^s \in \{0, 1\}$ with $q_i^s = 1$ if the patient shows up, and 0 otherwise. Uncertain no-shows (time-dependent and different between virtual and office visits) and service time (different between patient types and physicians) are realized based on our scenario generation (see Section 5 for details). The unit waiting time cost incurred for an office or virtual medical service that cannot start on time because of the delay in previous appointment operations is denoted as c_w^o and c_w^e, respectively. For all $i = 1, \cdots, n$, the waiting time of the patient assigned to appointment interval i in scenario s is denoted as w_i^s. Additionally, let variables o_k^s and z_k^s represent server k's overtime and idle time after finishing the server's assigned appointments in scenario s, respectively. The notations applied in our model are listed in Table 1.

The SMIP formulation of the multiserver time window allowance appointment scheduling problem for coordinating virtual and office visits where the total patient waiting time, idle time, and overtime of servers are minimized is as follows:

$$minE\left[\sum_{i=1}^{n}(c_w^o y_i^o w_i^s + c_w^e y_i^e w_i^s) + \sum_{k=1}^{m}(c_k^o o_k^s + c_k^z z_k^s)\right] \tag{1a}$$

$$s.t. \sum_{i=1}^{n} y_i^o = nr_o \tag{1b}$$

$$\sum_{i=1}^{n} y_i^e = nr_e \tag{1c}$$

$$y_i^e + y_i^o = 1, \forall i \in I \tag{1d}$$

$$\sum_{k=1}^{m} x_{ik} = 1, \forall i \in I \tag{1e}$$

$$a_i \le \overline{a}^o y_i^o + \overline{a}^e y_i^e, \forall i \in I \tag{1f}$$

$$\underline{a}^o y_i^o + \underline{a}^e y_i^e \le a_i, \forall i \in I \tag{1g}$$

$$a_p + w_p^s + M\left(2 - x_{pk} - x_{jk}\right) \ge a_j + w_j^s + q_j^s d_j^s, \forall p > j \in I, k \in K, s \in S \tag{1h}$$

$$z_k^s \ge a_i + w_i^s - \sum_{j \in I, j \ne i} q_j^s d_j^s x_{jk} - M(1 - x_{ik}), \forall i \in I, k \in K, s \in S \tag{1i}$$

$$o_k^s \ge \sum_{i \in I} q_i^s d_i^s x_{ik} + z_k^s - T, \forall k \in K, s \in S \tag{1j}$$

$$\left(o_k^s, z_k^s\right) \ge 0, \forall k \in K, s \in S \tag{1k}$$

$$w_i^s \ge 0, \forall i \in I, s \in S \tag{1l}$$

$$x_{ik}, y_i^o, y_i^e \in \{0, 1\}, \forall i \in I, k \in K \tag{1m}$$

Objective (1a) minimizes the total cost. The first term on the left calculates the sum of the waiting cost of office patients. The waiting cost of virtual visits is measured by the second term. The third term presents the total cost of the servers' overtime and idle time. Constraints (1b) and (1c) describe that there are exactly nr_o office visits and nr_e virtual patients to be scheduled. Constraints (1d) limit one patient type for each appointment interval. Constraints (1e) ensure that each patient is assigned to exactly one server. Constraints (1f) and (1g) restrict that each office and virtual appointment on server k starts within its requested time window, respectively. Constraints (1h) ensure that if appointment intervals $p \in I$ and $j \in I$ are set on the same physician and appointment interval $j \in I$ is scheduled before $p \in I$, then the service start time of the patient assigned to appointment interval $p \in I$ should start after the end of treatment of the patient assigned to appointment interval $j \in I$. Constraints (1i) and (1j) determine the idle time and overtime during the regular working session of server k, respectively. Constraints (1k), (1l), and (1m) define feasible ranges of the decision variables.

4. Solution approach

In this section, for a simple understanding, we first present the classical Benders decomposition with sample average approximation. By analyzing the structural characteristics of the SMIP formulation, the problem can be divided into a master problem and S subproblems associated with S scenarios based on the SAA approach. The master problem (*MP*) is as follows, where continuous variables δ_s are introduced to estimate the objective value of subproblem:

$$MP: min \sum_{s=1}^{S} \delta_s \tag{2a}$$

$$s.t. (1b) - (1g), (1m) \tag{2b}$$

Each generated scenario corresponds to one Benders subproblem. For scenario s, given values of assignment decisions (assigning appointment interval to servers $\boldsymbol{x}$, assigning patient type to appointment interval $\boldsymbol{y}$) and decisions about the patient's scheduled arrival time $\boldsymbol{a}$, one subproblem is obtained to minimize patient waiting time, server idle time, and overtime costs. The subproblem corresponding scenario s (SP^s) is as follows:

$$SP^s(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{\xi}(\boldsymbol{s})) : Q(\boldsymbol{x}, \boldsymbol{a}, \xi(\boldsymbol{s})) = min \sum_{i=1}^{n}(c_w^o y_i^o + c_w^e y_i^e) w_i^s + \sum_{k=1}^{m}(c_k^o o_k^s + c_k^z z_k^s) \tag{3a}$$

$$s.t. a_p + w_p^s + M\left(2 - x_{pk} - x_{jk}\right) \ge a_j + w_j^s + q_j^s d_j^s, \forall p > j \in I, k \in K \tag{3b}$$

$$z_k^s \ge a_i + w_i^s - \sum_{j \in I, j \ne i} q_j^s d_j^s x_{jk} - M(1 - x_{ik}), \forall i \in I, k \in K \tag{3c}$$

$$o_k^s \ge \sum_{i \in I} q_i^s d_i^s x_{ik} + z_k^s - T, \forall k \in K \tag{3d}$$

$$\left(o_k^s, z_k^s\right) \ge 0, \forall k \in K \tag{3e}$$

$$w_i^s \ge 0, \forall i \in I \tag{3f}$$

The Benders decomposition algorithm sequentially generates two sets of cuts optimality cuts and feasibility cuts. Feasibility cuts deliver essential conditions to ensure the feasibility of the primal subproblem, while the optimality cut provides the optimal value approximation of the primal subproblem when the subproblem is linear bounded. It is not necessary to generate feasibility cuts since the problem in (2a)–(2b), (3a)–(3f) has complete recourse (Proposition 1).

Proposition 1. *The SMIP problem (2a)–(2b), (3a)–(3f) has complete recourse.*

Proof. Observe that problem in (3a)–(3f) is always feasible no matter what realizations of the random vector $\xi(s)$ and decisions of the master problem variables $\boldsymbol{x}$,$\boldsymbol{y}$ and $\boldsymbol{a}$ are. As we observed, with the optimal decisions $\boldsymbol{x}$,$\boldsymbol{y}$ and $\boldsymbol{a}$, the objective function value and solution of problem

(3a)–(3f) can be calculated exactly when the uncertain parameters are realized with no need really to solve the optimization problem.

For simplicity of presentation, we provide more details about the classical Benders decomposition (CBD) in Appendix A. Then, the SBD algorithm based on SAA is developed to solve the SMIP problem. We take two types of measures for SBD. One accelerates the convergence rate by adding constraints with more accurate information concerning the subproblems in each iteration (i.e., cut aggregation and feasibility cuts). The other changes the termination criterion by suboptimizing the master problem. We create this variant by observing that the master problems obtain too little information from the subproblems to be worth rigorous optimization.

(1) Cut Aggregation.

In the BD algorithm, the structure of subproblems (i.e., one scenario corresponds to one subproblem) leads to S optimal cuts being generated and added to the master problem for each iteration. In this case, the size of the master problem can be large, which undermines computational tractability. In addition, some redundant cuts may contribute very little information to obtaining the optimal solution to the master problem. Aggregating these cuts does not cause a loss of information while avoiding unnecessary computation. To address this issue, we reformulate the master problem (2a)–(2b) and subproblem (3a)–(3f) to add the optimality cuts that are aggregated. We divide the subproblems based on servers to use as much information as possible from the subproblems without adding too many cuts into the master problem. Since the number of physicians scheduled simultaneously in practice is quite limited, which is far smaller than the number of scenarios, the aggregation greatly reduces the number of cuts and speeds up the solution procedure. We add all the cuts corresponding to the m sub-problems into the master problem and use the commercial solver Gurobi to solve the master problem. Presolve routines are typically incorporated in commercial solvers to reduce the model size before branch and cut procedure. Presolve operations, which are based on logical implications or dual information, include tightening bounds and constraints, removing redundant columns and rows, and fixing variables. We formulate the corresponding subproblem with respect to each server, which contains constraints for all scenarios below.

$$SP^k(\boldsymbol{x},\boldsymbol{a},\boldsymbol{\xi}): min 1/S\Big(\sum_{s=1}^{S}\sum_{i=1}^{n}(c_w^o y_i^o + c_w^e y_i^e)w_i^s x_{ik} + c_k^o o_k^s + c_k^z z_k^s\Big) \tag{4a}$$

$$s.t.\ (1h)-(1l) \tag{4b}$$

In this context, the master problem changes into the following form, with m aggregated optimal cuts added at each iteration. Each aggregated cut is formed by obtaining the average information of all scenarios.

$$min \sum_{k}^{m}\delta_k \tag{5a}$$

$$s.t.(2b) \tag{5b}$$

$$\begin{aligned}\delta_k \geq \frac{1}{S}\Bigg[&\sum_{s\in S}\sum_{j<p\in I} f_{jp}^{ks}\left(q_j^s d_j^s - M\left(2-x_{pk}-x_{jk}\right) - a_p + a_j\right)\\ &+\sum_{s\in S}\sum_{i\in I}\alpha_{ik}^s\left[a_i - \sum_{j\in I, j\neq i} q_j^s d_j^s x_{jk} - M(1-x_{ik})\right]\\ &+\sum_{s\in S}\beta_k^s\left(\sum_{i\in I} q_i^s d_i^s x_{ik} - T\right)\Bigg], \forall k\in K\end{aligned} \tag{5c}$$

(2) Feasibility cuts.

In the algorithm, variables δ_k replace the second-stage variables w_i^s, z_k^s, and o_k^s, resulting in rare information about the removed variables and corresponding operation costs contributing to the master problem. We use feasibility cuts proposed by Zhou et al. [34] to accelerate the algorithm. The feasible region is constrained by the set of feasibility cuts, which aids in a more effective resolution of the master problem. Feasibility cuts are established through solution to the newsvendor problem.

First, the individual cost C_i is defined by the following formula:

$$C_i = \begin{cases} E\left[\left(c_w^{i+1}\right)x_{i+1,k}w_{i+1} + c_k^z z_{ik}\right], i=1,2,\cdots,n-1, k\in K\\ E\left[c_k^o o_k + c_k^z z_{ik}\right], i=n, k\in K\end{cases} \tag{6a}$$

where z_{ik} denotes the idleness of server k after serving the patient assigned to appointment interval i satisfying $\sum_{i\in I} z_{ik} = z_k$.

Proposition 2. *For any given decisions $\boldsymbol{x}$, bounds on the individual cost C_i can be strengthened as follows:*

$$C_i \geq \begin{cases} x_{i+1,k}g_i, i=1,2,\cdots,n-1, k\in K\\ g_i, i=n\end{cases} \tag{6b}$$

where

$$g_i = \begin{cases} \min\limits_{s_i}\mathbb{E}\left[\left(c_w^{i+1}\right)[d_i-s_i]^+ + c_k^z[d_i-s_i]^-\right], i=1,\cdots,n-1, k\in K\\ \min\limits_{s_n}\mathbb{E}\left[c_k^o[d_i-s_i]^+ + c_k^z[d_i-s_i]^-\right], i=n, k\in K\end{cases} \tag{6c}$$

In the above formula, $s_i, i=1,\cdots,n$ are defined as the length of appointment interval scheduled for patient i. Note that $[a]^+ = max\{a,0\}$, and $[a]^- = max\{-a,0\}$.

Proof: For any $i = 1,2,\cdots,n-1$, we have $C_i \geq x_{i+1,k}\mathbb{E}[c_w^{i+1}w_{i+1} + c_k^z z_{ik}]$. We derive the bounds on $\mathbb{E}[c_w^{i+1}w_{i+1} + c_k^z z_{ik}]$. The following equality holds from the definition:

$$\begin{aligned}\mathbb{E}\left[c_w^{i+1}w_{i+1} + c_k^z z_{ik}\right] &= \mathbb{E}\left[c_w^{i+1}[w_i+d_i-s_i]^+ + c_k^z[w_i+d_i-s_i]^-\right]\\ &= \mathbb{E}_{w_i}\left\{\mathbb{E}_{d_i}\left[c_w^{i+1}[w_i+d_i-s_i]^+ + c_k^z[w_i+d_i-s_i]^-|w_i\right]\right\}\end{aligned}$$

Let $s_i^* = \arg\min\limits_{s_i}\min\mathbb{E}_{d_i}[\left(c_w^{i+1}\right)[d_i-s_i]^+ + c_k^z[d_i-s_i]^-]$, where s_i^* is the optimal solution to achieve g_i. We have the following inequality based on the definition:

$$g_i \leq \mathbb{E}_{d_i}[c_w^{i+1}[d_i-\widetilde{s}_i]^+ + c_k^z\left[d_i-\widetilde{s}_i\right]^-], \forall\widetilde{s}_i$$

For any realization of w_i, let $\widetilde{s}_i = s_i - w_i$ and substitute it into the above formula. Hence, we have:

$$g_i \leqslant \mathbb{E}_{d_i}\left[c_w^{i+1}[w_i+d_i-s_i]^+ + c_k^z[w_i+d_i-s_i]^-|w_i\right]$$

Then, the inequality still holds after taking expectation for w_i for the right side of above formula. Hence, we have: $g_i \leq \mathbb{E}\left[c_w^{i+1}w_{i+1} + c_k^z z_{ik}\right]$. That is, g_i is a bound of $\mathbb{E}\left[c_w^{i+1}w_{i+1} + c_k^z z_{ik}\right]$ for any s_i. In a similar way, C_i can also be bounded when $i = n$. The proof is complete.

The following feasibility cuts are added to the master problem to help bound δ_k the operational cost decomposed by each server.

$$\delta_k \geq \sum_{i-1}^{n-1} x_{i+1,k}g_i + g_n, \forall k\in K \tag{6d}$$

Note from Equation (6c) that it can be viewed as a general newsvendor problem with an optimal solution s_i^* and an optimal cost g_i. Thus, we can obtain the optimal solution s_i^* by the critical fractile of a cumulative probability distribution F_i for service time d_i. The corresponding calculation formula is $F_i\left(s_i^*\right) = c_w^{i+1}/(c_w^{i+1} + c_k^z), i = 1,\cdots,n-1$. In this context, the optimal cost g_i can be achieved easily by taking the integral. Notably, we have two types of visits with different unit waiting costs and service times. We treat patients in the same category, assuming all are either virtual patients or office visits. Then, we calculate s_i^* and the optimal cost g_i under the corresponding unit waiting cost and service time. We use the smaller optimal cost g_i for constructing feasibility cuts.

(3) ε- optimal strategy.

Since Benders decomposition comes with significant drawbacks, such as a weak master problem that restricts the dual cuts to the scenario sub-problems after the relaxation step, Geoffrion and Graves [43] considered that the master problem does not need to seek its optimal solution and can be stopped when the resulting feasible solution is better than the current optimal solution. The master problem is used only to

find a feasible solution to the problem. This strategy for solving the master problem suggests that the master problem does not provide a best lower bound on the optimal value of the initial problem (1a)–(1m). For adopting the strategy, first, we reformulate the relaxed master problem by incorporating information pertaining to the scenario sub-problems into the master problem to strengthen its formulation. The stochastic model's deterministic counterpart where all random variables (e.g. patients' show-up status and service time) are replaced by their expectation values is included in the master problem. The master problem is reformulated as follows:

$$min \sum_k^m \delta_k + \sum_k^m \sum_{i=1}^n \left(\left(c_w^o y_i^o + c_w^e y_i^e \right) w_i^v x_{ik} + c_k^o o_k^v + c_k^z z_k^v \right) \tag{7a}$$

$$a_p + w_p^v + M\left(2 - x_{pk} - x_{jk}\right) \geq a_j + w_j^v + \mu_j^q \mu_j^d, \forall p > j \in I, k \in K \tag{7b}$$

$$z_k^v \geq a_i + w_i^v - \sum_{j \in I, j \neq i} \mu_j^q \mu_j^d x_{jk} - M(1 - x_{ik}), \forall i \in I, k \in K \tag{7c}$$

$$o_k^v \geq \sum_{i \in I} \mu_j^q \mu_j^d x_{ik} + z_k^v - T, \forall k \in K \tag{7d}$$

$$(5b)–(5c), (6d) \tag{7e}$$

$$\left(o_k^v, z_k^v\right) \geq 0, \forall k \in K, w_i^v \geq 0, \forall i \in I \tag{7f}$$

where we define virtual waiting time $w_i^v (i = 1, \cdots, n)$ for patient i, virtual idle time $z_k^v (k = 1, \cdots, m)$ and virtual overtime $o_k^v (k = 1, \cdots, m)$ for server k. Objective (7a) includes the information related to the removed variables w_i^s, z_k^s, and o_k^s. Constraints (7b)–(7d) are obtained from Constraints (3b)–(3d), respectively, and provide estimates of the patient waiting time, server idle time, and overtime.

Then, a permissible error $\varepsilon > 0$ is introduced. The algorithm terminates when the current master problem cannot find a feasible solution below $UB - \varepsilon$. The current best feasible solution is the ε- optimal solution to the initial problem (1a)–(1m). We can easily find that this variant must converge to an ε-optimal solution within a limit number of iterations [43]. Under this ε- optimal strategy, the following new Benders cuts are added to the master problem:

$$\begin{aligned} &\sum_{k \in K} \left[\frac{1}{S} \left[\sum_{s \in S} \sum_{j<p \in I} f_{jp}^{ks} \left(q_j^s d_j^s - M\left(2 - x_{pk} - x_{jk}\right) - a_p + a_j \right) \right. \right. \\ &+ \sum_{s \in S} \sum_{i \in I} \alpha_{ik}^s \left[a_i - \sum_{j \in I, j \neq i} q_j^s d_j^s x_{jk} - M(1 - x_{ik}) \right] + \sum_{s \in S} \beta_k^s \left(\sum_{i \in I} q_i^s d_i^s x_{ik} \right. \\ &\left. \left. - T \right) \right] \Bigg] \\ &\leq (UB - \varepsilon) \end{aligned} \tag{8}$$

Applying the above-improved measures to the original Benders master problem, we can convert it to the Formulas (9a)–(9b):

$$min \sum_k^m \delta_k + \sum_k^m \sum_{i=1}^n (c_w^o y_i^o + c_w^e y_i^e) w_i^v x_{ik} + c_k^o o_k^v + c_k^z z_k^v) \tag{9a}$$

$$s.t. \ (7a)–(7f), (8) \tag{9b}$$

Finally, the stabilized Benders decomposition algorithm with enhanced measures can be expressed as follows:

SBD algorithm for the proposed problem

1: Initialization. Set $UB = +\infty$. Select a convergence tolerance parameter $\varepsilon \geq 0$.
2: Solve the master problem (9a)–(9b).
3: **if** the master problem is feasible **then**
4: Record the optimal solutions $(\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{a}^*, \boldsymbol{\delta}^*, \boldsymbol{\mu}^q)$ and the optimal objective value z^{MP}.
5: Stochastic service time and time-dependent no-shows scenario generation. Sample $S = 1000$ *i.i.d.* realizations $(q_1^s, d_1^s), \ldots, (q_n^s, d_n^s), s = 1, \ldots, S$ by given service time distribution and no-shows with mean $\boldsymbol{\mu}^q$.
6: **for** each $k \in [1, m]$ **do**

(*continued on next column*)

(*continued*)

SBD algorithm for the proposed problem

7: Solve the subproblem SP^k (4a)–(4b) with $\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{a}^*$ and record the optimal objective value z_k^{SP}.
6: Get the extreme point.
7: Add optimality Benders cut (5c) to the master problem.
8: **end for**
9: Calculate $\delta = 1/S(\sum_1^S z_s^{SP})$.
10: Set $UB = min\{UB, z^{MP} + \delta - \delta^*\}$.
11: Go to line 2.
12: **else if** the master problem is infeasible
13: terminate. return the current best feasible solution $\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{a}^*$ as the ε- optimal solution and UB as the ε- optimal value.
14: **end if**

The proposed model and method can be extended to solve the appointment scheduling problem by considering other features and situations, such as unpunctuality: earliness or lateness, cancelations, etc. [36,56]. To model patient unpunctuality, the key idea is to calculate the actual arrival time of the patient through $a_i' = a_i + u_i$, where a_i is the appointment time without unpunctuality studied previously, and $u_i \in [-U, U]$. Specifically, a negative u_i represents that patient i arrives earlier than her appointment time; otherwise, a positive u_i represents patient's lateness. U is the bound for earliness/lateness in the unpunctual time window. The main concept behind modeling patient cancellation is to view the patient as a "ghost" patient with 0 service time. As for the Benders decomposition algorithm, since BD is an effective framework for solving a large-scale mixed-integer program, it can be customized to other appointment models with various specific situations. The proposed acceleration techniques can also be extended to fit other models. The cut aggregation method can enhance the algorithm efficiency for multiserver appointment models with separable subproblems. The feasibility cuts can be applied for similar appointment scheduling problems with patient unpunctuality, or cancellations, etc. They can be used to these problems to obtain an individual cost lower bound by adjusting the service time or the appointment time. The ε- optimal strategy is a type of general strategy that can be incorporated into any BD framework.

5. Numerical studies

In this section, first, we verify the efficiency of our proposed algorithm through numerical studies. Then, we investigate the performance of the joint multiserver appointment scheduling model under homogeneous and heterogeneous physicians for virtual and office care, respectively, and we further compare it with the problem formulation as several separated single-server models. Moreover, we study a scenario where the system allows virtual visits to arrive during the specified period by adding time window constraints. Under various time-dependent show-up patterns of office and virtual patients, system operational performances in the above scenario are evaluated. Finally, we analyze the impact of cost weighting coefficients on the system performance measures. We exhibit representative scenarios and elaborate upon some necessary observations in each subsection.

5.1. Input data

In this subsection, we interpret the scenario generation before evaluating the performance of the algorithm and system. Our dataset consists of the real service time data records and simulated no-show data of office and virtual visits. We first collected the service time data of both types of patients who visited the Department of Endocrine at Shanghai Sixth People's Hospital from 3 June 2020 to 31 December 2020. Shanghai Sixth People's Hospital is a comprehensive hospital with 33 clinical departments and 9 technical departments. It has its own Internet hospital and provides both office and virtual medical services.

Table 2
Running time performance of the SBD algorithm.

Scale	m	n	Average Time		STD		MIN		MAX	
			SBD	Gurobi	SBD	Gurobi	SBD	Gurobi	SBD	Gurobi
Small	20	1	67.19	3.88	9.06	0.33	50.99	3.61	71.38	4.42
	12	2	45.92	1296.33	8.97	1032.69	36.1	286.76	53.34	2902.42
	10	3	56.58	2154.04	0.28	1189.03	56.38	670.69	57.06	3674.22
	14	2	47.91	4057.99	0.15	3406.32	47.75	1058.66	48.11	9432.24
	15	2	64.57	5213.09	13.96	5786.01	54.08	1348.05	79.99	15307.78
Medium	20	2	92.65	>18000	0.32	–	92.14	16939.76	92.95	>18000
	25	3	228.65	>18000	45.24	–	207.87	>18000	309.57	>18000
	20	5	380.40	>18000	75.06	–	345.5	>18000	514.66	>18000
	30	3	353.12	>18000	79.26	–	294.19	>18000	441.4	>18000
	40	2	268.26	>18000	60.29	–	223.67	>18000	335.4	>18000
Large	50	2	591.21	>18000	119.02	–	536.28	>18000	804.1	>18000
	40	3	920.02	>18000	206.22	–	768.09	>18000	1146.59	>18000
	60	2	513.01	>18000	0.94	–	512.25	>18000	514.58	>18000
	40	4	680.50	>18000	2.07	–	678.46	>18000	683.87	>18000
	50	5	1702.95	>18000	353.47	–	1312.94	>18000	1964.38	>18000

Specifically, the data shows that office service time ranges from 0 to 26 min, with the average being 9.53 min. The virtual service time varies between 0 and 36 min, with an average of 14.79 min. The 95 % confidence intervals for the means of office and virtual service times are [7.82,11.24] and [12.93,16.65] minutes, respectively. All the service time data used in numerical studies were generated based on the above real data characteristics.

Kong et al. [4] analyzed two datasets and found that patient attendance behavior is significantly impacted by the time of day. Potential reasons may be people's different life schedules, patient populations, and the culture of attitude toward time. We generate time-dependent no-show scenarios to incorporate the no-show's property of time-dependent. First, we define time-dependent no-show scenarios as follows. For each patient $i = 1, \cdots, n$, the random variable of the no-show status $Q \in \{0, 1\}$ follows the Bernoulli distribution with a parameter μ_i^q. That is, $Q \sim B(\mu_i^q)$. A scenario s for patient i contains her no-show information q_i(e.g., $q_i = 1$ represents show, and $q_i = 0$ represents no-show). q_i is a realization of Q, and the index set of scenarios is denoted by $S := \{s\}$. Because the show status of a real appointment is known at the time it was scheduled, we cannot know if the same appointment would have had the same show status if it is scheduled at another time. In this context, we use two piecewise linear functions to emulate the show-up pattern of office and virtual visits, similar to the observations in Kong et al. [4]. We consider the piecewise linear function for the office (virtual) patient time-dependent show-up pattern that has R segments with corresponding breakpoints $t_r^o(t_r^e), r = 0, \cdots, R$ and function values $pr_r^o(pr_r^e), r = 0, \cdots, R$. We can obtain its show-up probability $\mu_i^q, i = 1, \cdots, n$ for any given patient's scheduled arrival time $a_i, i = 1, \cdots, n$. In the proposed improved benders decomposition algorithm, based on the framework of the algorithm, the master problem is solved first, and the patient's scheduled arrival time $a_i, i = 1, \cdots, n$ is obtained. According to the piecewise linear function, the show-up probability μ_i^q corresponding to a_i is obtained. Based on $\mu_i^q, i = 1, \cdots, n$, a set of scenarios is generated following Bernoulli distribution. Then the solution of the decision variables of the master problem and these generated scenarios are substituted into the sub-problems to solve the sub-problems.

Service times and time-dependent no-shows for office and virtual visits with sample size $S = 1000$ are realized based on the above procedure. By conducting a survey of visits' show-up preference at different times if they have an virtual or office appointment, we estimate an increasing shape for virtual patient show-up pattern with show rate $pr_0^e = 0.45$ at time 0, $pr_1^e = 0.75$ at time $T/2$, and $pr_2^e = 0.95$ at time T, and a decreasing shape for office patient show-up pattern with show rate $pr_0^o = 0.9$ at time 0, $pr_1^o = 0.7$ at time $T/2$, and $pr_2^o = 0.4$ at time T. Through a discussion with physicians in practice, we determined the unit waiting time costs for virtual and office patients, unit costs for outpatient physician idle time and overtime to be $c_w^e = 1$, $c_w^o = 0.8$, $c_k^z = 1$, and $c_k^o = 1.5$, respectively. These settings represent our base case on which real-world inspired numerical studies are conducted. It can reflect our field study in the Department of Endocrine at Shanghai Sixth People's Hospital of China.

5.2. Algorithm performance

To study the efficiency of the algorithm, we generate many instances based on the real service data and no-show simulated data. We consider different numbers of servers and patients, with 25 % virtual and 75 % office visits. Fifteen pairs of (n, m) with five instances of each with $c_k^o = 0.7, 0.9, 1.1, 1.3, 1.5$ are studied. The length of the time limit is set at $T = (nr_o\mu_{ok}^d + nr_e\mu_{ek}^d)/m$, which coordinates with the number of servers and patients. Then, all the results from our SBD method are compared with the benchmark. Gurobi is used to solve the SAA-based stochastic programming problem as the benchmark. We calculate the minimum, average, and maximum running times of solving the problem. Table 2 summarizes the comparison results. The numerical instances are conducted on a PC with an Intel Core i7-10700 CPU 2.90 GHz and 16 GB memory. The master problem and subproblems of the proposed SBD algorithm are solved by calling Gurobi 9.0.2 on Python 3.7 with the default settings.

We first investigate the computation speed performance of the proposed SBD algorithm. We set the computational time limit to 18,000 s (e.g. five hours). The performance results of SBD's computation speed are summarized in Table 2. The results indicate that the stabilized BD significantly reduces the minimum, average, and maximum computational times. For small-size problems, our method can obtain ε- optimal solutions within 2 min, while Gurobi consumes nearly 1.5 h computation time on average. As the scale of the problem grows, running time becomes longer for both our method and the benchmark. The running time of the SBD algorithm grows slowly, while that of Gurobi grows rapidly with increasing instance size. The SBD algorithm performs much more stably. For instances of medium and large scales, the benchmark cannot find the optimal solution within the time limit (e.g. 18000 s). Our algorithm can solve all problem instances within 2000 s. Typically, our algorithm requires a surprisingly small number of iterations for convergence, even with very small values of the optimality tolerance ε. These findings indicate the stabilized BD algorithm is indeed much more efficient than the solver Gurobi.

Since Gurobi cannot output the optimal solution after running for 5 h, it is not necessary to limit the calculation time to 5 h to compare the quality of the solution. To make a fair comparison and save computing time, we limit the computing time to 1000 s, and set the number of scenarios to be 500. To study the proposed SBD's performance in

Table 3
Comparison of solution quality between SBD, CBD, and Gurobi on small-scale instances.

n	m	c_k^o	*Obj*		
			SBD	CBD	Gurobi
20	1	0.7	100.0	1368.1	92.3
		1.5	100.0	1396.3	92.3
12	2	0.7	40.0	487.9	35.9
		1.5	39.6	527.4	38.0
10	3	0.7	27.4	359.1	25.5
		1.5	24.7	397.5	28.5
13	2	0.7	45.8	529.9	42.7
		1.5	50.0	571.6	43.1
14	2	0.7	54.1	557.7	51.3
		1.5	52.3	603.1	50.5
15	2	0.7	60.7	685.7	55.5
		1.5	57.4	734.0	56.2

Table 4
Comparison of solution quality between SBD, CBD, and Gurobi on medium-scale instances.

n	m	c_k^o	*Obj*		
			SBD	CBD	Gurobi
20	2	0.7	91.9	1243.0	104.0
		1.5	105.9	1308.0	113.0
20	3	0.7	76.1	1201.4	127.8
		1.5	80.0	1278.7	136.0
25	3	0.7	99.3	2000.8	189.9
		1.5	106.7	2098.5	259.7
20	5	0.7	65.0	1168.1	140.5
		1.5	89.4	1255.2	279.9
30	3	0.7	131.7	2763.8	459.8
		1.5	167.7	2880.8	496.8
40	2	0.7	298.9	5224.5	493.0
		1.5	276.9	5355.9	497.6

Table 5
Comparison of solution quality between SBD, CBD, and Gurobi on large-scale instances.

n	m	c_k^o	*Obj*		
			SBD	CBD	Gurobi
50	2	0.7	367.4	7857.6	19367.2
		1.5	335.8	8021.7	1695.0
40	3	0.7	226.8	4968.6	986.4
		1.5	205.7	5124.6	1068.1
50	3	0.7	324.9	7539.4	1248.1
		1.5	286.6	7734.2	1322.5
60	2	0.7	501.3	11196.7	--
		1.5	488.2	11391.4	--
40	4	0.7	221.6	4840.6	785.4
		1.5	164.6	5009.0	864.2
50	5	0.7	273.3	7284.9	--
		1.5	247.8	7504.2	--

solution quality, we compare the associated overall system cost for best feasible solutions obtained by Gurobi, CBD and SBD after running 1000 s. The comparison results of small-, medium-, large-scale instances are summarized in Tables 3–5. Column "Gurobi" reports the solution by Gurobi after it reaches the time limit or out of memory. Compared with the classical Benders decomposition algorithm, the solution quality of the proposed SBD is significantly and consistently superior. This implies that the acceleration techniques take effects as CBD coverages rather slowly through iteration process and yields poor solutions. In comparison to Gurobi, for small-scale instances, we compute the gap between the solution value generated by the SBD (Obj_{SBD}) and the solution value generated by the Gurobi (Obj_{Gurobi}) as follows: $GAP_{SBD-Gurobi} = 100^*(Obj_{SBD} - Obj_{Gurobi})/Obj_{Gurobi}$. The SBD can generate near equally good solutions with an overall average gap of 5.85 % over all 12 instances. As the problem scale rises with an increasing number of patients and servers, the solutions generated by SBD are always better to those generated by Gurobi. And the improvement is more pronounced for large-scale instances than for medium-scale instances.

5.3. The performance of multiserver time window allowance schedules

5.3.1. The performance of multiserver model compared with single-server model

In this subsection, we aim to investigate, in terms of the performance of virtual and office patient scheduling systems, whether the joint multiserver model (MSAS) has advantages over the separated single-server model (SSAS) in previous studies [4,15,29]. To verify the effectiveness of the MSAS model, two sets of numerical studies have been implemented. One set is for homogeneous physicians with the same properties. The other set is for heterogeneous physicians with different service times, overtime and idle time costs. Separated single-server models are simulated for multiple servers by repeating single-server results multiple times and are compared to the MSAS with the same number of servers. We first present the analysis for homogeneous physicians in Appendix B, while each physician's behavior varies depending on the physician's habits, treatment methods, and patience in practice. Therefore, more examples are generated with heterogeneous servers based on our dataset in Section 5.1. In this situation, the service times of different physicians treating office and virtual visits are modelled by nonidentical normal distributions differentiated by their means $\mu_{ok}^d(\mu_{ek}^d)$, $k = 1, \cdots, m$ and standard deviations $\sigma_{ok}^d(\sigma_{ek}^d), k = 1, \cdots, m$ within the actual office and virtual service intervals. Different server parameter configurations are listed in Table 6. We examine a two-sever virtual and office patient scheduling problem with 40 visits (40 % virtual, 60 % office patients) and $T = (nr_o\sum_{k=1}^m \mu_{ok}^d/m + nr_e\sum_{k=1}^m \mu_{ek}^d/m)/m$. Seven experiments are implemented to study the impact of physician heterogeneity in service efficiency, service efficiency stability, overtime costs, and idle costs on the performances of the SSAS and the MSAS.

The comparison results of the system performances using the SSAS and the MSAS for each case are summarized in Table 7. Specifically, Cases 1 and 2 assume two heterogeneous physicians with different service efficiencies, which is achieved by modelling their service times as

Table 6
Parameter configurations for heterogeneous servers.

No. of Case	1		2		3		4		5		6		7		
No. of Server	1	2	1	2	1	2	1	2	1	2	1	2	1	2	3
μ_{ok}^d	8	8.5	8	9	8	8	8	8	8	8	8	8	8	8	8
μ_{ek}^d	13	13.5	13	14	13	13	13	13	13	13	13	13	13	13	13
Cov	0.2	0.2	0.2	0.2	0.2	0.4	0.2	0.8	0.2	0.2	0.2	0.2	0.2	0.4	0.8
σ_{ok}^d	1.6	1.7	1.6	1.8	1.6	3.2	1.6	6.4	1.6	1.6	1.6	1.6	1.6	3.2	6.4
σ_{ek}^d	2.6	2.7	2.6	2.8	2.6	5.2	2.6	10.4	2.6	2.6	2.6	2.6	2.6	5.2	10.4
c_k^o	1	1	1	1	1	1	1	1	1	1.5	1	5	1	1	1
c_k^z	1	1	1	1	1	1	1	1	1	1.5	1	5	1	1	1

Table 7
Objective value comparison results of MSAS and SSAS under heterogeneous servers.

No. of Case	1	2	3	4	5	6	7
SSAS	270.55	307.31	353.14	457.65	334.54	441.59	679.36
MSAS	236.52	237.04	321.2	411.89	260.4	321.68	524.18
Improvement	12.58 %	22.87 %	9.04 %	10.00 %	22.16 %	27.15 %	22.84 %

Table 8
Parameters settings of office and virtual patient show-up patterns.

Patient type	Show-up Pattern	$t_r^o(t_r^e)$	0	$T/2$	T
Office patient	I	pr_r^o	0.4	0.7	0.9
	II		0.9	0.7	0.4
	III		0.4	0.9	0.4
	IV		0.9	0.4	0.9
Virtual patient	I	pr_r^e	0.45	0.75	0.95
	II		0.95	0.75	0.45
	III		0.45	0.95	0.45
	IV		0.95	0.45	0.95

nonidentical distributions with various means. The difference in efficiency between the two servers in Case 2 is greater than that in Case 1. We keep the other parameters the same between Cases 1 and 2. The results indicate that the MSAS model decreases total costs by 12.58 % and 22.87 % for Cases 1 and 2, respectively. The percentage decrease in cost for Case 2 by using the MSAS model is 10.29 % greater than that for Case 1. In addition, we consider two heterogeneous physicians differentiated by their standard deviation in service times (e.g. Cases 3 and 4), unit overtime, and idle costs (e.g. Cases 5 and 6). The remaining server parameters remain the same. The differences between the two servers in Case 4 (Case 6) are more significant than those in Case 3 (Case 5). We also explore how the number of heterogeneous physicians affects the performance of the MSAS model and conduct a comparative analysis of them. The comparison results of Cases 3, 4, and 7 (see Table 7) indicate that more heterogeneous physicians will improve the cost reduction effect of the MSAS, resulting in a larger performance gap between it and the SSAS.

We can obtain the following conclusions from the results. In the virtual and office patient scheduling system, the multiple server model has a limited impact on cost reduction when the servers are

Table 10
The detailed comparison results under show-up pattern Combo 2.

Combo	Performance	Schedule 1	Schedule 2	Schedule 3	Schedule 4
2	*Obj*	215.86	201.21	273.37	211.22
	T_w^o	158.03	92.84	113.94	118.76
	T_w^e	34.12	83.38	65.76	67.83
	T_w	192.15	176.21	179.70	186.59
	T_o	0.05	0.00	2.86	0.00
	T_i	23.66	25.00	90.81	24.63

Table 9
The comparison results under 16 show-up pattern combinations.

Combo	Show-up Pattern		Schedule 1	Schedule 2	Schedule 3	Schedule 4
	Virtual patient	Office patient				
1	I	I	241.60	199.33*	298.78	215.47
2	II		215.86	201.21*	273.37	211.22
3	III		238.07	207.91*	278.27	226.93
4	IV		232.28	174.20*	287.17	197.62
5	I	II	225.41	181.68	171.88*	181.70
6	II		190.10	183.21	171.44	169.98*
7	III		222.24	171.67	168.97*	172.63
8	IV		216.99	177.50	177.46	177.01*
9	I	III	270.07	215.26*	268.42	232.43
10	II		197.12*	215.45	248.18	200.57
11	III		258.70	215.78*	251.80	252.60
12	IV		242.52	195.31*	260.43	210.96
13	I	IV	204.01	195.60	202.73	192.17*
14	II		159.53*	185.44	193.02	169.70
15	III		194.58	188.37*	193.00	188.47
16	IV		179.48	162.68*	201.23	166.40

* represents the lowest cost achieved under the optimal schedule.

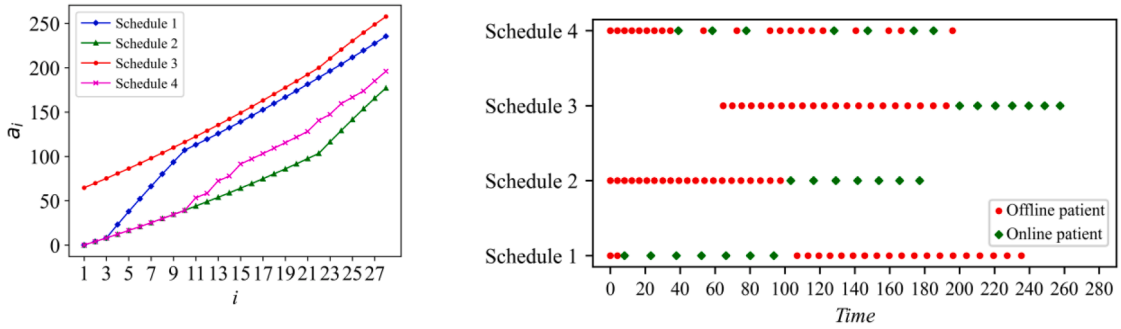


Fig. 1. Scheduling results of four types of time window allowance schedules under Combo 2.

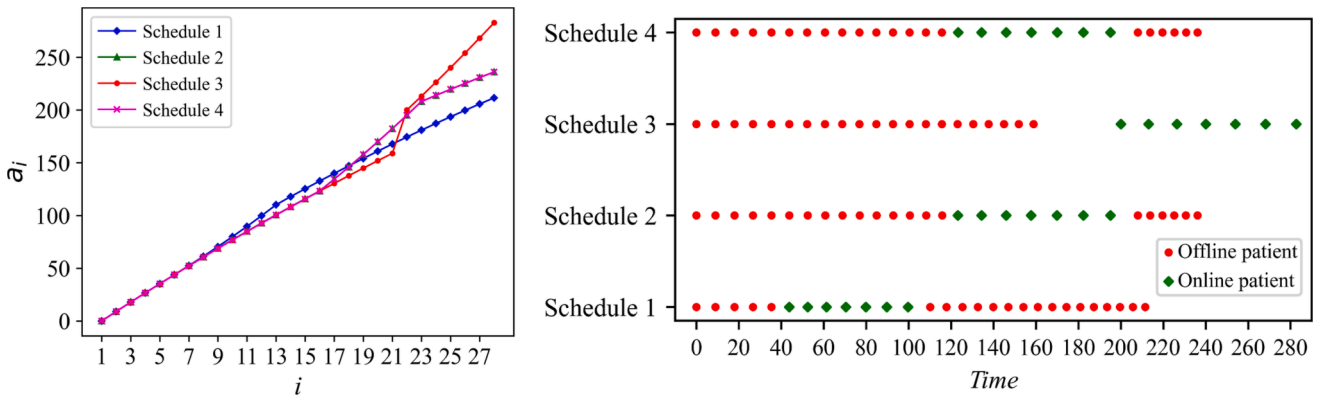


Fig. 2. Scheduling results of four types of time window allowance schedules under Combo 5.

Table 11
The detailed comparison results under show-up pattern Combo 5.

Com	Performance	Schedule 1	Schedule 2	Schedule 3	Schedule 4
5	*Obj*	225.41	181.68	171.88	181.70
	T_w^o	163.49	99.10	84.38	99.57
	T_w^e	35.89	56.77	28.63	56.38
	T_w	199.38	155.88	113.01	155.95
	T_o	0.00	0.05	6.02	0.05
	T_i	26.04	25.76	52.85	25.70

Table 12
The results for different ratio of virtual visits.

r_e	r_o	MTWAS–[0, T/2]	MTWAS–[T/2, T]	MTWAS–[0, T]
0.25	0.75	334.3	225.7*	347.1
0.35	0.65	358.7	254.9*	375.7
0.5	0.5	438.3	277.5*	421.8
0.65	0.35	490	408.7	383.3*
0.75	0.25	702.6	589.2	445*

* represents the lowest cost achieved under the optimal schedule.

homogeneous. By keeping other server configurations fixed, for more heterogeneous servers with more significant differences in terms of service efficiency, service efficiency stability, overtime costs, and idle costs, the benefits brought by the MSAS will be greater than those brought by the SSAS. Thus, the managerial insight that we provide to hospital practitioners is to choose the appropriate scheduling model based on the differentiation degree between the physicians being scheduled. Specifically, when there is no significant difference between the physicians being scheduled, the SSAS is recommended to obtain similar results with less computation time if fast scheduling is needed. In contrast, when there are significant differences between the physicians being scheduled, the MSAS should be selected to realize considerable cost savings from less patient waiting, physician idle and overtime costs.

5.3.2. Impact of appointment time windows under time-dependent show-up patterns

In this subsection, we consider a single server and 28 patients, with 25 % virtual and 75 % office visits. The length of the time limit is set at $T = 300$. To avoid the impact of the cost parameters on the result, we set the cost ratio as c_w^o: c_w^e: $c_k^z : c_k^o = 1{:}1{:}1{:}1$. To investigate how different specified appointment time windows for virtual visits under various

Table 13
The results for different combinations of cost coefficients.

Case	c_w^o	c_w^e	c_k^z	c_k^o	*Obj*	w^o	w^e	w	t_i	t_o	t_i^o
1	0.1	1	1	1	99.46	164.48	58.61	223.10	24.27	0.12	24.39
2	0.5	1	1	1	140.87	74.56	77.56	152.12	25.15	0.87	26.02
3	2	1	1	1	226.88	54.28	90.51	144.79	24.44	3.38	27.82
4	10	1	1	1	626.64	50.63	90.92	141.56	24.62	4.59	29.21
5	1	0.1	1	1	99.37	64.33	83.96	148.30	24.82	1.82	26.64
6	1	0.5	1	1	127.59	54.70	90.54	145.24	24.31	3.31	27.63
7	1	2	1	1	302.03	134.56	71.56	206.12	23.94	0.41	24.35
8	1	10	1	1	794.77	144.33	62.59	206.93	24.31	0.19	24.50
9	1	1	0.5	1	161.38	63.57	83.47	147.04	24.98	1.84	26.82
10	1	1	2	1	199.67	64.07	84.04	148.11	24.86	1.83	26.70
11	1	1	10	1	444.11	138.15	63.85	202.00	24.20	0.14	24.34
12	1	1	15	1	620.80	242.16	71.35	313.50	20.46	0.45	20.90
13	1	1	1	0.5	235.58	92.77	92.10	184.87	20.02	3.08	23.10
14	1	1	1	10	191.23	64.08	83.98	148.06	24.86	1.83	26.69
15	1	1	1	15	237.56	152.84	58.29	211.13	24.38	0.14	24.51
16	1	1	1	20	238.24	152.84	58.29	211.13	24.38	0.14	24.51

Table B1
Objective value comparison results of MSAS and SSAS under homogeneous servers.

m	2			3			4			5			6		
n	20	40	60	20	40	60	20	40	60	20	40	60	20	40	60
MSAS	87.87	273.01	411.36	76.61	213.65	392.83	72.65	213.35	359.57	70.64	187.05	309.13	88.57	164.38	293.02
SSAS	83.1	263.64	457.02	77.59	229.82	395.46	62.27	166.19	413.22	72.93	170.5	349.42	49.91	153.46	249.29

patient show-up patterns impact the performance of the integrated virtual and office appointment scheduling system, we consider four types of MTWAS with different specified appointment time windows for virtual patients: Schedule 1 (MTWAS–[0, T/3], allowing virtual visits to be scheduled during period [0, T/3] only), Schedule 2 (MTWAS–[T/3, 2 T/3], allowing virtual patients to be scheduled during period [T/3, 2 T/3] only), Schedule 3 (MTWAS–[2 T/3, T], allowing virtual visits to be scheduled during period [2 T/3, T] only) and Schedule 4 (MTWAS–[0, T], allowing virtual patients to be scheduled during the entire working session). Furthermore, four show-up patterns for both office and virtual visits are considered: Pattern I (increasing), Pattern II (decreasing), Pattern III (first increasing then decreasing), and Pattern IV (first decreasing then increasing). They are represented by two-piece piecewise linear functions, and the associated breakpoint parameters and function values are set in Table 8. Therefore, there are 16 combinations of show-up patterns (hereafter referred to as Combo). In this study, we compare the solution performance of the proposed time window allowance Schedules 1–3 with Schedule 4 extensively discussed in previous studies [15,28,29,36]. Numerical instances for the 16 show-up pattern Combos under the four types of MTWAS are generated according to our dataset in Section 5.1.

The total cost results for the 16 Combos of show-up patterns are shown in Table 9. The results show significant differences in the performance of different time window allowance schedules. Among them, when the show-up pattern of office visits is an increasing shape, the costs under the four schedules always satisfy Schedule 2 $<$ Schedule 4 $<$ Schedule 1 $<$ Schedule 3; that is, Schedule 2 can obtain the lowest cost, and Schedule 3 is the worst decision. Schedule 1 shows the worst performance when the show-up pattern of office patients is decreasing, regardless of the virtual visits' show-up pattern. Schedule 3 is most beneficial to the overall service system under Combos 5 and 7. Under Combos 6 and 8, there are no significant performance differences between Schedules 4 and 3 to be best decisions. Notably, when the show-up pattern for office patients is first increasing then decreasing and the Combos are 9, 11, and 12, the performance of Schedule 2 is significantly better than that of the other three types of time window allowance schedules.

For presentation brevity and illustration purposes, we present the results for Combos 2 and 5 to provide further explanations as two examples. Other cases of office and virtual show-up pattern combinations can be analyzed in a similar way. Fig. 1 reports the appointment schedules of the four types of time window allowance schedules under Combo 2. When the show-up patterns of the two types of visits is Combo 2, the appointment schedule under Schedule 1 starts with two office patients and consecutive virtual visits, followed by the remaining consecutive office visits. The arrival time of the two types of patients is scheduled later because virtual visits with longer service times are arranged in the front where the corresponding show probability is higher, while office patients are arranged in the latter position with the higher show probability. Conversely, under Schedule 2, the better result is achieved when office visits are scheduled first with office visits' lower show probability until all office visits are allocated. Then, subsequent virtual patients start to be arranged in the period with a relatively lower show-up probability. Consequently, the arrival times of all visits are scheduled earlier. For Schedule 3, its schedule result is similar to Schedule 2; all office visits will arrive before virtual visits. The difference is that the first office patient is scheduled to arrive much later at time 64.57 min, and the entire schedule is shifted later. As the increasing show-up pattern has a low show probability at the front of the timeline, if patients are scheduled at the front, the schedule is very tight, and the arrival interval between two consecutive patients is even smaller. In addition, given the time window constraints of virtual visits, the arrival time between the last office patient and the first virtual patient will definitely be far apart, resulting in a significant increase in idle time. Therefore, scheduling the first office patient at a later time reduces the risk of an over increasing idle time. For Schedule 4, the scheduled arrival times between patients of the same category present a more decentralized manner, with alternating scheduling between office and virtual visits.

Table 10 demonstrates the comparison results of each detailed cost in terms of system performance measures (e.g. virtual and office patient waiting costs T_w^e, T_w^o, server overtime costs T_o, idle time costs T_i, and total costs *Obj*) for different schedules. As shown in Table 10, Schedule 2 performs significantly better than the other three types of time window allowance schedules. Compared to Schedule 1, Schedule 2, which places a larger number of office visits in front of a smaller number of virtual visits, makes the reduction in waiting time for office patients more than the increase in waiting time for virtual visits. Consequently, total waiting time is reduced because patient service times are uncertain, and the First In First Out rule requires the previous service to be completed before the next service can begin, creating waiting times based on schedules. Accumulation and propagation of this stochastic uncertainty over time increase the risk that patients who arrive later in the overall schedule will wait longer. Therefore, with overtime and idle time being nearly equal, Schedule 2 is better than Schedule 1. Total patient waiting time under Schedule 3 is nearly the same as that under Schedule 2. However, idle time under Schedule 3 is greater than that under Schedule 2, resulting in a significant increase in total costs as well. Schedule 4 performs nearly the same as Schedule 1.

Notably, for all office and virtual show-up pattern combinations, Combo 5 represents the scenario that currently exists in the Department of Endocrine at Shanghai Sixth People's Hospital. As demonstrated in Fig. 2, the appointment schedule under Schedule 1 starts with five office visits and consecutive virtual patients, followed by remaining consecutive office visits. Again, it's still not a good decision to put most office patients in the back, resulting in a long waiting time. Schedule 2 and Schedule 4 have very similar schedules, and corresponding performance is almost the same. The appointment schedule under the two schedules starts with 15 consecutive office visits and consecutive virtual visits, followed by remaining subsequent office patients.

The performance evaluation corresponding to the above schedules is as shown in Table 11. In this case, Schedule 3 is the best decision, so we recommend scheduling virtual visits in the period [2 T/3, T]. Since we did not consider all potential scheduling periods, we only suggest scheduling office visits earlier and virtual patients later with their higher show-up probabilities. Scheduling the majority of office visits at the front results in a significant reduction in office patient waiting time. In addition, there is a wide interval between the last office and the first virtual visits due to the [2 T/3, T] time window constraint. This interval absorbs exactly office visits' service deferral accumulated by service times uncertainty without significantly increasing idle time. The waiting time for virtual visits is also reduced. Thus, the total cost is minimized.

From the above observations, we can conclude the following. Nearly in all the cases of integrated virtual and office appointment scheduling, the proposed multiserver time window allowance schedule has superior performance than traditional schedule without time windows in previous research. The performances of Schedules 1 and 2 are not significantly affected by the show-up patterns of virtual and office patients. In most cases, Schedule 1 performs poorly, whereas Schedule 2 results in the best or close to the best results. The performance of Schedule 3 is much more sensitive to the show-up patterns of virtual and office visits. Schedule 3 performs the worst with an increasing show-up pattern of office visits. Schedule 3 performs nearly the worst when office visits' show-up pattern is either first increasing then decreasing or first decreasing then increasing. However, when office patients' show-up pattern is decreasing, Schedule 3 is the best or near-best decision. These provide managers with specified appointment period decision guidelines when scheduling hybrid office and virtual visits with different show-up patterns. We prefer to suggest to schedule virtual visits in the period [T/3, 2 T/3] under all show-up pattern combinations because of its robustness for best or near-best results. Scheduling virtual visits in the period [2 T/3, T] with a decreasing office visits' show-up pattern is

recommended to achieve better system operational efficiency.

5.3.3. Impact of the ratio of virtual visits

We conduct experiments to investigate the impact of the ratio of virtual visits (RoV) on the time windows scheduling decisions and the overall cost. To generate the instances, we change the value of RoV from 0.25 to 0.75, while the other parameters are kept fixed. Table 12 presents the results for different ratio of virtual visits. When the RoV is not greater than a certain value (such as 0.5), the best schedule always arranges virtual visits at [T/2, T], which is not sensitive to the patient ratio. As the RoV increases, the overall costs pertaining to both [0, T/2] and [T/2, T] time window schedules increase. When the RoV continues to increase until it exceeds a certain threshold (such as 0.5), the best scheduling decision has changed. The schedule without time window generates the lowest overall cost. From these results, we can provide the practitioners the following suggestions: the length and position of the time window adopted can be adjusted according to the RoV. The lowest cost can be obtained by scheduling virtual visits within the time window with appropriate length and position. When the RoV reaches a certain level, it is better to use the schedule without time window restrictions.

5.3.4. Impact of cost weighting coefficients

In this subsection, we analyze the impact of the cost weighting coefficients. We vary the c values, where (c_w^o, c_w^e, c_k^z, and c_k^o) are the objective function coefficients of office patient waiting time, virtual patient waiting time, physicians' idle time, and overtime, respectively. We test 16 sets of different values of c on an instance set. Without loss of generality, we consider a single server and 25 patients, with 40 % virtual and 60 % office visits (a common scenario in the Department of Endocrine at Shanghai Sixth People's Hospital). The time limit is set at $T = 260$. Other parameters are the same as in Section 5.1. We report the sensitivity results in Table 13, which clearly demonstrates the results between office patient waiting time w^o, virtual patient waiting time w^e, server idle time t_i, and overtime t_o with reference to c values. When only c_w^o is increased, both office patient and total waiting times decrease significantly. However, virtual patient waiting time and physician overtime increase. When only c_w^e is increased, both office patient and total waiting times increase significantly. However, virtual patient waiting time and physician overtime present decreasing trends. When either c_w^o or c_w^e changes, the server idle time stays approximately the same. Server idle time and overtime values decline as the c_k^z and c_k^o values increase individually while fixing the remaining three parameters. In addition, the impact of c_k^z and c_k^o on their respective performance measures is not as substantial as that of c_w^o and c_w^e on the two types of patient waiting time. These results imply that minimizing the waiting time for one type of patient has no benefit on minimizing that for the other type of patient, meaning that office and virtual patient waiting times are conflicting measures. Moreover, the impact of c_w^o on office patient waiting time is much more apparent than that of c_w^e on virtual patient waiting time, which leads to the total waiting time and the office patient waiting time changing in the same direction. In addition, the results demonstrate that the total waiting time of visits and the sum of server overtime and idle time t_i^o are changing in the opposite direction. The model assigns appointment times in wider time intervals to make the total patient waiting time shorter, which leads to an increase in the sum of overtime and idle time values. This result indicates a trade-off between total patient waiting time and the sum of server overtime and idle time.

6. Conclusion

In this study, we investigate the appointment scheduling problem of virtual service with proposing MTWAS to manage virtual visits to receive services at a specific period considering stochastic service time and time-dependent no-show. Our modelling approach addresses the challenges in modelling several aspects simultaneously, whereas previous research focused only on parts of those factors. The problem jointly optimizes the decisions of assigning visits to multiple physicians and the patient appointment time. The objective is to minimize the sum of patients' waiting time, servers' idle time, and overtime costs. To solve the problem, we formulate it as a stochastic mixed-integer program. Several acceleration measures are taken into the Benders decomposition such that a stabilized Benders decomposition algorithm with SAA approach is proposed to obtain an ε-optimal solution. Finally, the numerical results demonstrate that our proposed algorithm outperforms the solver Gurobi. More computational studies are conducted to provide managers with managerial insights when scheduling multiple physicians for treating virtual and office visits in practice. In most cases of integrated virtual and office patients' scheduling, the superiority of our proposed multiserver time window allowance schedule solutions are more prominent than previous solutions. First, compared with separate scheduling with several single-server models, it is suggested to apply a joint multiple servers' system to schedule patients when there are significant differences among physicians for treating virtual and office patients. Second, when scheduling with hybrid office and virtual visits, under different show-up pattern combinations, determining the appropriate specified appointment period for virtual visits will significantly affect system performance.

Future work can be extended in several directions. First, a capacity allocation problem and appointment scheduling problem can be jointly optimized because decisions in both stages will affect each other. Second, more complicated patient behaviors or patient types can be investigated in our model, e.g. patient unpunctuality, cancellation, office walk-in patients, and office to virtual revisits. Third, Internet hospitals have obvious advantages in reducing the risk of cross-infection in the COVID-19 epidemic era and have no transportation costs. Hence, the benefits brought by Internet hospitals can be considered in the problem.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The first two authors were supported by the National Natural Science Foundation of China (Grant No. 72171144; 72293585; 71871138) and Shanghai Jiao Tong University Medical and Industrial Cross Project (YG2022QN004); the third author was supported by the China Europe International Business School (CEIBS) Healthcare Research Fund. Meanwhile, the authors would like to sincerely appreciate the editors and reviewers for their careful work and valuable suggestions, which significantly improved the quality of this paper.

Appendix A:. The supplementary of Benders decomposition

The optimality cuts are derived to add to the master problem. Let f_{jp}^{ks} be the optimal value of dual variables corresponding to the constraints (3b) of $SP^s(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{\xi})$, and α_{ik}^s and β_k^s be the optimal value of dual variables corresponding to the constraints (3c) and (3d) of $SP^s(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{\xi})$. We formulate the

subproblem $SP^s(\boldsymbol{x},\boldsymbol{a},\boldsymbol{\xi})$ in its dual form as:

$$max \sum_{k\in K}\sum_{j<p\in I} f_{jp}^{ks}\left(q_j^s d_j^s - M(2-x_{pk}-x_{jk}) - a_p + a_j\right) + \sum_{k\in K}\sum_{i\in I}\alpha_{ik}^s[a_i - \sum_{j\in I, j\neq i} q_j^s d_j^s x_{jk} - M(1-x_{ik})] + \sum_{k\in K}\beta_k^s(\sum_{i\in I} q_i^s d_i^s x_{ik} - T) \tag{A.1a}$$

$$s.t. -\sum_{k=1}^{m}\sum_{p=2}^{n} f_{1p}^{ks} - \sum_{k=1}^{m}\alpha_{1k}^s \le c_w^o y_1^o + c_w^e y_1^e \tag{A.1b}$$

$$-\sum_{k=1}^{m}\sum_{p=i+1}^{n} f_{ip}^{ks} + \sum_{k=1}^{m}\sum_{j=1}^{i-1} f_{ji}^{ks} - \sum_{k=1}^{m}\alpha_{ik}^s \le c_w^o y_i^o + c_w^e y_i^e, \forall i \in I, 1 < i < n \tag{A.1c}$$

$$\sum_{k=1}^{m}\sum_{j=1}^{n-1} f_{jn}^{ks} - \sum_{k=1}^{m}\alpha_{nk}^s \le c_w^o y_n^o + c_w^e y_n^e \tag{A.1d}$$

$$\sum_{i=1}^{n}\alpha_{ik}^s - \beta_k^s \le c_k^z, \forall k \in K \tag{A.1e}$$

$$\beta_k^s \le c_k^o, \forall k \in K \tag{A.1f}$$

$$f_{jp}^{ks} \ge 0, \forall p > j \in I, k \in K \tag{A.1 g}$$

$$\alpha_{ik}^s, \beta_k^s \ge 0, \forall i \in I, k \in K \tag{A.1 h}$$

By the objective function (A.1a) in the dual problem (A.1a)–(A.1h), we can derive the optimality cuts. We model the master problem of SMIP added the optimality cuts as the following formulation:

$$min 1/S \sum_{s\in S}\delta_s \tag{A.2a}$$

$$s.t.\ (2b)–(2b) \tag{A.2b}$$

$$\delta_s \ge \sum_{k\in K}\sum_{j<p\in I} f_{jp}^{ks}\left(q_j^s d_j^s - M(2-x_{pk}-x_{jk}) - a_p + a_j\right) + \sum_{k\in K}\sum_{i\in I}\alpha_{ik}^s[a_i - \sum_{j\in I, j\neq i} q_j^s d_j^s x_{jk} - M(1-x_{ik}) + \sum_{k\in K}\beta_k^s(\sum_{i\in I} q_i^s d_i^s x_{ik} - T), \forall s \in S \tag{A.2c}$$

where $\delta_s, \forall s \in S$ is continuous variable. Constraints (A.2c) are the optimality cuts.

Finally, the steps of the Benders decomposition algorithm are presented as the following pseudocode:

CBD algorithm for the proposed problem

1: Initialization. Set $UB = +\infty$, $LB = -\infty$. Select a convergence tolerance parameter $\varepsilon \ge 0$.
2: Solve the master problem. Record the optimal solutions $(\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{a}^*, \boldsymbol{\delta}^*, \boldsymbol{\mu}^q)$ and the optimal objective value z^{MP}. Set $LB = z^{MP}$.
3: Stochastic service time and time-dependent no-shows scenario generation. Sample $S = 1000$ *i.i.d.* realizations $(q_1^s, d_1^s), \ldots, (q_n^s, d_n^s), s = 1, \ldots, S$ by given service time distribution and no-shows with mean $\boldsymbol{\mu}^q$.
4: **for** each $s \in [1, S]$ **do**
5: Solve the subproblem SP^s (3a)–(3f) with $\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{a}^*$ and record the optimal objective value z_s^{SP}.
6: Get the extreme point.
7: Add optimality Benders cut (A.2c) to the master problem.
8: **end for**
9: Calculate $\delta = 1/S(\sum_1^S z_s^{SP})$.
10: Set $UB = min\{UB, z^{MP} + \delta - \delta^*\}$.
11: **if** $UB - LB \le \varepsilon$ **then**
12: Terminate. return $\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{a}^*$ as the optimal solution and UB as the optimal value.
13: **else**
14: Go to line 2.
15: **end if**

Appendix B:. Comparison results of MSAS and SSAS under homogeneous servers

The MSAS and SSAS models have been analyzed regarding the total costs when every physician is homogeneous with the same office and virtual service time, unit overtime and idle time cost. We consider different number of servers and patients, with 40 % virtual and 60 % office patients as illustrated in Table B1. The time limit is set at $T = (nr_o\mu_{ok}^d + nr_e\mu_{ek}^d)/m$. The optimization results are summarized in Table B1. The results indicate that the MSAS model performs better in seven cases, while the performance of the MSAS model is worse in the other eight cases.

References

[1] X. Zhong, A queueing approach for appointment capacity planning in primary care clinics with electronic visits, IISE Transactions 50 (11) (2018) 970–988.
[2] T. Cayirli, E. Veral, Outpatient scheduling in health care: a review of literature, Prod. Oper. Manag. 12 (4) (2003) 519–549.
[3] C. Zacharias, M. Pinedo, Appointment scheduling with no-shows and overbooking, Prod. Oper. Manag. 23 (5) (2014) 788–801.
[4] Q. Kong, S. Li, N. Liu, C.P. Teo, Z. Yan, Appointment scheduling under time-dependent patient no-show behavior, Manag. Sci. 66 (8) (2020) 3480–3500.
[5] N. Liu, S. Ziya, Panel size and overbooking decisions for appointment-based services under patient no-shows, Prod. Oper. Manag. 23 (12) (2014) 2209–2223.
[6] Y. Deng, S. Shen, Decomposition algorithms for optimizing multi-server appointment scheduling with chance constraints, Math. Program. 157 (1) (2016) 245–276.

[7] A. Mandelbaum, P. Momilovi, N. Trichakis, S. Kadish, R. Leib, C.A. Bunnell, Data-driven appointment-scheduling under uncertainty: the case of an infusion unit in a cancer center, Manag. Sci. 66 (1) (2020) 243–270.
[8] J. Castaing, A. Cohn, B.T. Denton, A. Weizer, A stochastic programming approach to reduce patient wait times and overtime in an outpatient infusion center, IISE Transactions on Healthcare Systems Engineering 6 (3) (2016) 111–125.
[9] N.B. Demir, S. Gul, M. Elik, A stochastic programming approach for chemotherapy appointment scheduling, Nav. Res. Logist. 68 (1) (2021) 112–133.
[10] X. Zhong, P. Hoonakker, P.A. Bain, A.J. Musa, J. Li, The impact of e-visits on patient access to primary care, Health Care Manag. Sci. 21 (4) (2017) 475–491.
[11] Z. Xiang, L. Jingshan, P.A. Bain, A.J. Musa, Electronic visits in primary care: modeling, analysis, and scheduling policies, IEEE Trans. Autom. Sci. Eng. 14 (3) (2017) 1451–1466.
[12] A. Bayram, S. Deo, S. Iravani, K. Smilowitz, Managing virtual appointments in chronic care, IISE Transactions on Healthcare Systems Engineering 10 (1) (2020) 1–17.
[13] H. Wen, J. Song, X. Pan, Physician recommendation on healthcare appointment platforms considering patient choice, IEEE Trans. Autom. Sci. Eng. 17 (2) (2020) 886–899.
[14] X. Pan, J. Song, J. Zhao, V.A. Truong, Online contextual learning with perishable resources allocation, IISE Transactions 52 (12) (2020) 1343–1357.
[15] R.R. Chen, L.W. Robinson, Sequencing and scheduling appointments with potential call-in patients, Prod. Oper. Manag. 23 (9) (2014) 1522–1538.
[16] L.W. Robinson, R.R. Chen, A comparison of traditional and open-access policies for appointment scheduling, Manuf. Serv. Oper. Manag. 12 (2) (2010) 330–346.
[17] X. Qu, Y. Peng, J. Shi, L. LaGanga, An MDP model for walk-in patient admission management in primary care clinics, Int. J. Prod. Econ. 168 (oct.) (2015) 303–320.
[18] X. Pan, N. Geng, X. Xie, J. Wen, Managing appointments with waiting time targets and random walk-ins, Omega 95 (2020), 102062.
[19] S. Wang, N. Liu, G. Wan, Managing appointment-based services in the presence of walk-in customers, Manag. Sci. 66 (2) (2020) 667–686.
[20] P.M. Koeleman, G.M. Koole, Optimal outpatient appointment scheduling with emergency arrivals and general service times, IISE Transactions on Healthcare Systems Engineering 2 (1) (2012) 14–30.
[21] J. Luo, V.G. Kulkarni, S. Ziya, Appointment scheduling under patient no-shows and service interruptions, Manuf. Serv. Oper. Manag. 14 (4) (2012) 670–684.
[22] A. Sauré, J. Patrick, S. Tyldesley, M.L. Puterman, Dynamic multi-appointment patient scheduling for radiation therapy, Eur. J. Oper. Res. 223 (2) (2012) 573–584.
[23] S. Yu, V.G. Kulkarni, V. Deshpande, Appointment scheduling for a health care facility with series patients, Prod. Oper. Manag. 29 (2) (2020) 388–409.
[24] A. Sauré, M.A. Begen, J. Patrick, Dynamic multi-priority, multi-class patient scheduling with stochastic service times, Eur. J. Oper. Res. 280 (1) (2020) 254–265.
[25] K.S. Shehadeh, A.E.M. Cohn, R. Jiang, Using stochastic programming to solve an outpatient appointment scheduling problem with random service and arrival times, Nav. Res. Logist. 68 (1) (2021) 89–111.
[26] S.J. Lee, G.R. Heim, C. Sriskandarajah, Y. Zhu, Outpatient appointment block scheduling under patient heterogeneity and patient no-shows, Prod. Oper. Manag. 27 (1) (2018) 28–48.
[27] R. Kolisch, S. Sickinger, Providing radiology health care services to stochastic demand of different customer classes, OR Spectr. 30 (2) (2007) 375–395.
[28] S. Zhou, Q. Yue, Sequencing and scheduling appointments for multi-stage service systems with stochastic service durations and no-shows, Int. J. Prod. Res. 2 (2021) 1–20.
[29] R. Jiang, S. Shen, Y. Zhang, Integer programming approaches for appointment scheduling with random no-shows and service durations, Oper. Res. 65 (6) (2017) 1638–1656.
[30] K. Muthuraman, M. Lawley, A stochastic overbooking model for outpatient clinical scheduling with no-shows, IIE Trans. 40 (9) (2008) 820–837.
[31] N. Liu, S. Ziya, V.G. Kulkarni, Dynamic scheduling of outpatient appointments under patient no-shows and cancellations, Manuf. Serv. Oper. Manag. 12 (2) (2010) 347–364.
[32] C. Zacharias, M. Pinedo, Managing customer arrivals in service systems with multiple identical servers, Manuf. Serv. Oper. Manag. 19 (4) (2017) 639–656.
[33] R. Hassin, S. Mendel, Scheduling arrivals to queues: a single-server model with no-shows, Manag. Sci. 54 (3) (2008) 565–572.
[34] S. Zhou, D. Li, Y. Yin, Coordinated appointment scheduling with multiple providers and patient-and-physician matching cost in specialty care, Omega 101 (2021), 102285.
[35] Y. Hur, J.F. Bard, D.J. Morrice, Appointment scheduling at a multidisciplinary outpatient clinic using stochastic programming, Nav. Res. Logist. 68 (1) (2021) 134–155.
[36] X. Pan, N. Geng, X. Xie, A stochastic approximation approach for managing appointments in the presence of unpunctual patients, multiple servers and no-shows, Int. J. Prod. Res. 59 (10) (2021) 2996–3016.
[37] X. Wu, S. Zhou, Sequencing and scheduling appointments on multiple servers with stochastic service durations and customer arrivals, Omega 106 (2022), 102523.
[38] B. Shnits, I. Bendavid, Y.N. Marmor, An appointment scheduling policy for healthcare systems with parallel servers and pre-determined quality of service, Omega 97 (2020), 102095.
[39] X. Qu, J. Shi, Modeling the effect of patient choice on the performance of open access scheduling, Int. J. Prod. Econ. 129 (2) (2011) 314–327.
[40] K. Morikawa, K. Takahashi, Scheduling appointments for walk-ins, Int. J. Prod. Econ. 190 (2017) 60–66.
[41] X. Qu, Y. Peng, J. Shi, L. LaGanga, An MDP model for walk-in patient admission management in primary care clinics, Int. J. Prod. Econ. 168 (2015) 303–320.
[42] A. Saremi, P. Jula, T. ElMekkawy, G. Wang, Appointment scheduling of outpatient surgical services in a multistage operating room department, Int. J. Prod. Econ. 141 (2) (2013) 646–658.
[43] A.M. Geoffrion, G.W. Graves, Multicommodity distribution system design by Benders decomposition, Manag. Sci. 20 (5) (1974) 822–844.
[44] H. Salah, S. Srinivas, Predict, then schedule: Prescriptive analytics approach for machine learning-enabled sequential clinical scheduling, Comput. Ind. Eng. 169 (2022), 108270.
[45] S. Srinivas, A.R. Ravindran, Designing schedule configuration of a hybrid appointment system for a two-stage outpatient clinic with multiple servers, Health Care Manag. Sci. 23 (2020) 360–386.
[46] D. Gupta, B. Denton, Appointment scheduling in health care: Challenges and opportunities, IIE Trans. 40 (9) (2008) 800–819.
[47] L.F. Dantas, J.L. Fleck, F.L.C. Oliveira, S. Hamacher, No-shows in appointment scheduling–a systematic literature review, Health Policy 122 (4) (2018) 412–421.
[48] J. Marynissen, E. Demeulemeester, Literature review on multi-appointment scheduling problems in hospitals, Eur. J. Oper. Res. 272 (2) (2019) 407–419.
[49] Y. Chen, Y.H. Kuo, P. Fan, H. Balasubramanian, Appointment overbooking with different time slot structures, Comput. Ind. Eng. 124 (2018) 237–248.
[50] G.C. Kaandorp, G. Koole, Optimal outpatient appointment scheduling, Health Care Manag. Sci. 10 (2007) 217–229.
[51] Y.H. Kuo, H. Balasubramanian, Y. Chen, Medical appointment overbooking and optimal scheduling: tradeoffs between schedule efficiency and accessibility to service, Flex. Serv. Manuf. J. 32 (2020) 72–101.
[52] T. Cayirli, E. Veral, H. Rosen, Designing appointment scheduling systems for ambulatory care services, Health Care Manag. Sci. 9 (2006) 47–58.
[53] C. Yan, G.G. Huang, Y.H. Kuo, J. Tang, Dynamic appointment scheduling for outpatient clinics with multiple physicians and patient choice, Journal of Management Science and Engineering 7 (1) (2022) 19–35.
[54] J. Feldman, N. Liu, H. Topaloglu, S. Ziya, Appointment scheduling under patient preference and no-show behavior, Oper. Res. 62 (4) (2014) 794–811.
[55] H.J. Alvarez-Oh, H. Balasubramanian, E. Koker, A. Muriel, Stochastic appointment scheduling in a team primary care practice with two flexible nurses and two dedicated providers, Serv. Sci. 10 (3) (2018) 241–260.
[56] B. Jiang, J. Tang, C. Yan, A stochastic programming model for outpatient appointment scheduling considering unpunctuality, Omega 82 (2019) 70–82.
[57] G. Xiao, M. Dong, J. Li, L. Sun, Scheduling routine and call-in clinical appointments with revisits, Int. J. Prod. Res. 55 (6) (2017) 1767–1779.
[58] Z. Chen, T. Zhou, X. Ming, X. Zhang, R. Miao, Configuration optimization of service solution for smart product service system under hybrid uncertain environments, Adv. Eng. Inf. 52 (2022), 101632.
[59] H. Qiu, D. Wang, Y. Yin, T.E. Cheng, Y. Wang, An exact solution method for home health care scheduling with synchronized services, Nav. Res. Logist. 69 (5) (2022) 715–733.
[60] X. Yu, S. Shen, B. Badri-Koohi, H. Seada, Time window optimization for attended home service delivery under multiple sources of uncertainties, Comput. Oper. Res. 150 (2023), 106045.
[61] T.L. Magnanti, R.T. Wong, Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria, Oper. Res. 29 (3) (1981) 464–484.
[62] N. Papadakos, Practical enhancements to the Magnanti-Wong method, Oper. Res. Lett. 36 (4) (2008) 444–449.
[63] J.A. Rodríguez, M.F. Anjos, P. Côté, G. Desaulniers, Accelerating Benders decomposition for short-term hydropower maintenance scheduling, Eur. J. Oper. Res. 289 (1) (2021) 240–253.
[64] A.S. Michels, T.C. Lopes, C.G.S. Sikora, L. Magatão, A Benders' decomposition algorithm with combinatorial cuts for the multi-manned assembly line balancing problem, Eur. J. Oper. Res. 278 (3) (2019) 796–808.
[65] H. Zhang, K. Yang, Y. Gao, L. Yang, Accelerating Benders decomposition for stochastic incomplete multimodal hub location problem in many-to-many transportation and distribution systems, Int. J. Prod. Econ. 248 (2022), 108493.
[66] H. Gong, Z.H. Zhang, Benders decomposition for the distributionally robust optimization of pricing and reverse logistics network design in remanufacturing systems, Eur. J. Oper. Res. 297 (2) (2022) 496–510.
[67] F. García-Muñoz, S. Dávila, F. Quezada, A Benders decomposition approach for solving a two-stage local energy market problem under uncertainty, Appl. Energy 329 (2023), 120226.
[68] R. Rahmaniani, T.G. Crainic, M. Gendreau, W. Rei, The Benders decomposition algorithm: A literature review, Eur. J. Oper. Res. 259 (3) (2017) 801–817.
[69] X. Shen, S.C. Du, Y.N. Sun, P.Z. Sun, R. Law, E.Q. Wu, Advance scheduling for chronic care under online or offline revisit uncertainty, IEEE Trans. Autom. Sci. Eng. (2023) 1–14.