Advance Scheduling for Chronic Care Under Online or Offline Revisit Uncertainty

Xiaoxiao Shen, Shi-Chang Du, *Member, IEEE*, Yan-Ning Sun, Poly Z. H. Sun, *Member, IEEE*, Rob Law, and Edmond Q. Wu, *Senior Member, IEEE*

***Abstract*— Chronic disease patients often require revisits for long-term care. Online medical services shift revisits to online, which can improve the access to chronic care and reduce the burden on offline medical services. However, whether Internet healthcare can truly match the medical supply and demand, one of the critical issues is the efficient advance scheduling of the integrated online and offline systems. This study investigates the advance scheduling problem for the first visit and revisit patients in chronic care. The uncertainty of revisit status (i.e., online or offline) and heterogeneity of online and offline revisits (i.e., revisit interval, continuity of care violation penalty) are considered. A stochastic mixed-integer programming model is formulated for assigning patients to a specific physician on a specific day over the course of a finite planning period. The aim is to minimize the expected sum of three cost components related to offline and online services: overtime and idle time, continuity of care violation penalty, and fixed setup. This study proposes a modified progressive hedging algorithm and applies a sequential decision-making framework to obtain rolling time advance schedules. Results of the numerical analysis demonstrate the effectiveness of our algorithm compared to both the published state-of-the-art Lagrangian decomposition embedded with surrogate subgradient method and the commercial solver Gurobi. The insight obtained from the experiments is that a capacity allocation scheme with all physicians assigned with both offline and online capacities would be a good choice for considerable cost savings.**

***Note to Practitioners*—Internet healthcare is becoming increasingly popular. Operation and management issues have arisen in the integrated online and offline appointment systems. A sequential decision-making method embedded with a stochastic programming model and a modified PHA is proposed to help decision-makers generate the first visit and revisit advance schedules for chronic care. The performance of this approach and the system is thoroughly verified. Results show that the developed decision technique can lessen the operational cost generated by scheduling and realize the goal of continuity of care. This study offers a useful tool to help with intelligent patient advance scheduling in an integrated management system of online and offline chronic care.**

***Index Terms*— Patient advance scheduling, online and offline healthcare, progressive hedging, chronic care, stochastic programming.**

Manuscript received 8 June 2023; accepted 28 August 2023. Date of publication 4 September 2023; date of current version 16 October 2024. This article was recommended for publication by Associate Editor H. K. Lee and Editor X. Xie upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 52275499 and in part by the National Key Research and Development Program of China under Grant 2022YFF0605700. *(Corresponding author: Shi-Chang Du.)*

Xiaoxiao Shen is with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240 China, and also with the Department of Mechanical Engineering, Politecnico di Milano, 20156 Milan, Italy (e-mail: sjtusxx98@sjtu.edu.cn).

Shi-Chang Du and Poly Z. H. Sun are with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: lovbin@sjtu.edu.cn; zh.sun@sjtu.edu.cn).

Yan-Ning Sun is with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China (e-mail: sunyanning@shu.edu.cn).

Rob Law is with the Asia–Pacific Academy of Economics and Management, University of Macau, Taipa, Macau 999078, China (e-mail: roblaw@um.edu.mo).

Edmond Q. Wu is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: edmondqwu@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2023.3310116.

Digital Object Identifier 10.1109/TASE.2023.3310116

I. Introduction

INTERNET hospitals are playing an increasingly important role in serving patients seeking medical treatment and medicine. Since the outbreak of the Coronavirus disease 2019 in Shanghai in 2022, over 100 Internet hospitals in the city, have provided convenient services online, including specialist consultation, revisit consultation, and prescription. The total volume of Internet services in the city has increased significantly at an average of about 50%–80% [1]. Internet hospitals alleviate the contradiction between the supply and demand of medical services. They can divert patient demand and relieve the pressure of offline hospital services. At the same time, the rise of Internet medical care has also greatly improved the efficiency and effectiveness of chronic disease management. According to the research data of a tertiary hospital [2], after using a diabetes management smart medical platform, the average hospital stay of inpatients was reduced by 3.6 days, and the average blood sugar compliance time was decreased by 1.3 days. Despite the observed benefits of Internet healthcare in chronic care, various ambiguities occur in intelligent scheduling of the integrated online and offline medical systems.

This study is motivated by one of our investigated hospitals that provide both online and offline services. In chronic disease management, the hospital practitioners struggle with generating efficient schedules in advance with fewer costs. Their current practice is to arrange appointments separately for the first visit and revisit patients. After the first visit, the patients are not scheduled for revisits until the diagnosis information of the first visit is obtained. At this point, service capacity is highly occupied by the demands of the first visit patients who were scheduled before the current decision point. Thus, physicians often work overtime or returning patients face delayed treatments. The hospital has difficulties in guaranteeing to promptly serve returning patients within the medically required interval. Hence, we investigate the advanced scheduling problem for chronic care in the integrated online and offline medical systems.

Moreover, even in existing literature, assigning patients to future appointment days during a multi-day scheduling window is typically considered as a difficult managerial issue. Advance scheduling has wide and varied application areas, including the arrangement for diagnostic procedures such as Magnetic Resonance Imaging or Computed Tomography scans [3], [4], [5], [6], [7], radiotherapy, physical therapy, treatment scheduling [8], [9], primary care clinics [10], [11], [12], [13], and surgical scheduling [14], [15], [16], [17], [18], [19], [20], [21]. The most common approach used to formulate the advance scheduling problem is the Markov decision process (MDP) [4], [5], [6], [7], [8], [9], [13], [14], [22], [23], [24], [25]. However, high-dimensional state and action spaces make exact solution methods intractable. Such problem is addressed using different approaches, such as policy iteration algorithm [9], [14] and approximate dynamic programming (ADP) [4], [26]. Unlike the above studies, Dai et al. [13] established the structural properties of the finite-horizon MDP model and optimal scheduling policy, as well as designed two efficient heuristic policies from the theoretical results. Mathematical programming approaches are also applied to formulate advance scheduling problems [15], [16], [18], [19], [20], [21]. The majority of the aforementioned papers consider uncertainties and use various techniques to solve the stochastic model, including the scenario-based approach sample average approximation (SAA) [17], [20], [21], distributionally robust optimization [15], [19], and robust optimization [27]. As far as we know, the present study is the first work to build an advance scheduling model that contributes to chronic disease management from an operational perspective. Our model considers the three-pronged nature of the hybrid online and offline services system. The model addresses the uncertainty of online or offline revisit status while also ensures the continuity of care for patients by assigning them to a specific physician on a specific day. We consider online and offline revisit patient heterogeneity in their revisit intervals and continuity of care violation penalties. Additionally, we use a modified version of the progressive hedging algorithm (PHA) based on a stochastic problem structure to search for near-optimal chronic care schedules.

A few studies have focused on the advance scheduling problem considering revisits. As an illustration of work in this field, a dynamic model is developed by Sauré et al. [8] for scheduling radiation therapy patients with multiple appointments. Patients are assumed to have a constant number of appointments at their initial appointment. The studies by Yu et al. [9], Bayram et al. [28], and Yu and Bayram [29] are most closely related to our research. Considering a patient's random number of visits and a constant interval in between, Yu et al. [9] prepared schedules for a series of appointments. The difference of our study is that we consider two groups of revisited patients from both online and offline. To maximize aggregate health benefits, Bayram et al. [28] determined which patients must be scheduled for office and virtual appointments by building a finite horizon stochastic dynamic program. Yu and Bayram [29] used a migration network to model chronic new and returning patients' flow, making decisions to allocate the capacity of office and virtual appointments to maximize long-run average earnings. In both cases [28], [29], the authors decided whether a patient receives care online or offline. However, we consider the online or offline status of revisits in a stochastic manner. In addition, to ensure continuity of care, we define decision variables that introduce physician subscription to set soft constraints, which differentiates our modeling approach from those of the abovementioned research.

One of the popular techniques for solving large-scale stochastic mixed-integer programming problems is the decomposition method. The optimization method PHA used in this study is under the umbrella of Lagrangian decomposition (LD) which is a specific instance of Lagrangian relaxation. LD was first proposed by Carøe and Schultz [33], and the idea behind LD is splitting the original problem into smaller scenario subproblems by relaxing non-anticipativity constraints and finding the optimal dual multipliers to solve the problem. Since then, many pieces of literature have further studied LD. This method has been widely applied in practice and innovatively developed in theory, which is mainly reflected in combining other methods to form a new integration method. For instance, a recent application of ambulance relocation and routing under stochastic demand is presented in [34], where an algorithm composed of the Lagrangian dual decomposition and branch-and-bound is developed to accelerate the solution time. Escudero et al. [35], [36], [37], [38] investigate the Lagrangian relaxation application in solving multistage stochastic mixed programs with risk-averse measures and enrich traditional LD by developing scenario cluster approach, scenario cluster submodels optimization, and Lagrangian multipliers updating schemes. A LD method is proposed that combines cutting planes and subgradient methods in [39]. Lara et al. [40] consider a temporal fixed-charge flow problem in transportation applications and propose a new algorithm that composes balanced graph partitioning, LD, and a linear programming filtering heuristic. Zeighami et al. [41] develop a new hybrid approach integrating alternating LD, column generation, and dynamic constraint aggregation to address the integrated crew pairing and personalized assignment problems in airline applications. Zhang et al. [42] address a rail traffic real-time optimization problem for collaborative decisions of train rescheduling and track emergency maintenance and develop a Lagrangian relaxation based decomposition algorithm. Amiri and Barkhi [43] propose a Lagrangian relaxation based method to solve the multiple knapsack problem with setups. A thorough review of the Lagrangian relaxation method can be referred to [44].

The main contributions of the present study are summarized as follows. First, we study the advance scheduling problem for chronic care with first and revisit patients that explicitly considers the uncertainty and heterogeneity of online and offline revisits, such as revisit interval and continuity of care violation penalty. To the best of our knowledge, this study is the first to establish a chronic care advance scheduling model with first and revisits in integrated online and offline systems. Second, in a sequential decision framework, a stochastic mixed integer programming model is constructed and a modified PHA is developed. Penalty update and Lagrangian multiplier update methods are proposed, which consider the convergence behavior of the variable values. Numerical results indicate that our method performs better than the state-of-the-art Lagrangian decomposition approaches and the commercial solver Gurobi. The proposed modified PHA can obtain satisfied solutions in a relatively quite short time particularly for these large scale instances of the problems. Finally, managerial insights for guiding practice are observed through numerical studies. For example, when the patient volume is large, the allocation scheme with all physicians assigned for both offline and online capacities is highly recommended for better performance.

The remainder of this paper is structured as follows. The problem description and the sequential stochastic mixed integer programming model are presented in Section II. Section III presents the solution approach. In Section IV, our numerical experiments are performed and managerial insights are drawn. Discussions of our findings, their shortcomings, and possible future research directions are also provided. Finally, Section V provides the conclusions.

II. Problem Formulation

In this section, we describe the advance scheduling problem considering chronic patients with both first visit and revisit in offline and online care. A sequential stochastic mixed-integer programming model is developed for this problem. Then, we analyze the complexities of the model.

A. Mathematical Model

For chronic healthcare services of diabetes, cardiovascular disease, high blood pressure, Alzheimer's disease, and asthma, patients need revisits to receive long-term treatment and constant monitoring of their health status. We consider the hybrid offline and online advance scheduling problem for first visit and revisit patients suffering from chronic diseases. The first K days are denoted as the arrival horizon, indexed by $k \in \mathbb{K} = \{1, 2, \ldots, K\}$. In this study, the terms "period" and "day" are used interchangeably. In each period k of the horizon, first visit patients come to make appointment requests on future days. Since the first visit generally requires diagnostic tests, we assume that all first visits are served at the offline hospital. The first visit patients that require revisits (FV with RV) are denoted by $i_1 \in \mathbb{I}_1 = \{1, 2, \ldots, I_1\}$. The remaining first visit patients are those who leave the system after the first visit to complete treatment without a revisit (FV without RV), denoted by $i_2 \in \mathbb{I}_2 = \{1, 2, \ldots, I_2\}$. These two types of patients have to be given the first appointment on day $l \in \left\{k+1, k+2, \ldots, k+M^1\right\}$ and day $l \in \left\{k+1, k+2, \ldots, k+M^2\right\}$, respectively, where M^1 and M^2 are two given positive integers, called their appointment Maximum Wait Time Targets (MWTTs).

We consider the uncertainties of patient's revisit status (i.e., online or offline) because of its dependence on the physician's diagnosis and patient's specific conditions. Patients with chronic disease-related diagnoses, long-term medication, and stable conditions are recommended to undergo online follow-up consultation, and patients with unclear or changing conditions are recommended for offline follow-up consultation. The definition of the scenario and the underlying assumption of the rule we used to generate scenarios are now described. For each FV with RV $i_1 = 1, \ldots, I_1$, the random variable of the revisit access status $Q \in \{0, 1\}$ follows the Bernoulli distribution with a parameter p_{i_1}. That is, $Q \sim B(p_{i_1})$. A scenario s for the FV with RV contains revisit access information about that patient i_1 receives returning care online or offline q_{i_1} (e.g., $q_{i_1} = 1$ represents online, and $q_{i_1} = 0$ represents offline). q_{i_1} is a realization of Q, and the index set of scenarios is denoted by $\mathbb{S} := \{s\}$. In addition, the revisit interval of patient i_1 depending on revisit status, $M^3_{i_1}$ (positive integer), is defined and restricted to $\underline{a}^o \leq M^3_{i_1} \leq \overline{a}^o$ for revisits offline and $\underline{a}^e \leq M^3_{i_1} \leq \overline{a}^e$ for online.

Set $\mathbb{J} = \{1, 2, \ldots, J\}$ represents the set of physicians and c_{jt} is the total capacities of physician $j \in \mathbb{J}$ available in day t. Whether physician $j \in \mathbb{J}$ is scheduled to perform offline (online) medical service on day t is denoted by $\alpha_{jt}(\beta_{jt})$, which equals one if yes and zero otherwise. The regular offline (online) capacity of physician $j \in \mathbb{J}$ available on day t is denoted by $c^o_{jt}(c^e_{jt})$. Thus, clearly, $c^o_{jt} \leq c_{jt}(c^e_{jt} \leq c_{jt})$ for all j. The fixed-length service times of FV with RV, online revisit, offline revisit, and FV without RV on physician j are denoted by b^1_j, b^2_j, b^3_j, and b^4_j, respectively. To ensure continuity of patient care, we allow patients to be served by the same physician for first visit and revisits as much as possible, and model this as soft constraints. The planner incurs penalty costs of $c^e_p(c^o_p)$ if an offline (online) revisit cannot be served by the same physician. The fixed setup cost of physician $j \in \mathbb{J}$ scheduled to perform offline (online) medical service each day is $c^o_u(c^e_u)$. We denote unit overtime or idle cost on physician performing offline and online services as c^o_b, c^e_b, respectively.

At the decision point, which is a precise period of the day given the current appointment schedule, the hospital scheduler arranges new first visit requests and their potential revisits over an L-day planning window. Patients arrive for their first visit during the entire day, therefore their status is not fully known until the end of the day, which we thus assume to correspond to decision points. Note that, following the literature [9], our model schedules potential revisits of first visit patients together at the decision point to ensure continuity of care. Therefore, the planning horizon is set as $L = max\left\{M^2, M^1 + max\ M^3_{i_1}\right\}$ to accommodate the longest appointment lead times possible for both first visit and revisit patients. Our assumption is that patients accept the appointment date provided by the scheduler because they do not have a strong preference for the day of their consultation.

Rolling time patient advanced scheduling is accomplished by using a sequential decision-making framework. In sequential decision-making, according to the principle that at the start of each decision time, on the basis of all the historical information, the optimization of advanced scheduling is

TABLE I
MODEL NOTATION

Indexes and Sets	Description
$k \in \mathbb{K} = \{1, \ldots, K\}$	Set of periods within an arriving horizon of K days, indexed by k
$l \in \mathbb{L} = \{1, \ldots, L\}$	Set of periods in the L-day planning window, indexed by l
$i_1 \in \mathbb{I}_1 = \{1, \ldots, I_1\}$	Set of FV with RV patients, indexed by i_1
$i_2 \in \mathbb{I}_2 = \{1, \ldots, I_2\}$	Set of FV without RV patients, indexed by i_2
$j \in \mathbb{J} = \{1, \ldots, J\}$	Set of physicians, indexed by j
$s \in \mathbb{S} = \{1, \ldots, S\}$	Set of scenarios, indexed by s
Parameters	**Description**
$M^1(M^2)$	FV with RV (without RV) patient MWTT
c_{jt}	Total capacities of physician $j \in \mathbb{J}$ available for use in day t
$\alpha_{jt}(\beta_{jt})$	Status of physician $j \in \mathbb{J}$ whether to perform offline (online) medical service on day t
$c^o_{jt}(c^e_{jt})$	Regular offline (online) capacity of physician $j \in \mathbb{J}$ available on day t
$q^s_{i_1}$	Revisit status of FV with RV patient i_1 (i.e., online or offline) in scenario s
p_s	Probability of scenario s
M	A large number
$M^3_{i_1}$	Revisit interval of FV with RV patient i_1
$\underline{a}^o(\underline{a}^e), \overline{a}^o(\overline{a}^e)$	Minimum and maximum revisit interval of offline (online) revisit in scenario s
$b^1_j, b^2_j, b^3_j, b^4_j$	Fixed service length of a FV with RV, online revisit, offline revisit and FV without RV on physician j
$c^o_u(c^e_u)$	Fixed setup cost of physician $j \in \mathbb{J}$ scheduled to perform offline (online) medical service on each day
$c^e_p(c^o_p)$	Penalty cost if an offline(online) revisit that cannot be served by the same physician as first visit
$c^o_b(c^e_b)$	Unit overtime or idle cost of physician performing offline (online) service
Decision variables	**Description**
$x^{\mathrm{I}}_{i_1jt}$	Binary variable, 1 if FV with RV patient i_1 is assigned to physician j on day t, 0 otherwise
$x^{\mathrm{II}}_{i_2jt}$	Binary variable, 1 if FV without RV patient i_2 is assigned to physician j on day t, 0 otherwise
$y^s_{i_1jt}(z^s_{i_1jt})$	Binary variable, 1 if FV with RV patient i_1 is assigned to physician j for an offline (online) revisit on day t in scenario s, 0 otherwise
$o^{os}_{jt}(o^{es}_{jt})$,$a^{os}_{jt}(a^{es}_{jt})$	Continuous variable, physician j's offline (online) overtime or idle time after finishing its appointments on day t in scenario s

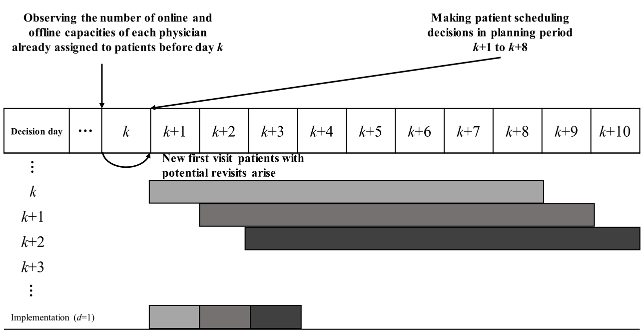

Fig. 1. A sequential decision-making framework for advance scheduling.

implemented for a finite planning horizon; however, only the decisions made in the present duration are carried out. Fig. 1 shows the appointment schedule in period k before the decision point can be observed, including the information on the number of online and offline capacities of each physician already assigned to first visit and revisit appointments on day $t\,(t = k{+}1, \ldots, k{+}L{-}1)$. At the end of period k, the scheduler must make decisions of assigning the new first visit patients who arrive on day k and their potential revisits to a certain physician $j\,(j \in \mathbb{J})$ on a certain day $t\,(t = k+1, \ldots, k+L)$. After each decision interval, the above-mentioned procedure is repeated and the updated information can be incorporated into the optimization at the present decision point. In this study, new appointment requests arrive on a daily basis, and thus the decision interval $d = 1$. Based on the current appointment schedule, we make the optimization for the patient scheduling from day $k + 1$ to day $k + L$; however, only the decisions of day k are implemented; then, on day $k + d$, a new decision optimization is carried out. Fig. 1 shows the structure of the proposed framework with $L = 8$ and $d = 1$.

We now present a stochastic mixed-integer formulation for the online and offline advance scheduling problem with both first visit and revisits. At a particular decision point k, the model is used to make decisions for the following L days. Table I lists a summary of the notations of this model.

The objective function aims to minimize the three-part total cost. The first part is the overtime and idle time costs of offline and online capacities. The second and third parts are the hospital's fixed costs for launching offline and online medical services and the continuity of care violation penalty cost of online and offline revisits, respectively. Constraints (1b)-(1c), as shown at the bottom of the next page, ensure that each first visit patient with revisit and without revisit receives a first appointment. Constraints (1d)-(1e) ensure that each online and offline revisit patient receives a revisit appointment. Constraints (1f)-(1i) enforce that physicians are assigned with offline and online appointments on scheduled days for such services. Constraints (1j)-(1m) place limitations on the specific day that a revisit patient may be allocated. When a revisit request is generated by a first visit, it must be scheduled after the first visit date and within the allowable revisit interval ($M^3_{i_1}$) for the revisit. These constraints also ensure continuity of care, with the first and revisit appointments treated by the same doctor. Note that the left and right endpoints of offline and online revisit intervals are set differently in constraints (1j)-(1m). Constraints (1n)-(1o) assure that the MWTT of each first visit patient with revisit and without revisit is satisfied. For physician j on day t during the planning horizon, constraints (1p)-(1q) determine the amount of online overtime and idle time, respectively, while similar constraints (1r)-(1s) calculate the offline costs. The integrality and non-negativity limitations on the decision variables are defined by constraints (1t).

B. Complexity of the Problem

The integrated online and offline advance scheduling problem for first visit and revisit can be formulated as a sequential stochastic mixed-integer programming model as described above. Small-scale problems can be solved directly by the most advanced MIP solvers (such as GUROBI). However, large-scale instances of the problem are computationally intractable due to the enormous number of variables and constraints in the model. Specifically, for K-day rolling planning, there are $|\mathbb{K}|\cdot(|\mathbb{J}|\cdot|\mathbb{L}|\cdot(|\mathbb{I}_1|\cdot(2\cdot|\mathbb{S}|+1)+|\mathbb{I}_2|+4\cdot|\mathbb{S}|))$ variables and $|\mathbb{K}|\cdot(2\cdot|\mathbb{I}_1|\cdot(1+3\cdot|\mathbb{S}|)+2\cdot|\mathbb{I}_2|+2\cdot|\mathbb{J}|\cdot$

$$\min \quad \sum_{s=1}^{S} p_s[\sum_{j\in\mathbb{J}}\sum_{t=1}^{T} c_b^o(o_{jt}^{os}+a_{jt}^{os})+c_b^e(o_{jt}^{es}+a_{jt}^{es})+c_u^o\alpha_{jt}+c_u^e\beta_{jt}+\sum_{i_1\in\mathbb{I}_1}\sum_{j\in\mathbb{J}} c_p^e q_{i_1}^s(\sum_{t=1}^{T} y_{i_1jt}^s-\sum_{t=1}^{T} x_{i_1jt}^{\mathrm{I}})+ c_p^o(1-q_{i_1}^s)(\sum_{t=1}^{T} z_{i_1jt}^s-\sum_{t=1}^{T} x_{i_1jt}^{\mathrm{I}})] \tag{1a}$$

$$s.t. \quad \sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} x_{i_1jt}^{\mathrm{I}}=1, \forall i_1\in\mathbb{I}_1 \tag{1b}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} x_{i_2jt}^{\mathrm{II}}=1, \forall i_2\in\mathbb{I}_2 \tag{1c}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} y_{i_1jt}^s=q_{i_1}^s, \forall i_1\in\mathbb{I}_1, s\in\mathbb{S} \tag{1d}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} z_{i_1jt}^s=1-q_{i_1}^s, \forall i_1\in\mathbb{I}_1, s\in\mathbb{S} \tag{1e}$$

$$x_{i_1jt}^{\mathrm{I}}\le\alpha_{jt}, \forall i_1\in\mathbb{I}_1, j\in\mathbb{J}, t=k+1,\ldots,k+L \tag{1f}$$

$$x_{i_2jt}^{\mathrm{II}}\le\alpha_{jt}, \forall i_2\in\mathbb{I}_2, j\in\mathbb{J}, t=k+1,\ldots,k+L \tag{1g}$$

$$y_{i_1jt}^s\le\beta_{jt}, \forall i_1\in\mathbb{I}_1, j\in\mathbb{J}, t=k+1,\ldots,k+L, s\in\mathbb{S} \tag{1h}$$

$$z_{i_1jt}^s\le\alpha_{jt}, \forall i_1\in\mathbb{I}_1, j\in\mathbb{J}, t=k+1,\ldots,k+L, s\in\mathbb{S} \tag{1i}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} ty_{i_1jt}^s\ge\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tx_{i_1jt}^{\mathrm{I}}+\underline{a}^e-M\big(1-q_{i_1}^s\big), \forall i_1\in\mathbb{I}_1, s\in\mathbb{S} \tag{1j}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} ty_{i_1jt}^s\le\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tx_{i_1jt}^{\mathrm{I}}+\overline{a}^e+M\big(1-q_{i_1}^s\big), \forall i_1\in\mathbb{I}_1, s\in\mathbb{S} \tag{1k}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tz_{i_1jt}^s\ge\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tx_{i_1jt}^{\mathrm{I}}+\underline{a}^o-Mq_{i_1}^s, \forall i_1\in\mathbb{I}_1, s\in\mathbb{S} \tag{1l}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tz_{i_1jt}^s\le\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tx_{i_1jt}^{\mathrm{I}}+\overline{a}^o+Mq_{i_1}^s, \forall i_1\in\mathbb{I}_1, s\in\mathbb{S} \tag{1m}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tx_{i_1jt}^{\mathrm{I}}\le M^1, \forall i_1\in\mathbb{I}_1 \tag{1n}$$

$$\sum_{j\in\mathbb{J}}\sum_{t=k+1}^{k+L} tx_{i_2jt}^{\mathrm{II}}\le M^2, \forall i_2\in\mathbb{I}_2 \tag{1o}$$

$$o_{jt}^{os}\ge\sum_{i_1\in\mathbb{I}_1} b_j^1x_{i_1jt}^{\mathrm{I}}+\sum_{i_1\in\mathbb{I}_1} b_j^3z_{i_1jt}^s+\sum_{i_2\in\mathbb{I}_2} b_j^4x_{i_2jt}^{\mathrm{II}}-c_{jt}^o\alpha_{jt}, \forall j\in\mathbb{J}, t=k+1,\ldots,k+L, s\in\mathbb{S} \tag{1p}$$

$$a_{jt}^{os}\ge c_{jt}^o\alpha_{jt}-\sum_{i_1\in\mathbb{I}_1} b_j^1x_{i_1jt}^{\mathrm{I}}-\sum_{i_1\in\mathbb{I}_1} b_j^3z_{i_1jt}^s-\sum_{i_2\in\mathbb{I}_2} b_j^4x_{i_2jt}^{\mathrm{II}}, \forall j\in\mathbb{J}, t=k+1,\ldots,k+L, s\in\mathbb{S} \tag{1q}$$

$$o_{jt}^{es}\ge\sum_{i_1\in\mathbb{I}_1} b_j^2y_{i_1jt}^s-c_{jt}^e\beta_{jt}, \forall j\in\mathbb{J}, t=k+1,\ldots,k+L, s\in\mathbb{S} \tag{1r}$$

$$a_{jt}^{es}\ge c_{jt}^e\beta_{jt}-\sum_{i_1\in\mathbb{I}_1} b_j^2y_{i_1jt}^s, \forall j\in\mathbb{J}, t=k+1,\ldots,k+L, s\in\mathbb{S} \tag{1s}$$

$$x_{i_1jt}^{\mathrm{I}}, x_{i_2jt}^{\mathrm{II}}, y_{i_1jt}^s, z_{i_1jt}^s\in\{0,1\}, o_{jt}^{os}, a_{jt}^{os}, o_{jt}^{es}, a_{jt}^{es}\ge 0, \forall i_1\in\mathbb{I}_1, i_2\in\mathbb{I}_2, j\in\mathbb{J}, t=k+1,\ldots,k+L, s\in\mathbb{S} \tag{1t}$$

$|\mathbb{L}|(|\mathbb{I}_1|\cdot(2\cdot|\mathbb{S}|+1)+|\mathbb{I}_2|+4\cdot|\mathbb{S}|))$ constraints. Even for a small instance with only 20 patients, 4 physicians, and 100 scenarios, the model contains more than 387,000 variables and 804,000 constraints. Furthermore, it is straightforward to identify the complexity of the problem, since mixed-integer programs are a class of NP-hard problems [45]. These difficulties drive our creation of an efficacious algorithm for resolving the problem.

III. Solution Approach

A. Progressive Hedging Algorithm

For the stochastic mixed-integer problem at hand, as the scenarios expand, the problem size becomes intractable due to the growing number of variables and constraints. In particular, the solution needs to be obtained in quite short time due to "online scheduling" settings. To address these issues, we develop a PHA based on the well-known Lagrangian relaxation technique. The successful application of PHA in resolving numerous combinatorial optimization issues over the past several decades has motivated our adoption of it. The extensive use of this approach was inspired by the first presentation in [46]. Since then, it has been used to solve a variety of issues, including the scheduling of chemotherapy appointments and surgeries, tank container operations, and routing of unmanned aerial vehicles. The success of the method can be attributed to its capacity to solve computationally challenging problems by decomposing them into more manageable scenario subproblems. While the PHA does not guarantee global optimality, it finds a good solution efficiently. Note that the PHA is a useful heuristic for stochastic mixed-integer models, as demonstrated by a substantial body of literature [31], [32], [47].

To prohibit anticipation of the future, we first establish non-anticipativity (NAC) constraints to ensure the consistency of first visit assignment decision variables over all possible scenarios. For each scenario, these variables must be copied to define non-anticipativity constraints. Given that the first visit assignment decisions of the patients' first visit with revisit ($x^{\mathrm{I}}_{i_1jt}$) and without revisit ($x^{\mathrm{II}}_{i_2jt}$) are independent of each other, non-anticipativity constraints must be used for both. For the first visit assignment decision variables, these constraints are expressed as:

$$x^{\mathrm{I}}_{i_1jts}=x^{\mathrm{I}}_{i_1jt},\forall i_1\in\mathbb{I}_1,\, j\in\mathbb{J},\, t=k+1,\ldots,k+L,\, s\in\mathbb{S}, \tag{2a}$$

$$x^{\mathrm{II}}_{i_2jts}=x^{\mathrm{II}}_{i_2jt},\forall i_2\in\mathbb{I}_2,\, j\in\mathbb{J},\, t=k+1,\ldots,k+L,\, s\in\mathbb{S}, \tag{2b}$$

where the $x^{\mathrm{I}}_{i_1jt}$ and $x^{\mathrm{II}}_{i_2jt}$ represent the consensus variables. These non-anticipativity restraints are loosened before the PHA execution and penalty terms are added to the objective function for their violation. The objective function is modified by adding the following formula:

$$\begin{aligned}&\sum_{i_1=1}^{I_1}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}\sum_{s=1}^{S}\mu^{\mathrm{I}}_{i_1jts}\big(x^{\mathrm{I}}_{i_1jts}-x^{\mathrm{I}}_{i_1jt}\big)\\&+\frac{\rho_1}{2}\sum_{i_1=1}^{I_1}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}\sum_{s=1}^{S}\big\|x^{\mathrm{I}}_{i_1jts}-x^{\mathrm{I}}_{i_1jt}\big\|^2\\&+\sum_{i_2=1}^{I_2}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}\sum_{s=1}^{S}\mu^{\mathrm{II}}_{i_2jts}\big(x^{\mathrm{II}}_{i_2jts}-x^{\mathrm{II}}_{i_2jt}\big)\\&+\frac{\rho_2}{2}\sum_{i_2=1}^{I_2}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}\sum_{s=1}^{S}\big\|x^{\mathrm{II}}_{i_2jts}-x^{\mathrm{II}}_{i_2jt}\big\|^2,\end{aligned} \tag{2c}$$

where $\mu^{\mathrm{I}}_{i_1jts}$, $\mu^{\mathrm{II}}_{i_2jts}$, $\forall i_1,\, i_2,\, j,\, t,\, s$ indicate the Lagrangian multipliers; ρ_1 and ρ_2 represent the penalty parameters; and $\|\cdot\|$ denotes the ordinary Euclidean norm. Given that $x^{\mathrm{I}}_{i_1jts}$, $x^{\mathrm{I}}_{i_1jt}$, $x^{\mathrm{II}}_{i_2jts}$, $x^{\mathrm{II}}_{i_2jt}$ are binary variables, the penalty component in (2c) is reformulated as follows:

$$\big\|x^{\mathrm{I}}_{i_1jts}-x^{\mathrm{I}}_{i_1jt}\big\|^2=x^{\mathrm{I}}_{i_1jts}-2x^{\mathrm{I}}_{i_1jts}x^{\mathrm{I}}_{i_1jt}+x^{\mathrm{I}}_{i_1jt}, \tag{2d}$$

$$\big\|x^{\mathrm{II}}_{i_2jts}-x^{\mathrm{II}}_{i_2jt}\big\|^2=x^{\mathrm{II}}_{i_2jts}-2x^{\mathrm{II}}_{i_2jts}x^{\mathrm{II}}_{i_2jt}+x^{\mathrm{II}}_{i_2jt}. \tag{2e}$$

To obtain a scenario separable formulation, we estimate variables $x^{\mathrm{I}}_{i_1jt}$, $x^{\mathrm{II}}_{i_2jt}$ by $\hat{x}^{\mathrm{I}}_{i_1jt}$, $\hat{x}^{\mathrm{II}}_{i_2jt}$, which is equal to the weighted sum of $x^{\mathrm{I}}_{i_1jts}$, $x^{\mathrm{II}}_{i_2jts}$ by using a proximal point method [30]:

$$\hat{x}^{\mathrm{I}}_{i_1jt}=\sum_{s\in\mathbb{S}}p_sx^{\mathrm{I}}_{i_1jts},\hat{x}^{\mathrm{II}}_{i_2jt}=\sum_{s\in\mathbb{S}}p_sx^{\mathrm{II}}_{i_2jts}. \tag{2f}$$

Replacing $x^{\mathrm{I}}_{i_1jt}$, $x^{\mathrm{II}}_{i_2jt}$ in (2d) and (2e) with $\hat{x}^{\mathrm{I}}_{i_1jt}$, $\hat{x}^{\mathrm{II}}_{i_2jt}$ addresses the quadratic term of the objective function. The resulting structure facilitates the model to become decomposable, with each scenario associated with a subproblem.

The implement steps of the PHA are shown as follows.

Step 1: Initialization. $n=1$, $\rho_1^{(n)}=\rho_2^{(n)}=\rho^0$, $\mu^{\mathrm{I}(n)}_{i_1jts}=\mu^{\mathrm{II}(n)}_{i_2jts}=0$, $\forall i_1,\, i_2,\, j,\, t,\, s$.

Step 2: Solve each scenario-based subproblems to obtain $x^{\mathrm{I}(n)}_{i_1jts}$, $x^{\mathrm{II}(n)}_{i_2jts}$, $\forall i_1,\, i_2,\, j,\, t,\, s$.

Step 3: Calculation of the consensus parameter $\hat{x}^{\mathrm{I}}_{i_1jt}$, $\hat{x}^{\mathrm{II}}_{i_2jt}$, $\forall i_1,\, i_2,\, j,\, t,\, s$.

Step 4: Update the penalty parameter and the Lagrangian multiplier. $\rho_1^{(n+1)}=\beta_1\rho_1^{(n)}$, $\rho_2^{(n+1)}=\beta_2\rho_2^{(n)}$, where $\beta_1,\beta_2>0$, $\mu^{\mathrm{I}(n+1)}_{i_1jts}=\mu^{\mathrm{I}(n)}_{i_1jts}+\rho_1^{(n)}(x^{\mathrm{I}(n)}_{i_1jts}-\hat{x}^{\mathrm{I}}_{i_1jt})$, $\mu^{\mathrm{II}(n+1)}_{i_2jts}=\mu^{\mathrm{II}(n)}_{i_2jts}+\rho_2^{(n)}(x^{\mathrm{II}(n)}_{i_2jts}-\hat{x}^{\mathrm{II}}_{i_2jt})$.

Step 5: Repeat Steps 2-4 until the following termination criteria are satisfied or the maximum number of iterations are reached:

$$\sum_{s=1}^{S}p_s\sqrt{\sum_{i_1=1}^{I_1}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}(x^{\mathrm{I}(n)}_{i_1jts}-\hat{x}^{\mathrm{I}}_{i_1jt})^2}\le\varepsilon,$$

and

$$\sum_{s=1}^{S}p_s\sqrt{\sum_{i_2=1}^{I_2}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}(x^{\mathrm{II}(n)}_{i_2jts}-\hat{x}^{\mathrm{II}}_{i_2jt})^2}\le\varepsilon.$$

B. Modifications on Progressive Hedging Algorithm

In this section, we outline our penalty parameter and Lagrangian multiplier update methods. The penalty parameter update method used is similar to that in literature [31]. Let Δ_{E1}, Δ_{E2} denote the convergence status that calculates the sum of squares of the difference between the solution of subproblems and consensus parameters over all scenarios at consecutive iterations. The solutions among all scenario subproblems are highly different and consistent parameter values are harder to achieve as $\Delta_{E1}(\Delta_{E2})$ increases. In this context, we increase penalty parameter $\rho_1(\rho_2)$ by multiplying it with a constant $\beta_1(\beta_2)$ greater than 1. Otherwise, a factor of $1/\beta_1(1/\beta_2)$ is applied to reduce the penalty parameter. The penalty updating method is defined by (3a)-(3b):

$$\Delta_{E1}^{(n)} = \sum_{i_1=1}^{I_1}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}\sum_{s=1}^{S}\left(x_{i_1jts}^{\mathrm{I}(n)} - \hat{x}_{i_1jt}^{\mathrm{I}(n)}\right)^2,$$

$$\Delta_{E2}^{(n)} = \sum_{i_2=1}^{I_2}\sum_{j=1}^{J}\sum_{t=k+1}^{k+L}\sum_{s=1}^{S}\left(x_{i_2jts}^{\mathrm{II}(n)} - \hat{x}_{i_2jt}^{\mathrm{II}(n)}\right)^2. \quad (3a)$$

$$\rho_1^{(n+1)} = \begin{cases} \beta_1\rho_1^{(n)}, & \text{if } \Delta_{E1}^{(n)} - \Delta_{E1}^{(n-1)} > 0, \\ (1/\beta_1)\rho_1^{(n)}, & \text{otherwise.} \end{cases}$$

$$\rho_2^{(n+1)} = \begin{cases} \beta_2\rho_2^{(n)}, & \text{if } \Delta_{E2}^{(n)} - \Delta_{E2}^{(n-1)} > 0, \\ (1/\beta_2)\rho_2^{(n)}, & \text{otherwise.} \end{cases} \quad (3b)$$

We use a Lagrangian multiplier update method inspired by a previous approach [32], which aims to obtain consistent solutions to all subproblems. At this point, the consensus parameter converges to one of two variable values: 1 or 0. We determine the target convergence value between these two values by the majority of the solutions in all scenario subproblems. Two threshold parameters denoted by α_i, $i = 1, 2$ are defined to help identify whether the majority of conditions are met. When the consistency parameter is greater than α_i, then the majority of conditions hold. Next, the Lagrangian multipliers are updated on the basis of the following rule. If $\hat{x}_{i_1jt}^{\mathrm{I}(n)}(\hat{x}_{i_2jt}^{\mathrm{II}(n)})$ is greater than $\alpha_1(\alpha_2)$, then the majority of the scenario subproblem solutions assign FV with RV (FV without RV) patient $i_1(i_2)$ to physician j on day t. In this context, the Lagrangian multipliers in subproblems with 0 values of relevant variables are decreased. Otherwise, $\hat{x}_{i_1jt}^{\mathrm{I}(n)}(\hat{x}_{i_2jt}^{\mathrm{II}(n)})$ is no greater than $\alpha_1(\alpha_2)$, then FV with RV (FV without RV) patient $i_1(i_2)$ is not scheduled to physician j on day t in the majority of the scenarios. Subsequently, in subproblems with the relevant variables equal to 1, the Lagrangian multipliers are raised. The Lagrangian multiplier update method can be formulated as (3c) and (3d), as shown at the bottom of the page.

IV. Numerical Experiments

A. Data Settings

We collected data at our collaborator hospital in Shanghai, China. The hospital provides comprehensive medical services, both offline and online, by operating its own Internet hospital. Similar to the literature [7], [9], we assume that the patient demand follows Poisson distribution. The average daily numbers of patient FV with RV and FV without RV are 28.5 and 18.1, respectively. The MWTTs for patient FV with RV and FV without RV are set to 3 and 5, respectively. The arriving horizon is set as $K = 5$. The rolling horizon procedure is carried out as follows. On day k, the advancing model is solved for a planning horizon of L days. Then, the schedule generated for day $k+1$ is implemented. The new arising patient demands are collected and the information of available capacities is updated before solving the model on day $k+1$ for the next L days. The schedule of this new model on day $k+1$ is then put into action. The model is once more solved for the following L days once the appropriate updates have been made to capacity and requests. This pattern is repeated for the remaining phases of the arriving horizon. Minimum and maximum revisit intervals of offline (online) revisits are set to 2(4) and 3(5), respectively. Therefore, the planning horizon is set to 12 days.

Four physicians are set up with offline and online medical services each day with a daily capacity of 480 minutes. During the first four days of the planning horizon, each physician has no initial available online capacity because it was occupied by previously scheduled revisit patients. The online capacities of each physician are initially available for 60 min on the following day. In the first two days of the planning horizon, the offline capacity initially available to each physician is 240 min because it is occupied by the pre-scheduled first visit or revisit patients. For the rest of the day, the total capacity minus the online capacity represents each physician's initial available offline capacity. The service duration for a FV with RV, online revisit, offline revisit and FV without RV is 60, 30, 40, and 40 minutes, respectively. We randomly generate the scenarios of patients' revisit status by setting the realized vector i-th component to one with a probability equal to 0.6. We set continuity of care violation penalties c_p^o, c_p^e as 1.5 and 1 for offline and online patients. The offline and online overtime or idle costs are set to 1.5 and 1, respectively. The fixed offline and online service setup cost c_u^o and c_u^e are set to 1.5 and 1, respectively. These settings represent our base case on which numerical experiments are implemented.

B. Algorithm Performance

In this experiment, we use five methods as benchmarks to evaluate the performance of the proposed PHA. The first method is to solve the SAA by using the standard branch-and-cut solution algorithm implemented in GUROBI. We propose the other four Lagrangian decomposition algorithms based on Lagrangian relaxation, which make it possible to tackle large-scale stochastic optimization problems by decomposing them into scenario subproblems [33].

$$\mu_{i_1jts}^{\mathrm{I}(n+1)} = \begin{cases} \mu_{i_1jts}^{\mathrm{I}(n)} + \rho_1^{(n)}\left(x_{i_1jts}^{\mathrm{I}(n)} - \hat{x}_{i_1jt}^{\mathrm{I}(n)}\right), & \text{if } \hat{x}_{i_1jt}^{\mathrm{I}(n)} < \alpha_1, x_{i_1jts}^{\mathrm{I}(n)} = 1, \\ \mu_{i_1jts}^{\mathrm{I}(n)} - \rho_1^{(n)}\left|x_{i_1jts}^{\mathrm{I}(n)} - \hat{x}_{i_1jt}^{\mathrm{I}(n)}\right|, & \text{if } \hat{x}_{i_1jt}^{\mathrm{I}(n)} > \alpha_1, x_{i_1jts}^{\mathrm{I}(n)} = 0. \end{cases} \quad (3c)$$

$$\mu_{i_2jts}^{\mathrm{II}(n+1)} = \begin{cases} \mu_{i_2jts}^{\mathrm{II}(n)} + \rho_2^{(n)}\left(x_{i_2jts}^{\mathrm{II}(n)} - \hat{x}_{i_2jt}^{\mathrm{II}(n)}\right), & \text{if } \hat{x}_{i_2jt}^{\mathrm{II}(n)} < \alpha_2, x_{i_2jts}^{\mathrm{II}(n)} = 1, \\ \mu_{i_2jts}^{\mathrm{II}(n)} - \rho_2^{(n)}\left|x_{i_2jts}^{\mathrm{II}(n)} - \hat{x}_{i_2jt}^{\mathrm{II}(n)}\right|, & \text{if } \hat{x}_{i_2jt}^{\mathrm{II}(n)} > \alpha_2, x_{i_2jts}^{\mathrm{II}(n)} = 0. \end{cases} \quad (3d)$$

Algorithm 1 Lagrangian Decomposition Embedded With Subgradient Method

Step 1. Initialize:
Lower and upper bounds, UB = ∞, LB = -∞;
Lagrangian multipliers, ρ^{I}, ρ^{II};
Iteration counter, $n = 0$, $r = 0$.
Step 2. Solving Lagrangian dual problem:
Solve scenario subproblems and sum subproblem objective values to get the lower bound LB_n; Update LB: if $\mathrm{LB}_n > \mathrm{LB}$, then $\mathrm{LB} = \mathrm{LB}_n$.
Step 3. Update UB:
Generate solutions x^{I}, x^{II} using the proposed heuristic rule;
For each patient, calculate how many scenarios the optimal assignment decision is physician j, and select the physician j that is the optimal allocation in most scenarios to form a feasible solution. The appointment day for the patient will be the earliest among the scenarios where the physician-patient allocation is decided;
Obtain the original problem objective value Obj_n and evaluate the solution feasibility. If x^{I} and x^{II} are not feasible, add an integer cut to exclude this infeasible solution, return to Step 2, and set $r = r + 1$;
If x^{I}, x^{II} are feasible and $Obj_n < \mathrm{UB}$, then $\mathrm{UB} = Obj_n$;
If x^{I}, x^{II} are not feasible and $r = r_{max}$, then use solutions x^{I}, x^{II} as input parameters, solve each scenario subproblem and obtain objective value Obj_n.
Step 4. Update Lagrangian multipliers using the subgradient method:
$\rho_{n+1}^{\mathrm{I}s} = \rho_n^{\mathrm{I}s} + \theta_n \frac{(UB-LB_n)}{\sum_s (v_{sn}^{\mathrm{I}})^2}$, $\rho_{n+1}^{\mathrm{II}s} = \rho_n^{\mathrm{II}s} + \theta_n \frac{(UB-LB_n)}{\sum_s (v_{sn}^{\mathrm{II}})^2}$, $\forall s \in S$,
where $v_{sn}^{\mathrm{I}} = x_{i_1 jts}^{\mathrm{I}} - x_{i_1 jt,s+1}^{\mathrm{I}}$, $v_{sn}^{\mathrm{II}} = x_{i_2 jts}^{\mathrm{II}} - x_{i_2 jt,s+1}^{\mathrm{II}}$, $s \in \mathbb{S} \setminus S$ pertaining to the sequential representation, $v_{sn}^{\mathrm{I}} = x_{i_1 jt1}^{\mathrm{I}} - x_{i_1 jts}^{\mathrm{I}}$, $v_{sn}^{\mathrm{II}} = x_{i_2 jt1}^{\mathrm{II}} - x_{i_2 jts}^{\mathrm{II}}$, $s \in \mathbb{S} \setminus 1$ pertaining to the asymmetric representation. The term $\theta_n \in (0,\ 2]$ is a step length parameter. By correcting the error in the estimation of the true optimal value, the term is used to change the step size's length.
Step 5. Stop and return x^{I}, x^{II} and UB, if UB-LB $< \varepsilon$ or any other condition is satisfied, such as the amount of running time or the number of iterations. If not, set $n = n + 1$ and go to Step 2.

We combine the subgradient method [48] and surrogate subgradient method [49], the most popular or state-of-the-art technique for obtaining solutions to the Lagrangian dual, with two different formulations for the NAC constraints in (2a) and (2b) (specifically the sequential formulation as given in (4a) and asymmetric formulation provided in (4b) [39]). In this context, we call the second method LDSS which is the combination of the subgradient method and the sequential NAC formulation, and the third method LDSA is composed of subgradient method and the asymmetric NAC formulation. The fourth and fifth methods are named as LDSSS and LDSSA, which are the combination of surrogate subgradient algorithm with sequential and asymmetric NAC formulation respectively. The penalty terms $v_s^{\mathrm{I}} = x_{i_1 jts}^{\mathrm{I}} - x_{i_1 jt,s+1}^{\mathrm{I}}$, $v_s^{\mathrm{II}} = x_{i_2 jts}^{\mathrm{II}} - x_{i_2 jt,s+1}^{\mathrm{II}}$, $s \in \mathbb{S} \setminus S$ associate with sequential NAC (4a). $v_s^{\mathrm{I}} = x_{i_1 jt1}^{\mathrm{I}} - x_{i_1 jts}^{\mathrm{I}}$, $v_s^{\mathrm{II}} = x_{i_2 jt1}^{\mathrm{II}} - x_{i_2 jts}^{\mathrm{II}}$, $s \in \mathbb{S} \setminus 1$ represent the penalty term associated with asymmetric NAC (4b). The compact formulations for Lagrangian relaxation, the Lagrangian dual problem, and scenario subproblem with respect to sequential and asymmetric NAC constraints can be found in [39]. The steps for LDSS, LDSA, LDSSS, and LDSSA these four Lagrangian decomposition algorithms are summarized in Algorithm 1 and Algorithm 2 (see more specific technical details of the algorithm in [39], [49]).

$$x_{i_1 jts}^{\mathrm{I}} = x_{i_1 jt,s+1}^{\mathrm{I}}, x_{i_2 jts}^{\mathrm{II}} = x_{i_2 jt,s+1}^{\mathrm{II}},$$
$$\forall i_1 \in \mathbb{I}_1, j \in \mathbb{J}, t = k+1, \ldots, k+L, s \in \mathbb{S} \setminus S. \tag{4a}$$
$$x_{i_1 jt1}^{\mathrm{I}} = x_{i_1 jt,s}^{\mathrm{I}}, x_{i_2 jt1}^{\mathrm{II}} = x_{i_2 jts}^{\mathrm{II}},$$
$$\forall i_2 \in \mathbb{I}_2, j \in \mathbb{J}, t = k+1, \ldots, k+L, s \in \mathbb{S} \setminus 1. \tag{4b}$$

To evaluate the efficiency of the proposed PHA, we design the testing instances as combinations of different levels of patient demands, probability of online revisit, and number of scenarios. Without loss of generality, we explore three groups of instances with rising patient volume: Low Traffic Case (daily arrival rate that is 0.5 times the base level), Medium Traffic case (base level daily arriving rate), and High Traffic Case (daily arriving rate that is 1.5 times the base level), which are common scenarios in the investigated hospital. We generate the possible scenarios based on a binomial distribution. In this advance scheduling context, we consider three typical p_{i_1} for the probability of online revisit based on history data, where p_{i_1}= 0.35, 0.6, or 0.85. The number of scenarios is drawn from [90, 3000]. Each instance group includes nine instances with different probabilities of online revisit and number of scenarios. As a result, 27 test instances in total are generated. Each instance is denoted by L/M/HT $- p_{i_1} - |\mathbb{S}|$, where $|\mathbb{S}|$ is the number of scenarios. Considering the "online scheduling" settings, the algorithm needs to be run repeatedly within a period and to obtain the optimal decision in a short time for each running to achieve real-time scheduling. In this context, we set the time restriction to 2880 seconds for each decision running using GUROBI, so the total time limitation is 14400s (e.g., 4h) for five decision runnings in each instance. The maximum number of iterations of the five methods PHA, LDSS, LDSA, LDSSS, and LDSSA are set to be the same to ensure the comparability of algorithm performance and to prevent pointless calculations. All experiments are carried out on a computer with a 2.90 GHz Intel Core i7-10700 CPU and 16 GB of memory. All models are solved by Gurobi 9.0.2 on Python 3.7.

Algorithm 2 Lagrangian Decomposition Embedded With Surrogate Subgradient Method

Step 1. Initialize lower and upper bounds, Lagrangian multipliers $\rho_0^{\mathrm{I}s}$, and $\rho_0^{\mathrm{II}s}$, and iteration counter $n = 1$, $r = 0$. Given the initial Lagrangian multipliers, solve scenario subproblems to obtain x_0^{I} and x_0^{II}.
Step 2. Set the surrogate dual to the dual and set the surrogate subgradient to the subgradien. Update Lagrangian multipliers using the subgradient information.
$\rho_1^{\mathrm{I}s} = \rho_0^{\mathrm{I}s} + \theta_n \frac{(UB-LB_n)}{\sum_s (v_{sn}^{\mathrm{I}})^2} v_{sn}^{\mathrm{I}}$, $\rho_1^{\mathrm{II}s} = \rho_0^{\mathrm{II}s} + \theta_n \frac{(UB-LB_n)}{\sum_s (v_{sn}^{\mathrm{II}})^2} v_{sn}^{\mathrm{II}}$, $\forall s \in S$.
Step 3. Update LB and UB using the same algorithmic logic of Step 2 and Step 3 in **Algorithm 1**.
Step 4. Update Lagrangian multipliers using the surrogate subgradient method:
Examine whether the surrogate optimality condition holds. Compute the objective value $L(\boldsymbol{\rho}_{n+1}^{\mathrm{I}}, \boldsymbol{\rho}_{n+1}^{\mathrm{II}}, \boldsymbol{x}_{n+1}^{\mathrm{I}}, \boldsymbol{x}_{n+1}^{\mathrm{II}})$ given $\boldsymbol{\rho}_{n+1}^{\mathrm{I}}$, $\boldsymbol{\rho}_{n+1}^{\mathrm{II}}$, $\boldsymbol{x}_{n+1}^{\mathrm{I}}$, $\boldsymbol{x}_{n+1}^{\mathrm{II}}$ and the objective value $L(\boldsymbol{\rho}_{n+1}^{\mathrm{I}}, \boldsymbol{\rho}_{n+1}^{\mathrm{II}}, \boldsymbol{x}_n^{\mathrm{I}}, \boldsymbol{x}_n^{\mathrm{II}})$ given $\boldsymbol{\rho}_{n+1}^{\mathrm{I}}$, $\boldsymbol{\rho}_{n+1}^{\mathrm{II}}$, $\boldsymbol{x}_n^{\mathrm{I}}$, $\boldsymbol{x}_n^{\mathrm{II}}$.
If $L(\boldsymbol{\rho}_{n+1}^{\mathrm{I}}, \boldsymbol{\rho}_{n+1}^{\mathrm{II}}, \boldsymbol{x}_{n+1}^{\mathrm{I}}, \boldsymbol{x}_{n+1}^{\mathrm{II}}) \geq L(\boldsymbol{\rho}_{n+1}^{\mathrm{I}}, \boldsymbol{\rho}_{n+1}^{\mathrm{II}}, \boldsymbol{x}_n^{\mathrm{I}}, \boldsymbol{x}_n^{\mathrm{II}})$, then the surrogate optimality condition does not hold and let $\boldsymbol{x}_{n+1}^{\mathrm{I}} = \boldsymbol{x}_n^{\mathrm{I}}$, $\boldsymbol{x}_{n+1}^{\mathrm{II}} = \boldsymbol{x}_n^{\mathrm{II}}$.
Retrieve the new set of Lagrangian multipliers.
$\boldsymbol{\rho}_{n+1}^{\mathrm{I}s} = \boldsymbol{\rho}_n^{\mathrm{I}s} + \theta_n \frac{(UB-LB_n)}{\sum_s (v_{sn}^{\mathrm{I}})^2} v_{sn}^{\mathrm{I}}$, $\boldsymbol{\rho}_{n+1}^{\mathrm{II}s} = \boldsymbol{\rho}_n^{\mathrm{II}s} + \theta_n \frac{(UB-LB_n)}{\sum_s (v_{sn}^{\mathrm{II}})^2} v_{sn}^{\mathrm{II}}$, $\forall s \in S$.
Step 5. Examine whether the termination criteria hold. If the termination criteria are satisfied, stop and return x^{I}, x^{II} and UB; otherwise, set $n = n + 1$ and go to Step 3.

Table II indicates the performances of PHA and benchmark methods for the proposed model with different datasets. Compared with the exact method by Gurobi 9.0.2, the PHA can output the satisfactory solution very quickly. The PHA can solve most of instances in a few hundred seconds instead of thousands of seconds by Gurobi. The proposed PHA can solve the problem with an average CPU time of 3940.4s, while the average CPU time taken by Gurobi is 15950.7s. The results where the cost value in the "PHA-Costs" column is less than that of in the "Gurobi SAA-Costs" column show that the solutions generated by our PHA are superior to those generated by Gurobi. This superior performance is noticeable in 11 cases. Compared with the Lagrangian decomposition algorithms embedded with the subgradient method, the performance of the PHA is significantly and consistently superior to that of the LDSS and LDSA methods in terms of both the quality of the solution and CPU time. Table III reports the gap between the solution value generated by the LDSSS and LDSSA (Obj_{LDSSS}, Obj_{LDSSA}) and the solution value generated by the PHA (Obj_{PHA}). The gap is computed as follows: $GAP_{PHA-LDSSS} = 100 * (Obj_{PHA} - Obj_{LDSSS})/Obj_{LDSSS}$, $GAP_{PHA-LDSSA} = 100 * (Obj_{PHA} - Obj_{LDSSA})/Obj_{LDSSA}$. In comparison to LDSSS and LDSSA, the Lagrangian decomposition methods embedded with the surrogate subgradient algorithm, the PHA can generate equally good solutions with an overall average gap of 3.60% and 3.35% over all 27 instances, respectively. It is worth noting that our algorithm is quite fast in solving the problem, taking less computation time than the LDSSS and LDSSA in each instance. Comparing the two methods LDSS and LDSA, we can find that the solution quality of LDSA with asymmetric NAC is better, but the solution time is longer than that of LDSS with sequential NAC. LDSSS and LDSSA perform equally well in solution value, while LDSSA with asymmetric NAC takes less running time than LDSSS with sequential NAC in most of instances. The comparison results of Columns "LDSS" and "LDSA" and Columns "LDSSS" and "LDSSA" indicate that the surrogate subgradient technique embedded within Lagrangian relaxation can produce better solutions with a longer time than subgradient technique embedded within Lagrangian relaxation.

As the problem scale rises with an increasing number of patients and scenarios, the running time of the PHA, Gurobi, LDSS, LDSA, LDSSS, and LDSSA all increase, and Gurobi has the highest growth rate among these methods. The use of Gurobi for Medium and High Traffic instances with 2000 and above scenarios is not feasible because of limitations in memory. Due to time limitations, it is not possible to use LDSS, LDSA, LDSSS and LDSSA for solving Medium Traffic instances with 2000 and above scenarios. High Traffic instances with 1000 scenarios cannot be solved within 4 hours by using LDSS, LDSA, LDSSS and LDSSA. Even feasible solutions to the much larger instances cannot be generated by the commercial solver Gurobi or the other four Lagrangian decomposition based algorithms. For the majority of cases, the PHA yields good solutions within 20 minutes. The algorithm scales up incredibly well and can solve High Traffic instances up to 3000 scenarios with near-optimal performance in an acceptable amount of time. Overall, the aforementioned facts lead us to the conclusion that the PHA is a powerful technique by obtaining good feasible solutions efficiently, and our algorithm's improvement effect becomes more pronounced when the problem scale is larger. The significant benefits demonstrate that the proposed PHA is an effective algorithm that can meet the needs of real-world applications.

C. Comparisons of Capacity Allocation Schemes Under Different Traffic Volumes

This section first presents the solution of the hybrid online and offline advance scheduling problem for the base case. The total costs are 10845.17. Fig. 2 indicates the results of each physician' service capacity allocation to first visit and revisit patients for each day in the planning horizon.

In our discussion, the hospital practitioners stated their desire to find an easy-to-implement capacity allocation method in the current hybrid online and offline system with first visit and revisit patients. Therefore, as shown in Fig. 3, we consider

TABLE II

COMPARISON RESULTS OF PHA AND PROPOSED BENCHMARKS

Instance	Gurobi SAA		PHA		LDSS		LDSA		LDSSS		LDSSA	
	Costs	Time (s)	Costs	Time (s)	Costs	Time (s)	Costs	Time (s)	Costs	Time (s)	Costs	Time (s)
LT-0.35-140	17952.3	56.1	17939.7	304.6	22781.3	2042.1	18094.2	2244.5	17936.0	2394.5	17938.8	2563.2
LT-0.35-190	17950.8	85.0	17936.0	429.4	22131.3	2827.7	18332.9	3268.3	17931.3	3123.4	17934.0	3516.8
LT-0.35-240	17924.1	141.4	17915.8	523.8	22175.2	3719.7	18080.1	4225.6	17911.8	4177.3	17914.0	4389.5
LT-0.6-150	18514.7	510.0	18472.1	332.3	22823.9	2208.5	18484.0	2616.4	18445.7	2772.1	18440.2	2715.0
LT-0.6-200	18539.3	1382.6	18499.2	447.1	22113.5	3049.9	19295.4	3379.7	18454.2	3746.2	18463.2	3588.9
LT-0.6-250	18550.8	4007.5	18522.0	571.1	21875.3	3770.4	18628.1	4182.4	18460.2	4492.3	18468.2	4772.5
LT-0.85-100	19406.3	1355.3	19328.0	233.1	22773.7	1424.4	19191.1	1764.8	19226.6	1840.9	19219.1	1950.3
LT-0.85-150	19463.4	6534.1	19393.3	355.9	24044.8	2224.9	19585.0	2628.9	19239.4	2685.5	19256.8	2846.8
LT-0.85-200	19460.0	8362.7	19395.4	495.0	23109.3	3044.6	20103.0	3565.1	19286.1	3788.6	19281.4	3864.4
MT-0.35-150	8807.2	5962.3	10620.3	730.0	25900.6	3996.5	12182.4	4759.4	10409.6	5229.1	10680.9	5442.8
MT-0.35-190	9401.2	6175.0	11476.7	915.7	26053.3	5451.0	11978.0	6400.5	10330.8	6985.1	10178.3	7279.3
MT-0.35-240	9523.3	6415.5	10978.1	1090.3	26510.2	7000.8	11568.8	7876.0	10949.7	9106.0	10793.5	10676.4
MT-0.6-200	9309.6	10222.0	11484.2	902.8	27672.5	6299.8	11487.2	7138.6	9827.5	7339.5	10053.8	7190.3
MT-0.6-250	10555.7	12001.7	10657.6	1209.8	27972.9	7136.6	10677.3	11115.3	9660.9	10703.9	9981.9	10309.6
MT-0.6-2000	–	–	11873.5	10487.6	–	–	–	–	–	–	–	–
MT-0.85-300	11526.6	1581.2	11657.5	1232.0	28977.4	8613.6	13379.4	10466.1	11254.1	12239.3	11278.0	9973.5
MT-0.85-500	11210.3	7127.5	12173.5	2077.2	29649.0	16640.4	13755.7	16359.1	11225.6	19375.8	11204.9	19948.5
MT-0.85-2500	–	–	14514.4	11299.5	–	–	–	–	–	–	–	–
HT-0.35-90	14930.0	5972.9	15743.7	550.7	34717.4	3779.5	17839.0	4374.2	15206.2	5394.5	15214.7	5193.1
HT-0.35-140	15015.7	6536.2	15819.9	866.5	35329.4	6167.4	18489.8	7009.9	15177.2	7795.6	15193.9	7146.6
HT-0.35-190	15124.0	11223.5	16010.4	1166.3	35300.1	8472.5	18316.3	8618.1	15196.6	12118.9	15201.2	11786.4
HT-0.6-1500	27375.4	15546.6	20506.9	9288.0	–	–	–	–	–	–	–	–
HT-0.6-2000	–	–	18493.7	15092.3	–	–	–	–	–	–	–	–
HT-0.6-2500	–	–	19760.1	17476.2	–	–	–	–	–	–	–	–
HT-0.85-500	19826.5	9412.8	18466.7	2961.1	39095.0	24210.8	19587.2	24589.0	17816.7	28315.0	17655.2	27554.1
HT-0.85-1000	20549.9	12826.4	20830.9	5744.3	39097.9	52184.6	19223.0	49463.3	–	–	–	–
HT-0.85-3000	–	–	21483.8	19608.4	–	–	–	–	–	–	–	–

TABLE III

COMPARISON RESULTS OF PHA, LDSSS AND LDSSA

Instance	LT-0.35-140	LT-0.35-190	LT-0.35-240	LT-0.6-150	LT-0.6-200	LT-0.6-250	LT-0.85-100	LT-0.85-150	LT-0.85-200
$Gap_{PHA-LDSSS}$	0.02%	0.03%	0.02%	0.14%	0.24%	0.33%	0.53%	0.80%	0.57%
$Gap_{PHA-LDSSA}$	0.01%	0.01%	0.01%	0.17%	0.19%	0.29%	0.57%	0.71%	0.59%
Instance	MT-0.35-150	MT-0.35-190	MT-0.35-240	MT-0.6-200	MT-0.6-250	MT-0.6-2000	MT-0.85-300	MT-0.85-500	MT-0.85-2500
$Gap_{PHA-LDSSS}$	2.02%	11.09%	0.26%	16.86%	10.32%	-	3.58%	8.44%	-
$Gap_{PHA-LDSSA}$	-0.57%	12.76%	1.71%	14.23%	6.77%	-	3.36%	8.64%	-
Instance	HT-0.35-90	HT-0.35-140	HT-0.35-190	HT-0.6-1500	HT-0.6-2000	HT-0.6-2500	HT-0.85-500	HT-0.85-1000	HT-0.85-3000
$Gap_{PHA-LDSSS}$	3.53%	4.23%	5.36%	-	-	-	3.65%	-	-
$Gap_{PHA-LDSSA}$	3.48%	4.12%	5.32%	-	-	-	4.60%	-	-

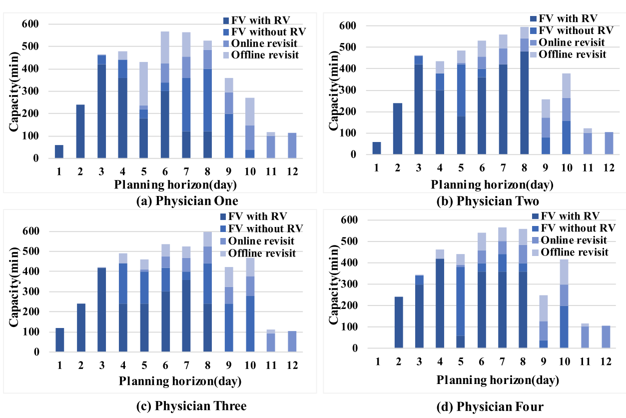

Fig. 2. Advance scheduling results for the base case.

four types of capacity allocation schemes with different numbers of physicians who are assigned with both offline and online capacities: Allocation I (four physicians with both offline and online capacities assigned), Allocation II (three physicians with both offline and online capacities assigned), Allocation III (two physicians with both offline and online capacities assigned), and Allocation IV (one physician with both offline and online capacities assigned). Furthermore, three scales of patient demand: Low Traffic Case, Medium Traffic case, and High Traffic Case as described before are considered. Other parameters are set unchanged as basic levels. Through these cases, we investigate how these four allocation schemes compare in performance under different patient demands.

Table IV reports the cost results for three situations (Low Traffic, Medium Traffic, and High Traffic). Increasing patient demand leads to a rise in offline overtime costs C_{Oo} and a decrease in idle cost C_{ao}. These observations are intuitive and hold in online capacity. As traffic volume increases, offline overtime and idle costs C_o decrease first and then increase, while online overtime and idle costs C_e continue increasing. The change in total cost TC is the same as that of C_o. That is, traffic volume has more sensitive effects on offline capacity than on online capacity, playing a lead role in the change in total cost.

In the low traffic case, Allocation IV slightly performs better compared with Allocation I, although the difference is not statistically significant. With increasing traffic volume, Allocation I consistently outperforms all other allocation schemes. Given the fact that more patients in larger systems are subject to schedule planning, Allocation I is more likely to improve the performance by implementing more intelligent approaches to capacity allocation and patient scheduling.

TABLE IV
COMPARISON RESULTS OF ALLOCATION SCHEMES UNDER DIFFERENT TRAFFIC VOLUMES

Traffic	Allocation	C_{Oo}	C_{ao}	C_o	C_{Oe}	C_{ae}	C_e	$C_P(C_{Pe}, C_{Po}=0)$	TC
Low traffic	I	0.00	17833.19	17833.19	47.58	670.98	718.56	33.73	18585.48
	II	4.80	17837.99	17842.79	68.64	692.04	760.69	30.74	18634.22
	III	0.00	17833.19	17833.19	59.90	683.30	743.20	35.01	18611.40
	IV	15.60	17848.79	17864.39	17.68	641.08	658.78	30.38	18553.55
Medium traffic	I	1293.82	8433.82	9727.63	833.36	233.36	1066.72	50.82	10845.17
	II	1808.88	8948.88	10757.76	769.88	169.88	939.76	56.54	11754.06
	III	2019.60	9159.60	11179.20	774.69	174.68	949.37	55.26	12183.83
	IV	1975.46	9115.46	11090.93	830.06	230.06	1060.13	57.42	12208.48
High traffic	I	8678.40	4814.40	13492.80	1925.08	47.08	1972.17	93.60	15558.57
	II	8709.60	4845.60	13555.20	1946.55	68.55	2015.10	94.10	15664.40
	III	8821.19	4957.20	13778.39	1932.88	54.88	1987.77	93.80	15859.96
	IV	8846.40	4982.40	13828.80	1918.84	40.84	1959.67	96.58	15885.05

TABLE V
SENSITIVITY ANALYSIS RESULTS OF COST COEFFICIENTS

Num. of cases	Cost type	Cost coefficient	C_o	C_e	$C_P(C_{Pe}, C_{Po}=0)$	TC
1	c_b^o	0.1	659.76	872.86	49.37	1581.99
2		0.5	3480.39	969.33	49.90	4499.62
3		1	6717.82	1033.93	55.56	7807.31
4		2	13305.59	1037.70	57.26	14400.55
5		5	33571.03	987.09	46.01	34604.13
6	c_b^e	0.1	6828.22	104.37	53.74	6986.33
7		0.5	5678.39	544.76	49.26	6272.41
8		1	6717.82	1033.93	55.56	7807.31
9		2	5809.59	1814.35	47.78	7671.72
10		5	6312.79	4399.68	49.06	10761.53
11	c_p^o	0.1	7163.20	1020.64	53.96	8237.80
12		0.5	7152.12	1030.37	53.72	8236.21
13		1	6717.82	1033.93	55.56	7807.31
14		2	6883.19	997.36	54.38	7934.93
15		5	6308.67	1014.07	52.36	7375.10
16	c_p^e	0.1	6568.79	1041.79	4.72	7615.30
17		0.5	6794.04	1032.74	27.69	7854.47
18		1	6717.82	1033.93	55.56	7807.31
19		2	7507.19	1026.29	97.20	8630.68
20		5	6549.34	1038.38	265.20	7852.92

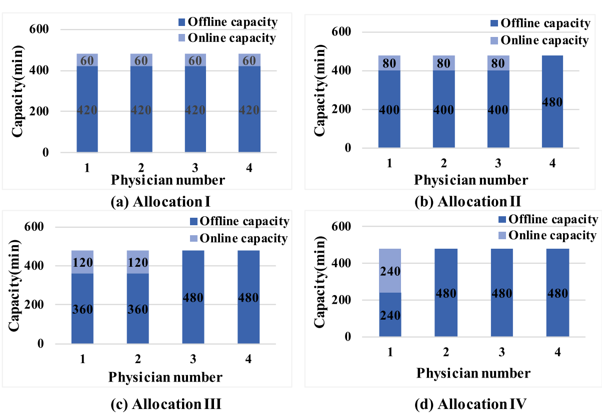

Fig. 3. Four types of allocation schemes.

D. Sensitivity Analysis of Cost Coefficients

In this section, we investigate the impact of the cost coefficients. We vary the unit costs of physician offline and online overtime or idle, continuity of care violation penalty of offline and online revisit separately and sequentially while fixing other cost coefficients as 1. Table V shows the system performance results including overtime and idle cost offline (C_o) and online (C_e), penalty of offline (C_{Po}) and online revisit (C_{Pe}), total penalty cost (C_P), and total cost (TC). A larger unit cost of c_b^o, c_b^e or c_p^e leads to an increase in the expected total cost. The results are intuitive. However, increases in unit cost of c_p^o lead to a decreasing trend in expected total costs. This is because the C_{Po} is always 0 regardless of the value of c_p^o. As c_p^o increases, offline overtime and idle cost decreases and results in a downward trend in the total costs. We can also note that the system performances are more sensitive to c_b^o and c_b^e than penalty cost of continuity of care violation c_p^o and c_p^e. In addition, both offline and online overtime and idle time show an upward trend with the increase of c_b^o. Conversely, increasing c_b^e makes both offline and online overtime and idle time tend to decrease.

E. Practical Implications

From the above observations, we can assume the value of introducing different capacity allocation schemes, which have a significant influence on the system performance. When the patient volume is large, the model suggests to apply Allocation I with all physicians assigned with both offline and online capacities to reduce the total costs. Another insight from the results is that hospitals must carefully and appropriately decide the parameters, especially the unit offline and online overtime or idle cost, according to their impact. We can suggest to hospital managers how to select parameters based on their purposes. For example, to obtain shorter offline overtime and idle time, we recommend practitioners to set a smaller c_b^o, a larger c_b^e, or a larger c_p^o. A smaller c_b^o or a larger c_b^e can also be adopted to achieve less online overtime and idle time.

F. Discussions

This study sets the foundations for the application of healthcare intelligence and digitalization. We provide intelligent decision-making techniques for the advance scheduling of chronic disease patients in integrated offline and online medical service systems. These efforts can improve the efficiency and effectiveness of chronic disease management. Although we attempt to formulate the model to be as realistic as possible, several limitations unavoidably occur and provide relevant interesting research directions. First, practical problems may occur such as no-shows, cancellations, and rescheduling, which can be addressed as follows. The clinic may overbook patients by "inflating" the typical capacity. That is, let us assume that the real effective regular capacity is q and the average no-show rate is p; then to handle no-shows, the clinic assumes they can serve a larger patient volume with a regular capacity of $q/(1-p)$. For rescheduling of cancellations or no-shows, the missed appointments on day k are then reviewed for scheduling on day $k+1$ together with new arising demands. Second, the sequencing and timing of each patient within the same day are not considered in the advance scheduling problem. One research direction is to consider physician brain cognitive fatigue state changes due to switching between online and offline in the joint model of advance and appointment scheduling. That is, suppose that a fixed or time-dependent switching cost is generated as a result of one working style conversion. Finally, the case where revisits have a higher priority to be admitted than the first visit patients when the service provider has an option of rejecting patients is worthy of further exploration.

V. Conclusion

In this study, we investigate the advance scheduling problem for managing first visits and revisits with chronic diseases while considering online and offline revisit uncertainty and heterogeneity. We obtain a patient schedule that minimizes the total cost, which includes offline and online overtime and idle costs and the continuity of care violation penalty of offline and online revisits. To schedule patients for a planning horizon at a particular decision moment, we develop a stochastic mixed integer programming model. Then, to obtain an effective solution, we reformulate the problem structure to be decomposable into scenario-based subproblems and solved by PHA. Next, the modified PHA with the penalty update and Lagrangian multiplier update is proposed. Finally, for rolling time advance scheduling, we use a sequential decision-making architecture that includes the stochastic programming model and modified PHA. The effectiveness of the proposed algorithm and the healthcare system is verified by the numerical experiments, which reveal that our approach has superior solving ability than the commercial solver Gurobi and the published state-of-the-art Lagrangian relaxation based methods. The numerical results also provide managerial insights. For example, a capacity allocation scheme with all physicians assigned with both offline and online capacities is a good choice to reduce costs. In the future, further investigations can focus on actual constraints of physician workload balance requirements and consider human factor costs caused by switching between online and offline work states.

References

[1] *Zhiyun Health "Internet + Medical Care" Empowers Chronic Disease Management and Unleashes More Power to Benefit the People*. Accessed: Sep. 5, 2022. [Online]. Available: http://www.cnfina.com/kuaixun/20220905_200337.html

[2] *Give Full Play to the Role of Internet Hospitals and Relieve the Contradiction Between Supply and Demand for Medical Treatment*. Accessed: Apr. 25, 2022. [Online]. Available: https://mp.weixin.qq.com/s/d5rYJG1MSlcgzHtjptWp8w

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

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

[5] H.-J. Schütz and R. Kolisch, "Approximate dynamic programming for capacity allocation in the service industry," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 239–250, Apr. 2012.

[6] V.-A. Truong, "Optimal advance scheduling," *Manage. Sci.*, vol. 61, no. 7, pp. 1584–1597, Jul. 2015.

[7] L. Zhou, N. Geng, Z. Jiang, and X. Wang, "Dynamic multi-type patient advance scheduling for a diagnostic facility considering heterogeneous waiting time targets- and equity," *IISE Trans.*, vol. 54, no. 6, pp. 521–536, 2022.

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

[9] S. Yu, V. G. Kulkarni, and V. Deshpande, "Appointment scheduling for a health care facility with series patients," *Prod. Oper. Manage.*, vol. 29, no. 2, pp. 388–409, Feb. 2020.

[10] H. Zhu, M. Hou, C. Wang, and M. Zhou, "An efficient outpatient scheduling approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 9, no. 4, pp. 701–709, Oct. 2012.

[11] X. Xie, Z. Fan, and X. Zhong, "Appointment capacity planning with overbooking for outpatient clinics with patient no-shows," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 864–883, Apr. 2022.

[12] J. Song, Y. Bai, and J. Wen, "Optimal appointment rule design in an outpatient department," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 1, pp. 100–114, Jan. 2019.

[13] J. Dai, N. Geng, and X. Xie, "Dynamic advance scheduling of outpatient appointments in a moving booking window," *Eur. J. Oper. Res.*, vol. 292, no. 2, Jul. 2021.

[14] D. Astaraky and J. Patrick, "A simulation based approximate dynamic programming approach to multi-class, multi-resource surgical scheduling," *Eur. J. Oper. Res.*, vol. 245, no. 1, pp. 309–319, Aug. 2015.

[15] A. Bansal, B. Berg, and Y.-L. Huang, "A distributionally robust optimization approach for coordinating clinical and surgical appointments," *IISE Trans.*, vol. 53, no. 12, pp. 1311–1323, 2021.

[16] W. Vancroonenburg, P. Smet, and G. V. Berghe, "A two-phase heuristic approach to multi-day surgical case scheduling considering generalized resource constraints," *Oper. Res. Health Care*, vol. 7, pp. 27–39, Dec. 2015.

[17] M. A. Kamran, B. Karimi, and N. Dellaert, "Uncertainty in advance scheduling problem in operating room planning," *Comput. Ind. Eng.*, vol. 126, pp. 252–268, 2018.

[18] A. Kumar, A. M. Costa, M. Fackrell, and P. G. Taylor, "A sequential stochastic mixed integer programming model for tactical master surgery scheduling," *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 734–746, Oct. 2018.

[19] K. S. Shehadeh and R. Padman, "A distributionally robust optimization approach for stochastic elective surgery scheduling with limited intensive care unit capacity," *Eur. J. oper. Res.*, vol. 290, no. 3, pp. 901–913, May 2021.

[20] J. Zhang, M. Dridi, and A. El–Moudni, "Column-generation-based heuristic approaches to stochastic surgery scheduling with downstream capacity constraints," *Int. J. Prod. Econ.*, vol. 229, Nov. 2020, Art. no. 107764.

[21] J. Zhang, M. Dridi, and A. El Moudni, "A two-phase optimization model combining Markov decision process and stochastic programming for advance surgery scheduling," *Comput. Ind. Eng.*, vol. 160, Oct. 2021, Art. no. 107548.

[22] A. Sauré, J. Patrick, and M. L. Puterman, "Simulation-based approximate policy iteration with generalized logistic functions," *INFORMS J. Comput.*, vol. 27, no. 3, pp. 579–595, Aug. 2015.

[23] A. Sauré, M. A. Begen, and J. Patrick, "Dynamic multi-priority, multi-class patient scheduling with stochastic service times," *Eur. J. Oper. Res.*, vol. 280, no. 1, pp. 254–265, Jan. 2020.

[24] H.-J. Schütz and R. Kolisch, "Capacity allocation for demand of different customer-product-combinations with cancellations, no-shows, and overbooking when there is a sequential delivery of service," *Ann. Oper. Res.*, vol. 206, no. 1, pp. 401–423, Jul. 2013.

[25] Y. Gocgun and M. L. Puterman, "Dynamic scheduling with due dates and time windows: An application to chemotherapy patient appointment booking," *Health Care Manage. Sci.*, vol. 17, no. 1, pp. 60–76, Mar. 2014.

[26] M. S. Parizi and A. Ghate, "Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking," *Comput. Oper. Res.*, vol. 67, pp. 90–101, Mar. 2016.

[27] H. Mahmoudzadeh, A. M. Shalamzari, and H. Abouee-Mehrizi, "Robust multi-class multi-period patient scheduling with wait time targets," *Oper. Res. Health Care*, vol. 25, Jun. 2020, Art. no. 100254.

[28] A. Bayram, S. Deo, S. Iravani, and K. Smilowitz, "Managing virtual appointments in chronic care," *IISE Trans. Healthcare Syst. Eng.*, vol. 10, no. 1, pp. 1–17, Jan. 2020.

[29] X. Yu and A. Bayram, "Managing capacity for virtual and office appointments in chronic care," *Health Care Manage. Sci.*, vol. 24, no. 4, pp. 742–767, Dec. 2021.

[30] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control Optim.*, vol. 14, no. 5, pp. 877–898, Aug. 1976.

[31] S. Gul, B. T. Denton, and J. W. Fowler, "A progressive hedging approach for surgery planning under uncertainty," *INFORMS J. Comput.*, vol. 27, no. 4, pp. 755–772, Nov. 2015.

[32] T. G. Crainic, X. Fu, M. Gendreau, W. Rei, and S. W. Wallace, "Progressive hedging-based metaheuristics for stochastic network design," *Networks*, vol. 58, no. 2, pp. 114–124, Sep. 2011.

[33] C. C. Carøe and R. Schultz, "Dual decomposition in stochastic integer programming," *Oper. Res. Lett.*, vol. 24, nos. 1–2, pp. 37–45, Feb. 1999.

[34] Y.-C. Lee, Y.-S. Chen, and A. Y. Chen, "Lagrangian dual decomposition for the ambulance relocation and routing considering stochastic demand with the truncated Poisson," *Transp. Res. B, Methodol.*, vol. 157, pp. 1–23, Mar. 2022.

[35] L. F. Escudero, M. A. Garín, G. Pérez, and A. Unzueta, "Lagrangian decomposition for large-scale two-stage stochastic mixed 0–1 problems," *TOP*, vol. 20, no. 2, pp. 347–374, Jul. 2012.

[36] L. F. Escudero, M. A. Garín, G. Pérez, and A. Unzueta, "Scenario cluster decomposition of the Lagrangian dual in two-stage stochastic mixed 0–1 optimization," *Comput. Oper. Res.*, vol. 40, no. 1, pp. 362–377, Jan. 2013.

[37] L. F. Escudero, M. A. Garín, and A. Unzueta, "Cluster Lagrangean decomposition in multistage stochastic optimization," *Comput. Oper. Res.*, vol. 67, pp. 48–62, Mar. 2016.

[38] L. F. Escudero, M. A. Garín, and A. Unzueta, "Scenario cluster lagrangean decomposition for risk averse in multistage stochastic optimization," *Comput. Oper. Res.*, vol. 85, pp. 154–171, Sep. 2017.

[39] F. Oliveira, V. Gupta, S. Hamacher, and I. E. Grossmann, "A Lagrangean decomposition approach for oil supply chain investment planning under uncertainty with risk considerations," *Comput. Chem. Eng.*, vol. 50, pp. 184–195, Mar. 2013.

[40] C. L. Lara, J. Koenemann, Y. Nie, and C. C. de Souza, "Scalable timing-aware network design via Lagrangian decomposition," *Eur. J. Oper. Res.*, vol. 309, no. 1, pp. 152–169, Aug. 2023.

[41] V. Zeighami, M. Saddoune, and F. Soumis, "Alternating Lagrangian decomposition for integrated airline crew scheduling problem," *Eur. J. Oper. Res.*, vol. 287, no. 1, pp. 211–224, Nov. 2020.

[42] H. Zhang, S. Li, Y. Wang, L. Yang, and Z. Gao, "Collaborative real-time optimization strategy for train rescheduling and track emergency maintenance of high-speed railway: A Lagrangian relaxation-based decomposition algorithm," *Omega-Int. J. Manage. Sci.*, vol. 102, Jul. 2021, Art. no. 102371.

[43] A. Amiri and R. Barkhi, "A lagrangean based solution algorithm for the multiple knapsack problem with setups," *Comput. Ind. Eng.*, vol. 153, Mar. 2021, Art. no. 107089.

[44] M. L. Fisher, "The Lagrangian relaxation method for solving integer programming problems," *Manage. Sci.*, vol. 50, no. 12, pp. 1861–1871, Dec. 2004.

[45] V. Nair et al., "Solving mixed integer programs using neural networks," 2020, *arXiv:2012.13349*.

[46] R. T. Rockafellar and R. J.-B. Wets, "Scenarios and policy aggregation in optimization under uncertainty," *Math. Oper. Res.*, vol. 16, no. 1, pp. 119–147, Feb. 1991.

[47] N. B. Demir, S. Gul, and M. Çelik, "A stochastic programming approach for chemotherapy appointment scheduling," *Nav. Res. Logistics (NRL)*, vol. 68, no. 1, pp. 112–133, Feb. 2021.

[48] M. Held and R. M. Karp, "The traveling-salesman problem and minimum spanning trees—Part II," *Math. Program.*, vol. 1, no. 1, pp. 6–25, Dec. 1971.

[49] M. A. Bragin, B. Yan, and P. B. Luh, "Distributed and asynchronous coordination of a mixed-integer linear system via surrogate lagrangian relaxation," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1191–1205, Jul. 2021.

Xiaoxiao Shen received the B.S. degree from the College of Information Management, Nanjing Agricultural University, Nanjing, China, in 2019. She is currently pursuing the Ph.D. degree in mechanical engineering with Shanghai Jiao Tong University, Shanghai, China. Her current research interests include analysis, modeling, and decision-making optimization for operations management problems in manufacturing systems and service systems. Her technical specialties include mathematical modeling, stochastic optimization, exact optimization algorithm, intelligent optimization algorithm, statistical modeling analysis, and simulation optimization. She is a member of the Operations Research Society of China and a Reviewer of *Computers and Industrial Engineering* journal.

Shi-Chang Du (Member, IEEE) received the B.S. and M.S.E. degrees in mechanical engineering from the Hefei University of Technology, Hefei, China, in 2000 and 2003, respectively, and the Ph.D. degree in industrial engineering and management from Shanghai Jiao Tong University, Shanghai, China, in 2008. He was with the University of Michigan, Ann Arbor, MI, USA, from 2006 to 2007, as a Visiting Scholar. He is currently a Professor with the Department of Industrial Engineering and Management, Shanghai Jiao Tong University. He has authored/coauthored 70 articles in journals, including the *ASME Journal of Manufacturing Science* and IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT. His current research interests include quality engineering, including quality control with the analysis of variation flow, monitoring, and the diagnosis of manufacturing process by support vector machine and Bayesian. His research has been supported by the National Natural Science Foundation of China, the Ministry of Science and Technology of China, and the Shanghai Association for Science and Technology. He is also an Area Editor of *Computers and Industrial Engineering* and an Associate Editor of *Flexible Services and Manufacturing Journal*.

Yan-Ning Sun was born in Shandong, China, in 1994. He received the B.S. and M.S. degrees from Northeastern University, China, in 2016 and 2019, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, China, in 2023. He is currently a Lecturer with the School of Mechatronic and Engineering and Automation, Shanghai University, Shanghai, China. His current research interests include the fundamental study of machine intelligence, complex network and complex systems, data-driven decision-making methods in complex industrial processes, and the development of industrial big data platform for intelligent manufacturing.

Poly Z. H. Sun (Member, IEEE) is currently pursuing the Ph.D. degree in industrial intelligent system with the Department of Industrial Engineering, Shanghai Jiao Tong University, Shanghai, China.

He has authored/coauthored over 30 research papers in top-tier refereed international journals and conferences, such as IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, and IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. His current research interests include intelligent transportation systems, human factors in driving, neuroergonomics, brain and cognitive status under maneuvers, physiological parameter measurement, bio-inspired vision mechanism, operations research, and their applications in human-in-the-loop transportation systems. He is a member of the IEEE ITSS, IEEE CIS, IEEE SMCS, and the Association for Computing Machinery. He has also been serving as a reviewer for several top-tier international journals, such as IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE/ASME TRANSACTIONS ON MECHATRONICS, and IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and the session chair for several conferences.

Rob Law is currently a University of Macau Development Foundation (UMDF) Chair Professor of Smart Tourism. He is also an honorary professor of several other reputable universities. Prior to joining the University of Macau in July 2021, he has worked in industry organizations and academic institutes. He is an active researcher. He has received more than 90 research related awards and accolades, and millions of USD external and internal research grants. He has edited four books and published more than 1,000 research articles (including 100's of articles in first-tier academic journals). His publications have received more than 53,500 citations, with H-index/i 10-index = 105/478 (http://scholar.google.com.hk/). In addition, he serves different roles for more than 200 research journals. He is the chair/a committee member of more than 180 international conferences.

Edmond Q. Wu (Senior Member, IEEE) received the Ph.D. degree in controlling theory and engineering from Southeast University, Nanjing, China, in 2009.

He is currently a Professor with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, China. He is also with the Science and Technology on Avionics Integration Laboratory, China National Aeronautical Radio Electronics Research Institute, Shanghai, China. His research interests include deep learning, fatigue recognition, and human–machine interaction. He is also an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, and IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS.