DBS-20-02

「Coordinated Appointment Scheduling with Multiple Providers and Patient-and-Physician Matching Cost in Specialty Care」

Shenghai Zhou

Jiangxi University of Finance and Economics

Debiao Li

Fuzhou University

Yong Yin

Graduate School of Business; Doshisha University

(Summary)

To achieve effective care, it is critical to match patients with capable physicians in specialty care. Motivated by the rising popularity of patient-and-physician matching applications in specialty care, this study optimizes the matching and appointment scheduling problems simultaneously in a stochastic environment, in which a decision-maker determines the patient-and-physician pair assignment and the starting times of services. We develop a stochastic optimization model to minimize the matching and operational costs (i.e., patients' waiting time costs, service providers' idle time and overtime costs). This paper is the first study that incorporates matching and appointment scheduling problems together. The benefits of combining these two problems are enormous. The experimental results show that the operational costs gap is as large as 51% between the ill-matched and the well-matched patient-and-physician scenarios. We first reformulate this problem as a two-stage optimization problem. With the analysis for the optimal solution of the second stage problem, a Benders decomposition algorithm is developed. To improve the efficiency of the proposed algorithm, we also prove a low bound of our problem and use it to construct a set of feasibility cuts. Then, we extend our method to incorporate no-shows. Our algorithm can solve problems efficiently, and it can obtain optimal solutions for medium-size problems within 2 or 3 minutes. In contrast, traditional optimal methods require nearly 2 hours. For large-size problems, our algorithm can obtain optimal solutions within 5 or 6 minutes, whereas traditional optimal methods cannot generate a result within 5 hours. Finally, numerical experiments are conducted to evaluate the performance of our proposed algorithm and to investigate the variation of the optimal solutions in different scenarios. To provide quality care as well as minimize the total cost of appointment scheduling in specialty care, we suggest that physicians should develop or train their specialties based on the local patients' disease pattern. We also disclose that the no-show has less influence on the service system when the weight of the matching cost is substantial.

Keywords: health care, Benders decomposition, appointment scheduling, multiple service providers, patient-and-physician matching.

Coordinated Appointment Scheduling with Multiple Providers and Patient-and-Physician Matching Cost in Specialty Care

Shenghai Zhou[a], Debiao Li[b,*], Yong Yin[c]

[a]School of Business Administration, Jiangxi University of Finance and Economics, Nanchang, China
[b]School of Economics and Management, Fuzhou University, Fuzhou, China
[c]Graduate School of Business, Doshisha University, Kyoto, Japan

Abstract

To achieve effective care, it is critical to match patients with capable physicians in specialty care. Motivated by the rising popularity of patient-and-physician matching applications in specialty care, this study optimizes the matching and appointment scheduling problems simultaneously in a stochastic environment, in which a decision-maker determines the patient-and-physician pair assignment and the starting times of services. We develop a stochastic optimization model to minimize the matching and operational costs (i.e., patients' waiting time costs, service providers' idle time and overtime costs). This paper is the first study that incorporates matching and appointment scheduling problems together. The benefits of combining these two problems are enormous. The experimental results show that the operational costs gap is as large as 51% between the ill-matched and the well-matched patient-and-physician scenarios. We first reformulate this problem as a two-stage optimization problem. With the analysis for the optimal solution of the second stage problem, a Benders decomposition algorithm is developed. To improve the efficiency of the proposed algorithm, we also prove a low bound of our problem and use it to construct a set of feasibility cuts. Then, we extend our method to incorporate no-shows. Our algorithm can solve problems efficiently, and it can obtain optimal solutions for medium-size problems within 2 or 3 minutes. In contrast, traditional optimal methods require nearly 2 hours. For large-size problems, our algorithm can obtain optimal solutions within 5 or 6 minutes, whereas traditional optimal methods cannot generate a result within 5 hours. Finally, numerical experiments are conducted to evaluate the performance of our proposed algorithm and to investigate the variation of the optimal solutions in different scenarios. To provide quality care as well as minimize the total cost of appointment scheduling in specialty care, we suggest that physicians should develop or train their specialties based on the local patients' disease pattern. We also disclose that the no-show has less influence on the service system when the weight of the matching cost is substantial.

Keywords: health care, Benders decomposition, appointment scheduling, multiple service providers, patient-and-physician matching

*Corresponding Email: debiaoli@fzu.edu.cn
Email addresses: zhoushenghai@jxufe.edu.cn (Shenghai Zhou), debiaoli@fzu.edu.cn (Debiao Li), yyin@mail.doshisha.ac.jp (Yong Yin)

1. Introduction

Specialty care clinics, in which specialists have extensive training and education, are designed to provide specific diagnoses and treatments. As modern medical development, specialties are distinguished even in the same clinics. For example, in otorhinolaryngology, specialty care physicians (SCPs) diagnose and treat a wide range of diseases around the ear, nose, and throat regions. If a patient chooses a physician who is not in the right expertise, effective care may not be achieved. In practice, patients select SCPs either through referrals from their primary care physicians (PCPs) or by themselves. In a specialty care system with referrals, e.g., the United States, referrals by PCPs are frequently failed and often led to medical errors [1, 2]. It can be worse without the help of referrals. Especially for patients, because they have limited medical education and knowledge, but have to choose SCPs by themselves. On the other side, SCPs would prefer to see more patients that fall into their areas of clinical expertise, such that their medical training and education can be realized. Therefore, it is very critical to match patients with capable physicians in specialty care, such that the quality of care can be guaranteed to some extent.

With the rapid development of information technology in healthcare, some applications have been invented to solve this patient and physician matching problem in specialty care recently. For example, in the United States, Specialty Care Connect (https://armadahealth.com/patient-physician-matching) powered by Armada Health had applied analytic models to connect patients to the right SCPs. In China, an intelligent matching system, named "Rui Zhi", has been invented by Tencent, which is one of the biggest internet-based technology corporations in China. It has been implemented in Guangzhou women and children's medical center since May 2018. The accuracy of diagnoses was claimed to be 94%, and the matching accuracy was reported to be 96%. Recently, a patient-and-physician matching index is proposed to measure the capability between patients and SCPs through an improved multi-disease pre-diagnosing Bayesian network model, which is based on given patient's symptoms and physician's specialties [3]. The experimental results show that the proposed patient-and-physician matching index increases the physician matching accuracy under various settings. Integrated with data science technologies and medical knowledge, these matching applications and mechanisms are designed to ensure the effectiveness of specialty care.

Apart from the effectiveness, timely, and accessible coordination in specialty care is also very critical. In order to improve the efficiency of specialty care operations, appointment scheduling is an essential and efficient way to maximize physicians' time utilities and improve patients' satisfaction by reducing waiting time. Appointment scheduling has been well studied in primary care. In specialty care, when we make a schedule, we should take the unique characteristic of specialty into consideration, i.e., the matching between patients and physicians, so as to make a more reasonable schedule. To the best of our knowledge, there is no appointment scheduling literature that considers the effectiveness through the patient and physician matching. To ensure the effectiveness and efficiency of specialty care, the matching mechanism should be embedded in the appointment scheduling system, such that timely access and efficiency can be achieved at the same time.

In an appointment scheduling problem in specialty, there are usually multiple physicians (with appointment scheduling terminology, physicians often refer to service providers) provide services. Thus, our studied appointment scheduling problem in specialty care is a multiple-provider appointment scheduling problem indeed. Due to the vast majority of work on matching between patients and physician preferences, in this work, we regard the matching cost between patients and physicians as the inputs to the appointment scheduling model.

To sum up, in this study, we integrate the matching cost between patients and physicians into appointment scheduling problem for the specialty care, in which the service times are stochastic. Our objective is to jointly optimize the assignment and job allowance for each patient, such that the weighted operational and matching costs are minimized. We first formulate our studied problem as a two-stage optimization problem based on the sample average approximation (SAA) approach. After analyzing the properties for the optimal solution of the second stage problem, an improved Benders decomposition algorithm is proposed. Finally, we conduct computational experiments to identify the efficiency of our proposed algorithm and investigate some managerial insights.

The contributions of this paper are summarized as follows. First, we adapt the patient-and-physician matching cost into the appointment scheduling problem. To the best of our knowledge, we are the first to jointly consider the matching problem and the appointment scheduling problem with multiple service providers in a stochastic environment. The experimental results in Section 5 give the managerial benefits of combining these two problems. The operational costs gap is as large as 51% between the ill-matched and the well-matched patients and physicians scenarios. Second, a two-stage stochastic program is formulated to minimize the weighted operational and matching costs. We analyze a low bound for the optimal objective value under any potential assignment. Third, we analyze the properties for the optimal solution of the second stage problem. On this basis, we propose an improved Benders decomposition algorithm with feasibility cuts to solve this problem efficiently. The experimental results of Section 5 show that our algorithm can obtain optimal solutions for medium-size problems within 2 or 3 minutes. In contrast, traditional optimal methods require nearly 2 hours. For large-size problems, our algorithm can obtain optimal solutions within 5 or 6 minutes, whereas traditional optimal methods cannot generate a result within 5 hours. Finally, several sensitivity analyses are conducted under different parameter settings. Our numerical results indicate the importance of the patient-and-physician matching. To provide quality care as well as minimize the total cost of appointment scheduling in specialty care, we suggest that physicians should develop or train their specialties based on the local patients' disease pattern.

The overview of this paper is organized as follows. The literature related to patient-and-physician matching and appointment scheduling is reviewed in Section 2. We formally describe our studied problem in Section 3. The details of our proposed improved Benders decomposition approach are stated in Section 4. We extend our method to incorporate no-shows in Section 5. In Section 6, we conduct some numerical experiments to verify our proposed method and examine some potential insights. Some managerial implications are summarized in Section 7, followed by the summary and future work in Section 8.

2. Literature Review

In this section, we review the literature that is most relevant to our research. In particular, we focus mainly on patient-and-physician matching problems and fundamental appointment scheduling problems.

2.1. Patient-and-physician matching

The patient-and-physician matching problems of specialty care are different from primary care and elective surgery. In primary care, access to services is the most critical factor for patient-and-physician matching [4]. Several studies also provide evidence that physicians' interpersonal skills affect patients' satisfaction [5] and treatment outcomes [6]. Gong et al. [7] propose a weighted average model to recommend physicians by considering the economic matching degree, medical domain matching degree, recommender influence, and region reference. A time-constraint probability factor graph model learns these features from a real-world medical data set, which was optimized by a constraint-based optimization framework. Liu et al [8] examine the patients' preferences and choice behavior in the scheduling appointment, which include the gender effect, speed, and physician of choice. Although the physician of choice is highly correlated to the quality of care, the detail of the physician of choice is not disclosed.

With the development of modern medical science, physicians tend to have distinguished specialties, although they are in the same department. The medical skills or specialties of physicians become more critical in patient-and-physician matching. Kinchen et al. [9] analyze the significant factors affecting the choice of specialists by primary care physicians through a survey. Medical skills, appointment timeliness, and quality of specialist communication are the three most important factors. Most existing literature studies the variations and their causes in referral decision making among PCPs. However, few explore the appropriateness of the referral decision, which is mainly defined by the medical skill [2].

Pan et al. [10] propose a dynamic preference learning algorithm to recommend physicians in specialty care by considering both patients' preferences and their heterogeneous illness conditions. Furthermore, it is assumed that general practitioners correctly evaluated patients' illness conditions, and there was no bias to refer physicians in specialty care. However, general practitioners may lack related expertise and have some biased information about SCP. To eliminate biases or mistakes of referrals in specialty care, a pre-diagnosing model is applied to gain a more accurate diagnosis of patients' disease(s). Given the specialty information of a physician, a patient-and-physician matching index is proposed to measure the quality influence during the matching [3].

However, the matching literature mentioned above only considers the isolated matching between patients and physicians but ignores the operational aspects (i.e., timely and accessible) during the specialty care visit. Thus, we need to further integrate the matching with the appointment scheduling problem.

2.2. Appointment Scheduling

In the literature, one classic appointment scheduling problem refers to the intra-day scheduling problem, which focuses on making appointment decisions on a given day. For this kind of appointment scheduling problem, usually, it is assumed that only one service provider provides service. The decision-maker needs to determine the start time of each appointment so as to balance the costs for both patients and service providers. Specifically, most works, including ours, study the objective of minimizing the total expected (weighted) costs of patients' waiting times, service provider's idle times, and overtime [11, 12]. If there is no session length constraint, some studies only take patients' waiting times and service providers' idle times into consideration in the objective function [13, 14, 15, 16]. While other studies, including ours, consider total expected (weighted) costs of patients' waiting times and service provider's overtime in the objective function [17, 18, 19]. Consider patients' behavior, some work also take patients' no-shows into consideration [17, 20, 21, 22]. Our work also consider this important patients' behavior.

The classic appointment scheduling problems with a single service provider are often considered in a stochastic environment, such as random service times [11, 23] and random no-shows [21, 22]. In order to handle those uncertainties, most works focus on developing some stochastic optimization models [11, 24, 25, 21, 22]. On this basis, the developed stochastic models are often approximated as corresponding (mixed-integer) linear programs through Sample Average Approximation (SAA) approach [26, 27, 28]. For those (mixed-integer) linear programs, when the sample size is small (e.g., ≤ 500), they often can be solved directly [27, 28]. Nevertheless, when the sample size is large, it is difficult to achieve a high accuracy level solution within a reasonable computational time. In this case, some efficient algorithms are developed, such as the Benders decomposition algorithm [29, 30, 26], and simulation-based sequential algorithms [31]. In this work, we first formulate our studied problem as a stochastic program. And then, we also exploit the SAA approach to handle the stochastic service times. On this basis, we develop an improved Benders decomposition algorithm to solve the problem.

However, the vast majority of the literature focuses on service systems that only involve one service process with one service provider [11, 24, 25, 17, 23]. In addition to systems with only one service provider, systems with many service providers also prevail in practice. However, studies on appointment scheduling problems with multiple service providers are limited. To the best of our knowledge, the only appointment scheduling works with multiple service provider systems are those by [32, 33, 34, 35, 36]. Sickinger et al. [34] consider two CT-scan machines in a radiology department, in which two machines are regarded as two identical parallel providers with identical deterministic service times. Similarly, Zacharias et al. [36] also assume the service providers are identical, and the service time is also identical deterministic. In contrast, Alvarez et al. [32] consider stochastic service times in a two-stage service system in which two identical parallel service providers in the first stage. Their purpose is to minimize the total weighted costs of patients' waiting and service providers' idling. Zheng et al. [33] consider no-shows in their model; however, their model cannot be easily adapted to the overtime case. Soltani et al. [35] also consider stochastic service times, and patients no-shows for an appointment scheduling problem with multiple service providers. Different from Zheng et al [33], they consider both patients' waiting time, and service providers' idle time and overtime in the objective function.

However, all existing appointment scheduling problems do not consider the quality-related matching factors between patients and physicians. As we mentioned in the introduction section, the patient-and-physician matching is critical to improving the effectiveness in specialty care. Therefore, we integrate the multiple-provider appointment scheduling problem with the patient-and-physician matching problem in this paper.

3. Problem Formulation

In this paper, we consider generic specialty care with k service providers (in the rest of this paper, we use the terminology "service provider" to represent the specialty care physician). There are totally n patients needed to be scheduled within a session length T. Before making a schedule, the decision-maker has the following information: (1) The matching costs ($m_{i,j}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, k$), which denote the cost for patient i match physician j, are assumed to be known in advance. In this work, we take the matching cost as input and it can be derived from the patient-and-physician matching index proposed by [3]. Li et al. [3] measure the matching indexes between patients and physicians based on the symptom-specialty relationship through a trained Bayesian network pre-diagnosing model. We transform the matching indexes into matching costs. Generally speaking, a higher matching index indicates a better-matched patient and physician. If a patient is assigned to a capable physician with a high matching index, it most likely leads to a better health-care outcome, which has a higher potential to reduce healthcare costs. Thus, when we transform matching indexes into matching costs, we may assume the higher the matching index, the lower the matching cost. For simplification. in this work, we assume the matching costs are known. (2) Through some preliminary classification, the service time of patient i at service provider j, $d_{i,j}$ $(i = 1, 2, \cdots, n, j = 1, 2, \cdots, k)$ is an independent, not necessary identically distributed random variable, which is known to the decision-maker.

With the above information, the decision-maker of the specialty care needs to determine (1) assignment problem and (2) appointment scheduling problem. Specifically, the assignment problem means how to assign those n patients to k service providers. We use $x_{i,j}$ $(i = 1, 2, \cdots, n, j = 1, 2, \cdots, k)$ to denote the decision variable for assignment problem. If patient i $(i = 1, 2, \cdots, n)$ is assigned to service provider j $(j = 1, 2, \cdots, k)$, $x_{i,j} = 1$, otherwise $x_{i,j} = 0$. Given the assignment, the appointment scheduling problem means, for each service provider, how to determine the start times (or equivalently the job allowances) for assigned patients. We use $s_{i,j}$ $(i = 1, 2, \cdots, n, j = 1, 2, \cdots, k)$ to denote the decision variable (i.e., job allowance) for appointment scheduling problem. And we have $s_{i,j} = 0$ if $x_{i,j} = 0$, which means if patient i is not assigned to service provider j, then we do not leave any job allowance for that patient i at service provider j. We assume that once the assignment is fixed, the patients assigned to each service provider would go through their service according to their index order, i.e., the service provider would handle patient i before patient i' if $i < i'$.

With any given assignment, for each service provider, the corresponding appointment scheduling degenerates into a classic appointment scheduling. For convenience and clarity, in the rest of this paper, we refer to a patient in the appointment system as a job and use the terms job and patient interchangeably.

Due to the randomness of the stochastic service times, patients' waiting or service providers' idling might come up. We use the term $\bar{W}_{i,j}$ $(i = 1, 2, \cdots, n, j = 1, 2, \cdots, k)$ to denote the actual waiting time of patient i before it has been seen by service provider j. Then we have $\bar{W}_{i,j} = 0$ if $x_{i,j} = 0$.In order to derive the waiting times logically, we introduce the virtual waiting time $W_{i,j}$ $(i = 1, 2, \cdots, n, j = 1, 2, \cdots, k)$ for patient i at service provider j. The virtual waiting time $W_{i,j}$ indicates the waiting time of patient i before she/he is seen by service provider j, regardless of the actual assignment of patient i. Given any assignment $x_{i,j} = 1$, patient i may actually need to wait for service provider j to be served but must not wait for other service provider $h \neq j$. With the definition of virtual waiting time $W_{i,j}$, the actual waiting time $\bar{W}_{i,j}$ can be achieved by multiplying the virtual waiting time $W_{i,j}$ by the assignment $x_{i,j}$, i.e., $\bar{W}_{i,j} = x_{i,j} W_{i,j}$.

Next, we present how to derive virtual waiting times $W_{i,j}$ recursively. In the classic single provider appointment scheduling problem, the actual waiting time is determined through the waiting time, the service time, and the job allowance of its preceding appointment recursively. However, for our problem, the actual service time is expressed as $x_{i,j} \cdot d_{i,j}$. Thus, in order to derive the virtual waiting times $W_{i,j}$, we can modify the service time of our problem and formulate the virtual waiting time $W_{i,j}$ as follows:

$$\begin{aligned} &W_{i,j} = \max\{0, x_{i-1,j}d_{i-1,j} + W_{i-1,j} - s_{i-1,j}\} \quad i = 2, \cdots n, j = 1, \cdots k \\ &W_{1,j} = 0 \quad j = 1, \cdots k \end{aligned} \tag{3.1}$$

Similarly, we define the virtual idle time $I_{i,j}$ ($i = 1, 2, \cdots n, j = 1, 2, \cdots k$), , which denotes the idleness of service provider j after serving patient i. Note that for any service provider j, the summation $\sum_{i=1}^{n} I_{i,j}$ equals to the actual idleness for service provider j. Thus, the notation of virtual idle time $I_{i,j}$ is enough to depict our performance indicator and we do not need to define the actual idle time. In the rest of this paper, we will omit the term "virtual" for idle time $I_{i,j}$. Then we also derive the idle time $I_{i,j}$ recursively as follows:

$$I_{i-1,j} = \max\{0, -x_{i-1,j}d_{i-1,j} - W_{i-1,j} + s_{i-1,j}\} \quad i = 2, \cdots n+1, j = 1, 2, \cdots k \tag{3.2}$$

Note that we restrict all services that should be finished within a session length T for all service providers. As a result, the service system may incur some overtime for some service providers. We define O_j ($j = 1, \cdots k$) to represent the overtime for service provider j, then it is derived as follows:

$$O_j = \max\{0, x_{n,j}d_{n,j} + W_{n,j} - s_{n,j}\} \quad j = 1, 2, \cdots k \tag{3.3}$$

For the service system, again, $m_{i,j}$ states the cost for patient i being assigned to service provider j. Let c_i^W denote the unit waiting time cost for patient i. And let c_j^I and c_j^O denote the unit idle time, and overtime cost for service provider j, respectively. We define a weight λ to balance the matching cost and the weighted operational costs. The objective is to optimize the assignment and appointment scheduling simultaneously, such that the expected weighted operational (i.e., waiting costs, idling costs, and overtime costs) and matching cost is minimized, as shown in equation (3.4):

$$\sum_{j=1}^{k} \mathbb{E}\left[\sum_{i=1}^{n} (c_i^W x_{i,j} W_{i,j} + c_j^I I_{i,j}) + c_j^O O_j\right] + \lambda \sum_{i=1}^{n} \sum_{j=1}^{k} m_{i,j} x_{i,j} \tag{3.4}$$

In the objective function (3.4), the weight λ is used to balance different dimensions of matching cost and the weighted operational costs. In practice, we may first test different values of λ to justify the corresponding matching and operational costs, and then select an appropriate value of λ to implement. Thus, in our numerical analyses section, we take the values of λ from 0 to 100. When $\lambda = 0$, it indicates that only the operational cost is considered. Besides, in the numerical analyses, we demonstrate the effect of the matching cost by increasing λ, which is the same as the standardization of the cost coefficients.

Next, we construct constraints for our problem. For decision variables $x_{i,j}$ and $s_{i,j}$, we define the following constraints:

$$\begin{aligned} &\sum_{i=1}^{n} x_{i,j} \geq 1 \quad j = 1, 2, \cdots, k \\ &\sum_{j=1}^{k} x_{i,j} = 1 \quad i = 1, 2, \cdots, n \\ &\sum_{i=1}^{n} s_{i,j} = T \quad j = 1, 2, \cdots, k \end{aligned} \tag{3.5}$$

The first equation in (3.5) makes sure there should be at least one patient assigned to each service provider, because there is no necessary to keep one service provider idle in a session length. The second equation in (3.5) ensures each patient should be assigned to one service provider. The third equation in (3.5) indicates all appointments should be scheduled within session length T. In addition, we need to make sure $s_{i,j} = 0$ if patient i is not assigned to service provider j. Thus, we define the following inequalities to achieve this purpose:

$$s_{i,j} \le M x_{i,j} \quad i = 1, 2, \cdots n, j = 1, 2, \cdots k \tag{3.6}$$

where M is a big number.

Through recursive equations (3.1), (3.2), and (3.3), we derive the following equalities:

$$\begin{aligned} & W_{i,j} - I_{i-1,j} = x_{i-1,j} d_{i-1} + W_{i-1,j} - s_{i-1,j} \quad i = 2, 3, \cdots n, j = 1, 2, \cdots, k \\ & O_j - I_{n,j} = x_{n,j} d_n + W_{n,j} - s_{n,j} \quad j = 1, 2, \cdots, k \end{aligned} \tag{3.7}$$

With performance indicators derived in equations (3.1),(3.2) and (3.3), the objective is to jointly optimize the assignment and appointment scheduling, such that the expected weighted operational (i.e., waiting costs, idling costs and overtime costs) and matching cost is minimized. Thus, our studied problem can be formulated as the following stochastic model ($M0$):

$$\begin{aligned} \text{(M0)} \qquad & \min_{\mathbf{x},\mathbf{s}} \sum_{j=1}^{k} \mathbb{E}\left[\sum_{i=1}^{n} (c_i^W x_{i,j} W_{i,j} + c_j^I I_{i,j}) + c_j^O O_j \right] + \lambda \sum_{i=1}^{n} \sum_{j=1}^{k} m_{i,j} x_{i,j} \\ & s.t. \ (3.5), (3.6), (3.7) \\ & \quad x_{i,j} \in \{0, 1\}, s_{i,j} \ge 0 \end{aligned} \tag{3.8}$$

4. Proposed Method

To solve this stochastic problem, we first exploit the SAA method to handle the stochastic service times. On this basis, an improved Benders decomposition algorithm with feasibility cuts is developed to solve the approximated problem efficiently.

4.1. SAA-based Formulation

As indicated in the literature, the SAA method is an efficient scenario-based method for solving stochastic programming problems, and has been widely used to solve appointment scheduling problems [23, 37, 26]. Specifically, with given service time distribution $\mathbf{d}$, we randomly generate H i.i.d. realizations. Then, the stochastic program (3.8) can be approximated by the following deterministic program (DP):

$$
\begin{aligned}
\text{(DP)} \qquad & \min_{\mathbf{x},\mathbf{s},\mathbf{W},\mathbf{I},\mathbf{O}} \frac{1}{H}\sum_{h=1}^{H}\sum_{j=1}^{k}\left[\sum_{i=1}^{n}(c_i^W x_{i,j}W_{i,j}^h + c_j^I I_{i,j}^h) + c_j^O O_j^h\right] + \lambda\sum_{i=1}^{n}\sum_{j=1}^{k} m_{i,j}x_{i,j} \\
& s.t. \ \sum_{i=1}^{n} x_{i,j} \geq 1 \quad j = 1, 2, \cdots, k \\
& \quad \sum_{j=1}^{k} x_{i,j} = 1 \quad i = 1, 2, \cdots, n \\
& \quad \sum_{i=1}^{n} s_{i,j} = T \quad j = 1, 2, \cdots, k \\
& \quad s_{i,j} \leq M x_{i,j} \quad\ i = 1, \cdots n, j = 1, \cdots, k \\
& \quad W_{i,j}^h - I_{i-1,j}^h = x_{i-1,j}d_{i-1}^h + W_{i-1,j}^h - s_{i-1,j} \quad i = 2, \cdots n, j = 1, \cdots, k, h = 1, \cdots, H \\
& \quad O_j^h - I_{n,j}^h = x_{n,j}d_n^h + W_{n,j}^h - s_{n,j} \quad j = 1, \cdots, k, h = 1, \cdots, H \\
& \quad W_{1,j}^h = 0 \quad j = 1, \cdots, k, h = 1, \cdots, H \\
& \quad x_{i,j} \in \{0, 1\}, s_{i,j} \geq 0
\end{aligned}
\tag{4.1}
$$

In the above deterministic program based on SAA, $d_{i,j}^h$ denotes the realization of service time ($d_{i,j}$) under realization h, the variables $W_{i,j}^h$, $I_{i,j}^h$ and O_j^h denote the corresponding waiting time, idle time and overtime under realization h, respectively. Note that except for the original decision variables $\mathbf{x}$ and $\mathbf{s}$, we let the performance indicators $\mathbf{W}, \mathbf{I}$ and $\mathbf{O}$ as new decision variables, to linearize those performance indicators.

For the above deterministic program, it is a non-linear mixed-integer linear program. We can even introduce a big-M method to reformulate it as a mixed-integer linear program. However, when the problem size is large, it is difficult to achieve an optimal solution in a reasonable computational time. According to our preliminary test, when $H = 1000$, $n = 40, k = 4$, respectively, the optimal solution of the corresponding MILP cannot be achieved within 5 hours. Thus, we propose an improved Benders decomposition to solve above DP efficiently.

4.2. Improved Benders Decomposition

The Benders decomposition method is suitable for some large scale problems with special structure. Our problem has the structure that the subproblem can be solved to optimality without actually solving the corresponding problem, which is suitable for the Benders decomposition. Before we introduce the Benders decomposition method, we first analyze a lower bound for the objective function, which helps to generate some feasibility cuts for the proposed algorithm. We define the individual cost $C_{i,j}$ for the original problem ($M0$) as follows:

$$
C_{i,j} = \begin{cases} \mathbb{E}[c_{i+1}^W x_{i+1,j}W_{i+1,j} + c_j^I I_{i,j}] & i = 1, 2, \cdots, n-1 \\ \mathbb{E}[c_j^O O_j + c_j^I I_{i,j}] & i = n \end{cases} \qquad j = 1, \cdots k \tag{4.2}
$$

Lemma 1. For any given assignment $\mathbf{x}$, the individual cost $C_{i,j}$ is bounded as follows:

$$C_{i,j} \geq \begin{cases} x_{i+1,j} g_{i,j} & i = 1, 2, \cdots, n-1 \\ g_{i,j} & i = n \end{cases} \qquad j = 1, \cdots k \tag{4.3}$$

where

$$g_{i,j} = \begin{cases} \min\limits_{s_{i,j}} \mathbb{E}[c^W_{i+1}[d_{i,j} - s_{i,j}]^+ + c^I_j[d_{i,j} - s_{i,j}]^-] & i = 1, 2, \cdots, n-1 \\ \min\limits_{s_{n,j}} \mathbb{E}[c^O_j[d_{i,j} - s_{i,j}]^+ + c^I_j[d_{i,j} - s_{i,j}]^-] & i = n \end{cases} \qquad j = 1, \cdots k \tag{4.4}$$

where $[a]^+ = \max\{a, 0\}$ and $[a]^- = \max\{-a, 0\}$.

Proof: For $i = 1, 2, \cdots, n-1, j = 1, \cdots k$, we always have $C_{i,j} \geq x_{i+1,j}\mathbb{E}[c^W_{i+1}W_{i+1,j} + c^I_j I_{i,j}]$. We now bound the right hand $E[c^W_{i+1}W_{i+1,j} + c^I_j I_{i,j}]$. Based on the definition, we have

$$\begin{aligned} \mathbb{E}[c^W_{i+1}W_{i+1,j} + c^I_j I_{i,j}] =& \mathbb{E}\left[c^W_{i+1}[W_{i,j} + d_{i,j} - s_{i,j}]^+ + c^I_j[W_{i,j} + d_{i,j} - s_{i,j}]^-\right] \\ =& \mathbb{E}_{W_{i,j}}\left\{\mathbb{E}_{d_{i,j}}\left[c^W_{i+1}[W_{i,j} + d_{i,j} - s_{i,j}]^+ + c^I_j[W_{i,j} + d_{i,j} - s_{i,j}]^- \middle| W_{i,j}\right]\right\} \end{aligned}$$

Suppose $s^*_{i,j}$ is the optimal solution to achieve $g_{i,j}$ such that $s^*_{i,j} = \arg\min_{s_{i,j}} \mathbb{E}_{d_{i,j}}\left[c^W_{i+1}[d_{i,j} - s_{i,j}]^+ + c^I_j[d_{i,j} - s_{i,j}]^-\right]$. It follows that

$$g_{i,j} \leq \mathbb{E}_{d_{i,j}}\left[c^W_{i+1}[d_{i,j} - \tilde{s}_{i,j}]^+ + c^I_j[d_{i,j} - \tilde{s}_{i,j}]^-\right], \forall \tilde{s}_{i,j}$$

Therefore, for any realization of $W_{i,j}$, let $\tilde{s}_{i,j} = s_{i,j} - W_{i,j}$, we have

$$g_{i,j} \leq \mathbb{E}_{d_{i,j}}\left[c^W_{i+1}[W_{i,j} + d_{i,j} - s_{i,j}]^+ + c^I_j[W_{i,j} + d_{i,j} - s_{i,j}]^- \middle| W_{i,j}\right]$$

By taking expectation for random variable $W_{i,j}$ for above equation, the inequality still holds, which implies $\mathbb{E}[c^W_{i+1}W_{i+1,j} + c^I_j I_{i,j}]$ is bounded by $g_{i,j}$ for any $s_{i,j}$.

Similarly, we can also bound $C_{i,j}$ when $i = n$. This completes the proof. □

Note that in Lemma 1, $g_{i,j}$ corresponds to the optimal cost of a general Newsvendor problem [38]. Given the cumulative distribution function $F_{i,j}$ for $d_{i,j}$, the optimal solution $s^*_{i,j}$ can be achieved through $F_{i,j}(s^*_{i,j}) = \frac{c^W_{i+1}}{c^W_{i+1} + c^I_j}$ for $i = 1, \cdots, n-1, j = 1, \cdots, k$. Thus, the optimal cost $g_{i,j}$ can be calculated easily.

With Lemma 1, for any given assignment $\mathbf{x}$, the operational cost is bounded as $\sum_{j=1}^{k}[\sum_{i=1}^{n-1} x_{i+1,j} g_{i,j} + g_{n,j}]$. Furthermore, Lemma 1 helps to bound some variables in the master problem in the Benders decomposition algorithm. We will introduce it later.

We observe that for any solution $(\mathbf{x}, \mathbf{s})$, the operational costs are decomposable by scenario h and service provider j. And they can be determined through recursive equations (3.1), (3.2) and (3.3). Due to this kind of special structure, we further reformulate problem (4.1) as the following two-stage optimization problem:

$$
\begin{aligned}
&\min_{\mathbf{x},\mathbf{s}} \sum_{j=1}^{k} Q_j(\mathbf{x},\mathbf{s}) + \lambda \sum_{i=1}^{n}\sum_{j=1}^{k} m_{i,j}x_{i,j} \\
&s.t.\ \sum_{i=1}^{n} x_{i,j} \geq 1 \quad j = 1,2,\cdots,k \\
&\qquad \sum_{j=1}^{k} x_{i,j} = 1 \quad i = 1,2,\cdots,n \\
&\qquad \sum_{i=1}^{n} s_{i,j} = T \quad j = 1,2,\cdots,k \\
&\qquad s_{i,j} \leq Mx_{i,j} \quad i = 1,\cdots n, j = 1,\cdots,k \\
&\qquad x_{i,j} \in \{0,1\}, s_{i,j} \geq 0
\end{aligned} \tag{4.5}
$$

where

$$
\begin{aligned}
Q_j(\mathbf{x},\mathbf{s}) = &\min_{\mathbf{W},\mathbf{I},\mathbf{O}} \frac{1}{H}\sum_{h=1}^{H}\left[\sum_{i=1}^{n}(c_i^W x_{i,j}W_{i,j}^h + c_j^I I_{i,j}^h) + c_j^O O_j^h\right] \\
&s.t.\ W_{i,j}^h - I_{i-1,j}^h = x_{i-1,j}d_{i-1,j}^h + W_{i-1,j}^h - s_{i-1,j} \quad i = 2,\cdots n, h = 1,\cdots,H \\
&\qquad O_j^h - I_{n,j}^h = x_{n,j}d_{n,j}^h + W_{n,j}^h - s_{n,j} \quad h = 1,\cdots,H \\
&\qquad W_{1,j}^h = 0 \quad h = 1,\cdots,H
\end{aligned} \tag{4.6}
$$

The above problems (4.5) and (4.6) called master problem (MP) and subproblem(SP), respectively. As we mentioned, with the optimal solution obtained from problem (4.5), the optimal cost and solution of problem (4.6) can be achieved through recursive equations (3.1), (3.2) and (3.3) without actually solving the optimization problem. Thus, the remaining problem is how to use the solution of the second stage problem to verify the optimality of the master problem.

Next, we analyze the optimal solution of the dual problem of $Q_j(\mathbf{x},\mathbf{s})$, which helps to find the optimal solution for the first problem (4.5). For each scenario h, let $\alpha_{i,j}^h(i = 1,2,\cdots n, h = 1,2,\cdots H)$ be the dual decision variable for the second problem (4.6), the dual form of the operational cost for service provider j under scenario h can be derived as follows:

$$
\begin{aligned}
&\max_{\alpha} \sum_{j=1}^{k}\sum_{i=1}^{n}(x_{i,j}d_{i,j}^h - s_{i,j})\alpha_{i,j}^h \\
&s.t.\ \alpha_{i-1,j}^h - \alpha_{i,j}^h \leq c_i^W x_{i,j} \quad i = 2,\cdots n, h = 1,\cdots H \\
&\qquad -\alpha_{i,j}^h \leq c_j^I \quad i = 2,\cdots n, h = 1,\cdots H \\
&\qquad \alpha_{n,j}^h \leq c_j^O \quad h = 1,\cdots H
\end{aligned} \tag{4.7}
$$

By the strong dual theorem, we can obtain the optimal solution of the dual problem (4.7) without actually solve it:

$$\alpha_{i,j}^h = \begin{cases} \alpha_{i+1,j}^h + c_{i+1}^W x_{i+1,j} & \text{if } W_{i+1,j}^h > 0 \\ -c_j^I & \text{otherwise} \end{cases} \qquad i = 1, \cdots n-1, h = 1, \cdots H \tag{4.8}$$

$$\alpha_{n,j}^h = \begin{cases} c_j^O & \text{if } O_j^h > 0 \\ -c_j^I & \text{otherwise} \end{cases} \qquad h = 1, \cdots H \tag{4.9}$$

To solve the two-stage program (4.5) efficiently, we propose an improved Benders decomposition algorithm based on the optimal solution of (4.8) and (4.9). Following the idea of general Benders decomposition, the procedures of the Benders decomposition algorithm for our studied two-stage program are as follows: (1) Formulate the master problem MP as a relaxation form by replacing the optimal value $Q_j(\mathbf{x}, \mathbf{s})$ with a new decision variable θ_j ($\theta_j \geq 0$). And then solve the new master problem and find an optimal solution $\mathbf{s}$ and sends it to the SP. (2) Evaluate whether the optimal solution obtained in the new master problem violates the optimality. If it does, then add optimality cuts generated by incorporating the optimal solution of the SP to the MP and go back to procedure (1); otherwise, the solution is globally optimal. (3) Repeat the above two procedures until an optimal solution is found.

In this work, we also improve the standard Benders decomposition by adding the feasibility cuts $\theta_j \geq \sum_{i=1}^{n-1} x_{i+1,j} g_{i,j} + g_{n,j} \quad j = 1, \cdots n$, which derives from Lemma 1. The set of feasibility cuts restrict the feasible region, which helps to solve the problem more efficiently. As for the optimality cuts $\{L(\mathbf{x}, \mathbf{s}) \geq 0\}$ that come from the optimal solution of the dual of the SP, due to the special structure of SP, the optimal solution can be achieved easily without actually solving the optimization problem.

The pseudocode of the proposed improved Benders decomposition is presented in Algorithm 1.

Algorithm 1

Step 1: Input: Service time realization $\mathbf{d}$, parameters c^W, c^I, c^O, λ, and $\mathbf{m}$. Set the set of optimality cuts $\{L(\mathbf{x}, \mathbf{s}) \geq 0\} = \emptyset$.

Step 2: Solve the master problem

$$\begin{aligned}
(MP) \qquad & \min_{\mathbf{x},\mathbf{s},\theta} \sum_{j=1}^{k} \theta_j + \lambda \sum_{i=1}^{n} \sum_{j=1}^{k} m_{i,j} x_{i,j} \\
& s.t. \sum_{j=1}^{k} x_{i,j} = 1 \quad i = 1, \cdots n \\
& \sum_{i=1}^{k} x_{i,j} \geq 1 \quad j = 1, \cdots n \\
& \sum_{i=1}^{n} s_{i,j} = T \quad j = 1, 2, \cdots, k \\
& s_{i,j} \leq M x_{i,j} \quad i = 1, \cdots n, j = 1, \cdots k \\
& \theta_j \geq \sum_{i=1}^{n-1} x_{i+1,j} g_{i,j} + g_{n,j} \quad j = 1, \cdots n \\
& x_{i,j} \in \{0, 1\}, s_{i,j} \geq 0 \\
& L(\mathbf{x}, \mathbf{s}) \geq 0
\end{aligned}$$

and record an optimal solution $(\mathbf{x}^*, \mathbf{s}^*, \theta^*)$.

Step 3: Given $(\mathbf{x}^*, \mathbf{s}^*)$ obtained from above MP, calculate the corresponding performance indicators $\mathbf{W}, \mathbf{I}$ and $\mathbf{O}$ by recursive equations (3.1), (3.2) and (3.3) . And record $\sum_{j=1}^{k} Q_j^*(\mathbf{x}^*, \mathbf{s}^*)$. On this basis, determine the optimal solution (α) to problem 4.7 for each scenario through recursive equations (4.8) and (4.9).

Step 4: If $\sum_{j=1}^{k} \theta_j^* \geq \sum_{j=1}^{k} Q_j^*(\mathbf{x}^*, \mathbf{s}^*)$, then

stop and return $(\mathbf{x}^*, \mathbf{s}^*, \theta^*)$ as an optimal solution.

Else

add the set of cuts $\theta_j \geq \frac{1}{H} \sum_{h=1}^{H} \left[\sum_{i=1}^{n} d_{i,j}^h \alpha_{i,j}^h x_{i,j} - \sum_{i=1}^{n} \alpha_{i,j}^h s_{i,j} \right]$ $(j = 1, \cdots, k)$, to the set of optimality cuts $\{L(\mathbf{x}, \mathbf{s}) \geq 0\}$. And go to Step 2.

End if.

5. Incorporating No-shows

In this section, the proposed method is extended to solve the matching and appointment scheduling problem by considering no-shows. Let p_i denote the show up probability for patient i, which is known to the decision-maker. Let z_i indicate whether patient i shows up for her appointment (i.e., $z_i = 1$ with probability p_i) or not (i.e., $z_i = 0$ with probability $1 - p_i$). And we also assume that the no-shows are independent for patients.

The key idea is to treat the no-show patient as a "ghost" patients with 0 service time. Let $\tilde{d}_{i,j}$ denote the service time in the presence of no-shows. We can calculate the actual service time in the system through $\tilde{d}_{i,j} = z_i d_{i,j}$, where $d_{i,j}$ is the service time without no-shows studied previously. By abusing notations, let $W_{i,j}, I_{i,j}$, and O_j denote the corresponding virtual waiting time, idle time, and over time, respectively. Then, we have

$$\begin{aligned}
&W_{i,j} = \max\{0, x_{i-1,j}\tilde{d}_{i-1,j} + W_{i-1,j} - s_{i-1,j}\} \quad i = 2, \cdots n, j = 1, \cdots k \\
&I_{i-1,j} = \max\{0, -x_{i-1,j}\tilde{d}_{i-1,j} - W_{i-1,j} + s_{i-1,j}\} \quad i = 2, \cdots n+1, j = 1, 2, \cdots k \\
&O_j = \max\{0, x_{n,j}\tilde{d}_{n,j} + W_{n,j} - s_{n,j}\} \quad j = 1, 2, \cdots k \\
&W_{1,j} = 0 \quad j = 1, \cdots k
\end{aligned} \tag{5.1}$$

In the presence of no-shows, if one patient is a no-show, we can regard the actual waiting time for him/her as zero. Thus, in the objective function, we only need to count the waiting time of those patients who actually show up. On this basis, the total cost under the no-show case is

$$\sum_{j=1}^{k} \mathbb{E}\left[\sum_{i=1}^{n} (c_i^W x_{i,j} z_i W_{i,j} + c_j^I I_{i,j}) + c_j^O O_j\right] + \lambda \sum_{i=1}^{n} \sum_{j=1}^{k} m_{i,j} x_{i,j} \tag{5.2}$$

Note that z_i is independent of $(z_1, \cdots, z_{i-1})$, it must be independent of $W_{i,j}$. Thus, $\mathbb{E}\left[c_i^W x_{i,j} z_i W_{i,j}\right] = p_i \mathbb{E}\left[c_i^W x_{i,j} W_{i,j}\right]$. Let $\tilde{c}_i^W = p_i c_i^W$, then the above equation (5.2) is equivalent to

$$\sum_{j=1}^{k} \mathbb{E}\left[\sum_{i=1}^{n} (\tilde{c}_i^W x_{i,j} W_{i,j} + c_j^I I_{i,j}) + c_j^O O_j\right] + \lambda \sum_{i=1}^{n} \sum_{j=1}^{k} m_{i,j} x_{i,j} \tag{5.3}$$

Note that equation (5.3) has the same form as equation (3.4), except the notations $\tilde{c}_i^W$ and $\tilde{d}_{i,j}$. Therefore, the proposed method in section 4 can be applied for the case with no-shows.

6. Numerical Analyses

In this section, we conduct numerical experiments to evaluate the performance of our proposed algorithm, and study the influence of different parameters on the optimal assignment and schedule. Specifically, we intend to compare the computational time between our proposed method and the benchmark. The benchmark is solving the corresponding MILP through CPLEX directly, which we will introduce later. Moreover, we investigate the optimal solutions under different scenarios $(\lambda, (n, k))$ and the effect of no-shows. We assume an i.i.d. normal distribution for the service time, i.e., $d_{i,j} \sim N(\mu, \sigma^2), i = 1, \cdots n, j = 1, \cdots k$, which has been widely used in the appointment scheduling literature [11, 24]. Following the literature[21, 20, 36], we set the unit waiting time, idling time and overtime costs as $c_i^W = 0.2, c_j^I = 1, c_j^O = 1.5$.

6.1. Performance of Improved Decomposition Algorithm

In this subsection, we study the performance of our proposed decomposition algorithm. We first reformulate the original deterministic problem ($\boldsymbol{DP}$) as a deterministic mixed-integer linear program ($\boldsymbol{MILP}$). And then we solve the MILP directly with CPLEX and use the corresponding results as the benchmark. Through the big-M transformation, the deterministic model ($\boldsymbol{DP}$) can be reformulated as the following mixed-integer linear program:

$$
\begin{aligned}
\text{(DMILP)} \qquad & \min_{\mathbf{x},\mathbf{s},\mathbf{W},\bar{\mathbf{W}},\mathbf{I},\mathbf{O}} \frac{1}{H} \sum_{h=1}^{H} \sum_{j=1}^{k} \left[\sum_{i=1}^{n} (c_i^W \bar{W}_{i,j}^h + c_j^I I_{i,j}^h) + c_j^O O_j^h \right] + \lambda \sum_{i=1}^{n} \sum_{j=1}^{k} m_{i,j} x_{i,j} \\
& s.t. \ \sum_{i=1}^{n} x_{i,j} \geq 1 \quad j = 1, 2, \cdots, k \\
& \sum_{j=1}^{k} x_{i,j} = 1 \quad i = 1, 2, \cdots, n \\
& \sum_{i=1}^{n} s_{i,j} = T \quad j = 1, 2, \cdots, k \\
& s_{i,j} \leq M x_{i,j} \quad i = 1, \cdots n, j = 1, \cdots, k \\
& W_{i,j}^h - I_{i-1,j} = x_{i-1,j} d_{i-1,j}^h + W_{i-1,j}^h - s_{i-1,j} \quad i = 2, \cdots n, j = 1, \cdots, k, h = 1, \cdots, H \\
& O_j^h - I_{n,j}^h = x_{n,j} d_{n,j}^h + W_{n,j}^h - s_{n,j} \quad j = 1, \cdots, k, h = 1, \cdots, H \\
& \bar{W}_{i,j}^h \geq W_{i,j}^h + (x_{i,j} - 1) M \quad i = 2, \cdots n, j = 1, \cdots, k, h = 1, \cdots, H \\
& \bar{W}_{i,j}^h \leq W_{i,j}^h \quad i = 2, \cdots n, j = 1, \cdots, k, h = 1, \cdots, H \\
& W_{1,j}^h = 0 \quad j = 1, \cdots, k, h = 1, \cdots, H \\
& x_{i,j} \in \{0, 1\}, s_{i,j} \geq 0
\end{aligned}
\tag{6.1}
$$

Throughout this section, we randomly generate $H = 1000$ i.i.d. realizations based on given service time distribution. And then we solve the MILP directly with CPLEX. Finally, we illustrate the superiority of our method by comparing the running time with the benchmark.

The parameters are presented as follows.

- The service times of all jobs at each server $d_{i,j}$ follow a normal distribution N(20, 16).
- The matching cost $m_{i,j}$ of patient i for physician j is generated by a uniform distribution $U[0, 1]$.
- The number of patients and physicians appear as pairs (n, k). We consider four pairs, i.e., $(20, 2)$, $(30, 3)$, $(40, 4)$ and $(50, 5)$.
- The session length is set at $T = 1.5 \cdot \mu \cdot n/k$;

For each pair (n, k), we generate 5 problem instances, for each problem instance, the λ is randomly generated from interval $[10, 1000]$. As a result, there are a total of 20 problem instances in our computational experiments. All instances are solved by calling CPLEX 12.6 on Matlab R2016a that runs on a PC with an Intel i5-4590 CPU and 12 GB memory. In our computational experiments, we set the limit of computational times to 5 hours (i.e., $18,000$ seconds). For our developed algorithm, we set the absolute tolerance at 0.01. Because the optimal objective value from our method is almost the same (the absolute tolerance is within 0.1) with that from the benchmark (those instances can be solved within 5 hours), we do not display the optimal value in this work. We compare the average, minimum and maximum computational times, which are summarized in Table 1.

Table 1: Comparison on the computational time between our method and benchmark (in second)

patients-physician	Our method			Benchmark		
pairs	Min. time	Avg. time	Max. time	Min. time	Avg. time	Max. time
(20,2)	67	76	90	854	947	1,042
(30,3)	107	149	196	4,670	7,127	15,374
(40,4)	181	213	269	15,929	18,000	18,000
(50,5)	278	320	369	18,000	18,000	18,000
overall	158	189	231	9,863	11,018	13,104

As shown in Table 1, the average, minimum and maximum computational times of the proposed method are significantly shorter than the benchmark. Computational time increases with the increasing of n and k in pair (n, k) for both our method and the benchmark. As (n, k) increases, the computational time of our method increases slowly, while that of the benchmark increases rapidly. When (n, k) reaches (40, 4) and (50, 5), the benchmark cannot obtain the optimal solution before reaching the time limit (i.e., Ave.time =18,000), while our method can solve all problem instances to optimality in a reasonable computational time. These facts indicate that our method is indeed much more efficient than the benchmark. Moreover, there is no significant difference when $\lambda \geq 100$ according to the experimental results. Therefore, we test λ ranged from 0 to 100 in the following sections.

6.2. Analysis of λ in different scenarios

In this subsection, we analyze the effect of λ in different scenarios on the optimal solution and the objective value. Specifically, we fix the pair of patients and physicians as (20, 2) and test different values of matching cost $\mathbf{m}$ and λ. During the numerical experiments, we construct four different scenarios for the matching between patients and physicians, i.e., the number of patients who match the first physician best is $2, 4, 6, 8$, respectively, which we use $1:9$, $2:8$, $3:7$ and $4:6$ to represent those four scenarios. We refer to scenario $1:9$ as the most imbalanced patients-physician scenario. As for the matching cost $\mathbf{m}$, we randomly generate from $U[0.1, 0.3]$ if one patient match with the best-matched physician; otherwise, we randomly generate from $U[0.8, 1]$. Other parameters setting are the same as subsection 6.1.

We first study the variant of the total cost in different scenarios as λ increases. As shown in Figure 1, all matching patterns appear an increasing trend as λ increases. Furthermore, the more imbalanced of patients-physician matching, the higher total the cost it has. This because the imbalanced matching pattern would result in a higher matching cost to the optimal solution, and further amplify the total cost through λ.

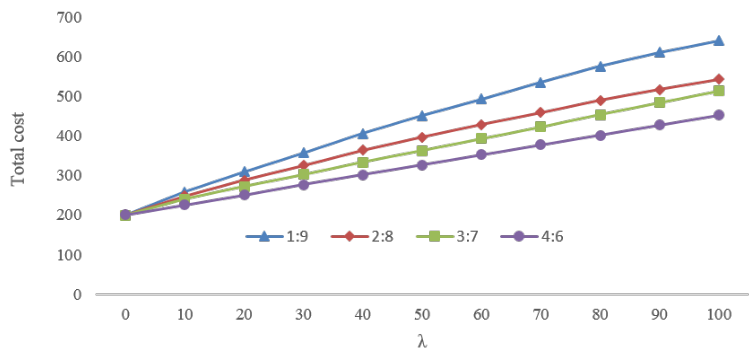
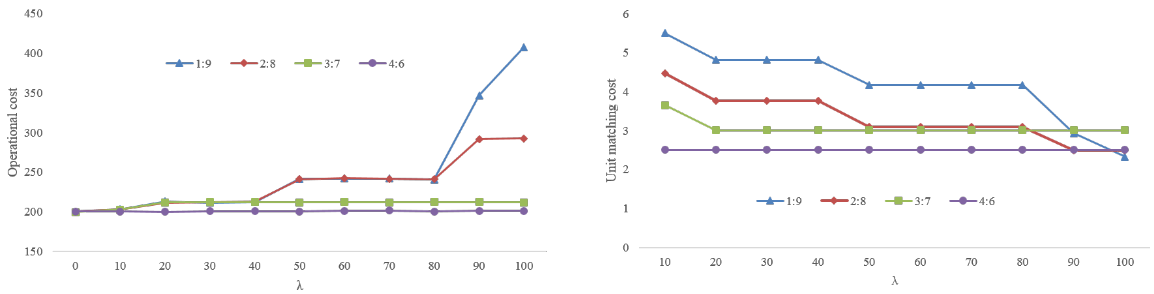


Figure 1: Comparison of the optimal total cost

Then, we break down the total cost and further study the variant of its corresponding operational and unit matching costs. The operational cost refers to the total expected weighted costs of patients' waiting times and service providers' idle times and overtimes under the optimal solution, i.e., $\sum_{j=1}^{k} \mathbb{E}\left[\sum_{i=1}^{n} (c_i^W x_{i,j}^* W_{i,j} + c_j^I I_{i,j}) + c_j^O O_j\right]$. The unit matching cost refers to the total matching cost without weighted by λ under the optimal solution, i.e., $\sum_{j=1}^{k} \sum_{i=1}^{n} m_{i,j} x_{i,j}^*$. As shown in Figure 2, both the operational cost and unit matching cost of the scenario $4:6$ remains stable as λ changes. However, for other matching scenarios, the operational costs increase as λ increases, while the unit matching costs exhibit an opposite trend. The reason is that when λ is large, the optimal assignment would result in an unbalance operational pattern to avoid a considerable matching cost, thus leads to higher operational costs. But for the more balanced matching scenario (e.g., scenario $4:6$), the weight λ has no significant impact on the assignment. Besides, both the operational cost and unit matching cost witness a more significant fluctuation when the matching scenario is unbalanced (e.g., scenario $1:9$). However, for the operational cost, it seems that the more stable matching scenario, the less operational cost it has. The difference of operational costs between scenarios $1:9$ and $4:6$ goes up to 51%, while the difference of unit matching costs between them goes up to 54%.

Moreover, we study the pattern of patients assigned to the first physician as λ changes for different matching scenarios. As shown in Figure 3, we can see that the number of patients assigned



a. Operational cost

b. Unit matching cost

Figure 2: Comparison of the operation cost and the unit matching cost

to the first physician exhibits a non-increasing trend as λ increases for all four matching scenarios. This result indicates that when the weight of matching is large, the optimal solution would assign more patients to the most matched physician. Otherwise, the optimal solution would balance the operational workload for each physician. Furthermore, when the scenario is more balanced (i.e., scenario $4:6$), λ is not a significant factor to the patient assignment. Another interesting observation is that there exist some overlaps for four scenarios and scenario $1:9$ overlaps all other scenarios. It may due to scenario $1:9$ is the most imbalanced case. When λ increases, patients tend to be assigned to the first physician, such that a higher matching cost is avoided.

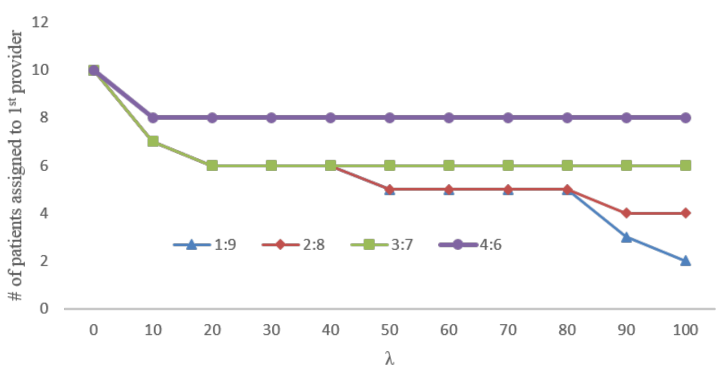
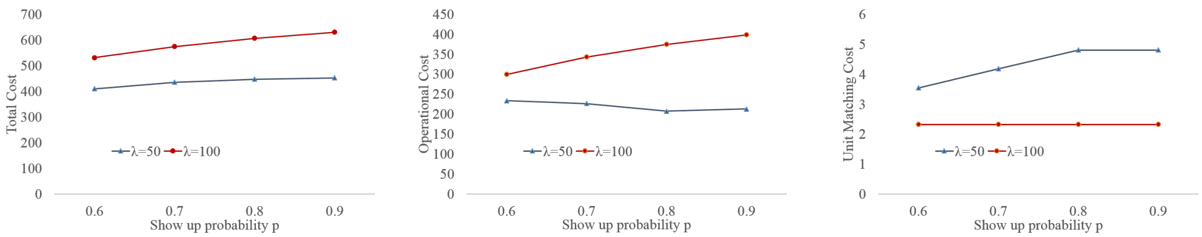


Figure 3: Comparison of the number of patients assigned to the first physician

6.3. Analyses of λ with no-shows

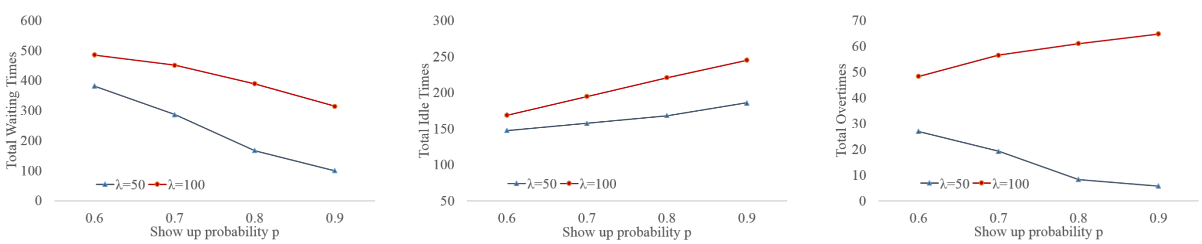
We further analyze the effect of no-shows on the optimal solution and the objective value. We assume an i.i.d. show up indicator, i.e., $p_i = p$, and set the show up probability p takes values from 0.6 to 0.9 with an increment of 0.1. By considering the no-shows, we reset the session length as $T = 1.5\mu pn/k$, such that it can adapt according to patients' no-show behaviors. From Figure 2a , we observe a more substantial fluctuation of the operational cost when the matching scenario is unbalanced. Thus, we set (n, k) at $(20, 2)$ and choose scenarios $1 : 9$ to analyze objective value and performance indicators. Moreover, there is a small variation of the operational cost when $\lambda \leq 50$. Therefore, to obtain an apparent comparison result for performance indicators, we test two different values of λ (i.e., $\lambda = 50, 100$), and examine the impact of λ on performance indicators. The other parameters (e.g., m, d, c_i^W, c_j^I, c_j^O) are the same with scenario $1 : 9$ in section 6.2.



a. Total cost b. Operational cost c. Unit matching cost

Figure 4: Comparison of the optimal cost with no-shows

The optimal total, operational, and unit matching costs are presented in Figure 4. As the same as we discussed in section 6.2, λ has the a positive impact on the total and operational costs and negative impact on the unit matching cost. It is intuitive that the total cost increases as the show up probability p increase, as shown in Figure 4a. We also observe the same effect of the show up probability on the unit matching cost when $\lambda = 50$. However, when λ is substantial (e.g., $\lambda = 100$), the show up probability has no influence on the unit matching cost, as shown in Figure 4c. It is more surprising to observe that the operational cost has an increasing trend when $\lambda = 100$ while an decreasing trend when $\lambda = 50$. Therefore, we decompose the operational cost and further analyze the total waiting times, idle times, and overtimes, as shown in Figure 5.



a. Total waiting times b. Total idle times c. Total overtimes

Figure 5: Comparison of operational indicators

From Figures 5a and 5b, we can see that the total waiting times under both values of λ decrease as the show up probability p increases, while the total idle times witness an opposite trend. This phenomenon can be explained as follows. In each configuration combination of (λ, p), the average job allowance for each patient can be approximated as $1.5\mu = 30$ roughly (Note that we set the session length as $T = 1.5 \cdot \mu \cdot p \cdot n/k$). However, the average service time is $\mu = 20$. It would be more idleness than waiting for the service system. Thus, as the show up probability increases, the number of show up patients increases, which leads to the decreasing of waiting times and increasing of idle times. This kind of trend is amplified as λ increase. Furthermore, from Figure 5c, we can observe that the total overtimes decrease when $\lambda = 50$ and increase when $\lambda = 100$ as the show up probability increases. The server system has a higher probability to suffer an overload when $\lambda = 100$ than $\lambda = 50$. Since the average job allowance is considerably longer than the average service time, the more show up patients, the more idleness, thus less overtimes when $\lambda = 50$. However, when $\lambda = 100$, the optimal assignment has a high chance to overload for one server. Therefore, more overtime may occur as the number of show up patients increases. As a result, when $\lambda = 100$, the total overtime increases as the show up probability p increases.

7. Managerial Implications

In this section, we summarize some managerial insights from the numerical results of this study. First, a more balanced supply and demand result in a lower total cost. From Figure 1, we can observe that for any fixed λ, the more imbalanced scenario would result in a higher total cost. This result indicates that if we want to achieve a lower total cost for the system, a more balanced scenario, i.e., scenario 4 : 6, is more desirable. Furthermore, as λ increases, the total cost increase for all scenarios. We also can observe that the more unbalanced scenario, i.e., scenario 1 : 9, the faster-increasing speed of total cost. This result implies that when matching becomes more important, the service provider should pay more effort to balance physicians' specialties and patients' diseases.

Second, the weight of the matching cost has a positive impact on the operational cost. From Figure 2, we can see that there is no significant difference for the operational costs among different scenarios when λ is small (e.g., $\lambda \leq 10$). As λ increases, the marginal operational cost increases while the marginal unit matching cost decreases. The operational cost increases at a swift speed as matching becomes more critical. It is because the overtime cost of matched physicians and waiting time of patients increases significantly while other mismatched physicians are idle for the unbalanced scenario, i.e., scenario 1 : 9.

Third, we would suggest the service provider to develop or train the specialty set based on the disease pattern of local patients, such that the workload among physicians can be more balanced. From Figure 3 we can see that when the weight λ is large (e.g., $\lambda = 100$), the workload among different service providers is imbalanced, which means there may be a waste of resources to some degree. Therefore, the specialty set of physicians is critical to balance the physician workload.

Fourth, when the weight of the matching cost is not substantial, we would suggest penalizing the no-show patient, such that the operational cost is minimized. However, when the weight of the matching cost is considerable, which means the matching is more critical to the health care quality, the patient has a higher motivation to see the matched physician. Therefore, the no-show has less influence on the service system.

8. Conclusion and future work

In this paper, we jointly optimize a matching problem and appointment problem with multiple service providers, in which the decision-maker determines how to assign patients to physicians and when to start serving patients for each service provider. We assume the service times are stochastic and all patients would arrive at the service system punctually at their scheduled times. The objective is to minimize total weighted matching costs and operational costs (patients' waiting time costs, service providers' idle time and overtime costs). To solve the problem, we first reformulate the studied problem as a two stage optimization problem based on the SAA approach. And then the properties for the optimal solution of the second stage problem are analyzed. On this basis, an improved benders decomposition algorithm is proposed to solve this problem efficiently. We also extend our method to incorporate no-shows. Finally, we conduct computational experiments to evaluate the efficiency of our proposed algorithm and investigate the variation of the optimal solutions yielded in different scenarios.

Our studies mainly show that: (1) the integrated matching problem and appointment scheduling can be formulated as a two-stage optimization problem; (2) the improved Benders decomposition algorithm is efficient to solve our studied problem in a reasonable time; (3) the optimal solution would assign patients to the most matched physicians if the matching cost dominate operational cost, otherwise, the optimal solution would balance the workload of physicians as much as possible; and (4) the no-show has less influence on the service system when the weight of the matching cost is substantial.

Our work can be extended in several aspects. We assume a fixed sequence for patients at each service provider, due to different type of patients, it may be valuable to optimize the sequence for patients at each service provider. Furthermore, unpunctuality is also inevitable in practice, which may lead to patients arriving out of order. In the future, we may also handle these stochastic factors.

Acknowledgment

The first author is supported by the the Science and Technology Foundation of Jiangxi Educational Committee (#GJJ190287). The second author is funded by NSFC #71801051, and the third author is funded by Omron research fund.

References

[1] M. Weiner, G. El Hoyek, L. Wang, P. R. Dexter, A. D. Zerr, A. J. Perkins, F. James, R. Juneja, A web-based generalist–specialist system to improve scheduling of outpatient specialty consultations in an academic center, Journal of General Internal Medicine 24 (6) (2009) 710–715.

[2] A. Mehrotra, C. B. Forrest, C. Y. Lin, Dropping the baton: specialty referrals in the united states, The Milbank quarterly 89 (1) (2011) 39–68.

[3] D. Li, S. Chen, X. Chen, C.-A. Chou, Learning and optimizing for the patient-physician matching problem in specialty care, SSRN (2019) Available at SSRN: https://ssrn.com/abstract=3450184.

[4] E. D. Güneş, H. Yaman, B. Çekyay, V. Verter, Matching patient and physician preferences in designing a primary care facility network, Journal of the Operational Research Society 65 (4) (2014) 483–496.

[5] R. Crow, H. Gage, S. Hampson, J. Hart, A. Kimber, L. Storey, H. Thomas, The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature., Health Technology Assessment 6 (32) (2002) 1–244.

[6] R. Beck, R. Daughtridge, P. Sloane, Physician–patient communication in the primary care office: a systematic review., The Journal of the American Board of Family Medicine 15 (2002) 25–38.

[7] J. Gong, H. Cheng, L. Wang, Individual doctor recommendation in large networks by constrained optimization, International Journal of Web Services Research 12 (2015) 16–28.

[8] N. Liu, S. R. Finkelstein, M. E. Kruk, D. Rosenthal, When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling, Management Science 64 (5) (2018) 1975–1996.

[9] K. Kinchen, L. Cooper, D. Levine, N.-Y. Wang, N. Powe, Referral of patients to specialists: Factors affecting choice of specialist by primary care physicians, Annals of family medicine 2 (2004) 245–52.

[10] X. Pan, J. Song, F. Zhang, Dynamic recommendation of physician assortment with patient preference learning, IEEE Transactions on Automation Science and Engineering 16 (1) (2019) 115–126.

[11] B. Denton, D. Gupta, A sequential bounding approach for optimal appointment scheduling, IIE Transactions 35 (11) (2003) 1003–1016.

[12] S. De Vuyst, H. Bruneel, D. Fiems, Computationally efficient evaluation of appointment schedules in health care, European Journal of Operational Research 237 (3) (2014) 1142–1154.

[13] E. N. Weiss, Models for determining estimated start times and case orderings in hospital operating rooms, IIE transactions 22 (2) (1990) 143–150.

[14] H.-S. Lau, A. H.-L. Lau, A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities, IIE Transactions 32 (9) (2000) 833–839.

[15] H.-Y. Mak, Y. Rong, J. Zhang, Sequencing appointments for service systems using inventory approximations, Manufacturing & Service Operations Management 16 (2) (2014) 251–262.

[16] A. Kuiper, B. Kemper, M. Mandjes, A computational approach to optimized appointment scheduling, Queueing Systems 79 (1) (2015) 5–36.

[17] R. Hassin, S. Mendel, Scheduling arrivals to queues: A single-server model with no-shows, Management Science 54 (3) (2008) 565–572.

[18] Q. Kong, C.-Y. Lee, C.-P. Teo, Z. Zheng, Scheduling arrivals to a stochastic service delivery system using copositive cones, Operations Research 61 (3) (2013) 711–726.

[19] H.-Y. Mak, Y. Rong, J. Zhang, Appointment scheduling with limited distributional information, Management Science 61 (2) (2014) 316–334.

[20] T. Cayirli, K. K. Yang, S. A. Quek, A universal appointment rule in the presence of no-shows and walk-ins, Production and Operations Management 21 (4) (2012) 682–697.

[21] L. W. Robinson, R. R. Chen, A comparison of traditional and open-access policies for appointment scheduling, Manufacturing & Service Operations Management 12 (2) (2010) 330–346.

[22] C. Zacharias, M. Pinedo, Appointment scheduling with no-shows and overbooking, Production and Operations Management 23 (5) (2014) 788–801.

[23] M. A. Begen, R. Levi, M. Queyranne, A sampling-based approach to appointment scheduling, Operations Research 60 (3) (2012) 675–681.

[24] L. W. Robinson, R. R. Chen, Scheduling doctors' appointments: optimal and empirically-based heuristic policies, IIE Transactions 35 (3) (2003) 295–307.

[25] G. C. Kaandorp, G. Koole, Optimal outpatient appointment scheduling, Health Care Management Science 10 (3) (2007) 217–229.

[26] B. Jiang, J. Tang, C. Yan, A stochastic programming model for outpatient appointment scheduling considering unpunctuality, Omega 82 (2019) 70–82.

[27] Y. Chen, Y.-H. Kuo, P. Fan, H. Balasubramanian, Appointment overbooking with different time slot structures, Computers & Industrial Engineering 124 (2018) 237–248.

[28] Y.-H. Kuo, H. Balasubramanian, Y. Chen, Medical appointment overbooking and optimal scheduling: tradeoffs between schedule efficiency and accessibility to service, Flexible Services and Manufacturing Journal (2019) 1–30.

[29] R. R. Chen, L. W. Robinson, Sequencing and scheduling appointments with potential call-in patients, Production and Operations Management 23 (9) (2014) 1522–1538.

[30] G. Xiao, M. Dong, J. Li, L. Sun, Scheduling routine and call-in clinical appointments with revisits, International Journal of Production Research 55 (6) (2017) 1767–1779.

[31] I. Bendavid, Y. N. Marmor, B. Shnits, Developing an optimal appointment scheduling for systems with rigid standby time under pre-determined quality of service, Flexible Services and Manufacturing Journal 30 (1-2) (2018) 54–77.

[32] 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, Service Science 10 (3) (2018) 241–260.

[33] B. Zheng, S. W. Yoon, M. T. Khasawneh, et al., An overbooking scheduling model for outpatient appointments in a multi-provider clinic, Operations Research for Health Care 6 (2015) 1–10.

[34] S. Sickinger, R. Kolisch, The performance of a generalized bailey–welch rule for outpatient appointment scheduling under inpatient and emergency demand, Health care management science 12 (4) (2009) 408.

[35] M. Soltani, M. Samorani, B. Kolfal, Appointment scheduling with multiple providers and stochastic service times, European Journal of Operational Research 277 (2) (2019) 667–683.

[36] C. Zacharias, M. Pinedo, Managing customer arrivals in service systems with multiple identical servers, Manufacturing & Service Operations Management 19 (4) (2017) 639–656.

[37] C. Mancilla, R. Storer, A sample average approximation approach to stochastic appointment sequencing and scheduling, IIE Transactions 44 (8) (2012) 655–670.

[38] E. N. Weiss, Models for determining estimated start times and case orderings in hospital operating rooms, IIE Transactions 22 (2) (1990) 143–150.