Two-Stage Robust Telemedicine Assignment Problem with Uncertain Service Duration and No-Show Behaviours

Menglei Ji * [1], Shanshan Wang † [2], Chun Peng ‡ [2], and Jinlin Li § [1]

[1]Beijing Institute of Technology, Beijing, China 100081
[2]HEC Montréal & GERAD, Montréal, Canada H3T 2A7

Abstract

The current pandemic of COVID-19 has caused significant strain on medical center resources, which are the main places to provide the rapid response to COVID-19 through the adoption of telemedicine. Thus it is crucial for healthcare managers to make an effective assignment plan for the patients and telemedical doctors when providing telemedicine services. Motivated by this, we present the first comprehensive study of a two-stage robust telemedicine assignment problem when three different sources of uncertainty are incorporated, including uncertain service duration, no-show behaviours of both patients and telemedical doctors. From an algorithmic viewpoint, we propose an efficient nested column-and-constraint generation (C&CG) solution scheme that decomposes the model into an outer level problem and an inner level problem. Our results show that we can solve the problems of realistic sizes within a reasonable time (e.g., up to 100 patients, 10 telemedical doctors, and 200 scenarios within two hours). On the empirical side, we demonstrate how the hyper-parameters make a balance between cost management and the coverage level of the served patients in the presence of three different sources of uncertainty. Our comparison with a two-stage stochastic programming model implies that our model is not overly conservative and seems to provide a relatively cheaper modeling alternative that requires much less information support when hedging against three different sources of uncertainty under a worst-case situation.

Key words: Telemedicine assignment, two-stage robust optimization, mixed-integer linear program, nested column-and-constraint generation, uncertain service duration, no-show behaviours.

1 Introduction

Telemedicine can be defined as the use of technology by medical professionals to deliver healthcare services (e.g., consulting, remote diagnosis, treatment, monitoring, and related follow-ups) for a remote location (Myers, 2003), which in some sense can provide rapid access to specialists who are not immediately available in person. The rise of the Internet Age brought with its profound changes in the practice of telemedicine. The proliferation of smart devices, capable of high-quality video transmission, opened up the possibility of delivering remote healthcare to patients in their homes, workplaces, or assisted living facilities as an alternative to in-person visits for both primary and specialty care. Telemedicine can be integrated into the healthcare system as an approach to maximize the efficiency of healthcare delivery. As we all know, the current pandemic of coronavirus disease 2019 (COVID-19) has caused significant strain on medical center resources. These hospitals can provide the necessary response to COVID-19 through the rapid adoption of digital tools and technologies such as telemedicine. Thus telemedicine/telehealth is contributing significantly to healthcare delivery during the COVID-19 crisis to meet patients' needs, reduce the transmission of the virus, and protect medical practitioners from infection while minimizing their exposure to the public and in-person visits. It is also considered as the doctors' first line of defense to control the spread of the coronavirus,

*Email: jimenglei@bit.edu.cn
†Email: shanshan.wang@hec.ca
‡Email: chun.peng@hec.ca
§Email: jinlinli@bit.edu.cn

keeping social distancing and providing services by phone or videoconferencing (e.g., Zhai et al., 2020; Jnr, 2020; Hollander and Carr, 2020).

Recently, National Telemedicine Center of China (NTCC) has explored various forms of collaborative services with provincial hospitals (called tertiary hospitals) and 120 regional cooperative hospitals (called primary hospitals). Compared with the provincial hospitals, regional cooperative hospitals are relatively small, as well as having less skilled doctors and less advanced medical equipment. In 2020, the number of telemedicine services is up to nearly 150 cases per day. Once both primary doctors and patients apply for the telemedicine service, NTCC will assign the doctors from tertiary hospitals to patients and schedule a teleconsultation service according to the available resources and needs in tertiary hospitals. Thus, similar to the classical healthcare assignment problems (e.g., Denton et al., 2010; Zhang et al., 2018a, 2020; Wang et al., 2021a,b), from a strategic/tactical level of viewpoint, our work will specifically focus on identifying the optimal assignment for the telemedical doctors and the patients within a fixed period under a highly uncertain decision-making environment, given that there are different sources of uncertainty from both telemedical doctors' and patients' sides, which in some sense makes the decision-making process more challenging. For instance, patients with scheduled appointments may not show up for their services for some reasons, such as referral, sudden death, equipment failure, and so on. Furthermore, most telemedical doctors in tertiary hospitals still provide the traditional outpatient medical services at the same time during the weekday, so they might not show up to provide the telemedicine service in the case of some emergency events that have a higher priority. This is one of the important features in telemedical healthcare resources allocation problems. Such cases regularly occur in practice, based on our field survey in several Chinese tertiary hospitals. We finally remark that it will be very interesting to incorporate the actual scheduling decisions and downstream resources constraints in the model, and we leave this for future research.

In the past decades, healthcare resource allocation problem under uncertainty has been widely studied in the literature (e.g., see review papers by Cardoen et al., 2010; Ferrand et al., 2014; Zhu et al., 2019). However, in terms of telemedicine operations management, there are very limited papers on this topic in the literature. Unlike the existing studies, we study a two-stage robust telemedicine assignment problem between telemedical doctors and patients while simultaneously incorporating three different sources of uncertainty in our mathematical model, including the no-show behaviours of both telemedical doctors and patients, as well as uncertain service duration. Our proposed model aims to minimize the weighted total cost of the normal case and worst case where some telemedical doctors do not show up due to unexpected requests. We aim to explore the optimal assignment of the telemedical doctors and patients in the first stage (normal case without telemedical doctor no-show). Once the information of the telemedical doctor no-show is realized, we then expect to re-schedule the assignment of the patients in the second stage.

Perhaps, the closest related studies in the literature consist in Erdogan et al. (2018) and Ji et al. (2020). Erdogan et al. (2018) study the telemedicine appointment problem that are quite different from our optimization model and beyond the scope of our research, while Ji et al. (2020) study a two-stage chance-constrained telemedicine assignment model with uncertain service duration and telemedical doctor no-show. Our work can be considered as an extension of Ji et al. (2020) in terms of the modeling framework and solution algorithm for the problems of a more realistic size.

To summarize, this paper addresses the resolution of a novel two-stage robust telemedicine assignment problem when taking into account different sources of uncertainty. We expect that this research could provide a solution tool for healthcare managers to better schedule the telemedicine services, especially during the Covid-19 pandemic. Although we address three different sources of uncertainty in a telemedical environment, these uncertainties do exist in a regular in-person doctor visit environment. We believe our proposed model and solution scheme can also be directly extended to the traditional doctor-patient assignment problems. Specifically, our contributions are summarized as follows:

- From a modeling viewpoint, we study a telemedicine assignment problem for the telemedical doctors and patients under three different sources of uncertainty, including patient no-show, uncertain service duration and no-show behaviour of the telemedical doctors. To our best knowledge, we are the first to study the robust telemedicine assignment problem with three sources of uncertainty.

- Methodologically speaking, we propose a two-stage robust telemedicine assignment model that minimizes the weighted total cost of the normal case and worst case when hedging against three different sources of uncertainty. In particular, we employ a "budgeted" uncertainty set to capture the no-show behaviour of the telemedical doctors, and a finite number scenarios to describe the no-show behaviour and service duration of the patients.

- From an algorithmic viewpoint, we develop an efficient nested C&CG method with symmetry-breaking constraints to solve our model. Specifically, we decompose the problem into two levels, i.e., outer level (first-stage problem) and inner level (recourse problem). We employ the C&CG to solve the recourse problem to identify the worst-case scenarios of the no-show behaviour of the telemedical doctors. We explore two approaches to solve the inner level problem, namely, Karush Kuhn Tucker (KKT) condition and strong duality. The numerical results show that our improved C&CG method can efficiently solve the problems (i.e., up to 100 patients, 10 telemedical doctors, and 200 scenarios) within two hours.

- Empirically speaking, we conduct an extensive numerical study to verify our proposed modeling framework. Our analyses imply that: 1) when the patient no-show rate is high, we might allow a limited number of telemedical doctors who show up to serve as many patients as possible while having a relatively cheaper expected total cost; 2) the higher patient no-show rates result in a decrease in the total expected cost, and the total expected cost will significantly increase when the number of telemedicine doctors who fail to show up exceeds a relatively high proportion (e.g., 60% in our numerical study); 3) our model is not overly conservative and seems to provide a relatively cheaper modeling alternative that requires much less information support when hedging against three different sources of uncertainty under a worst-case situation.

The remainder of the paper is organized as follows: Section 2 provides a brief review of relevant literature. Section 3 presents the modeling framework for a two-stage robust telemedicine assignment problem under three sources of uncertainty. Section 4 presents a nested C&CG solution scheme. We present the numerical results in Section 5 and give some concluding remarks in Section 6. Finally, we leave the supplementary materials in the appendix.

2 Literature Review

In this section, we briefly review two classes of literature that are closely related to our research, including the classical ORs assignment, surgery allocation problems that consider different sources of uncertainty (i.e., service duration, no-show behaviours, etc.) in Section 2.1, and the existing studies about telemedicine operations management in Section 2.2.

2.1 Literature on ORs Assignment and Surgery Allocation under Uncertainty

The telemedicine patients assignment problem is closely related to well-studied ORs allocation and scheduling problem. Recently, Cardoen et al. (2010), Ferrand et al. (2014) and Zhu et al. (2019) provide a comprehensive survey of research on ORs assignment and scheduling, as well as surgery allocation problems under uncertainty. We also remark that, since our work does not consider the appointment scheduling and sequencing decisions of the patients, we might not pay much attention to this streamline of literature in this section. Generally speaking, most of existing studies employ stochastic programming (SP) and robust optimization (RO) paradigms to model the inherent nature of uncertainty, e.g., uncertain surgery duration.

As one of the powerful tools to deal with uncertainty, SP (Kleywegt et al., 2002; Birge and Louveaux, 2011) has been widely used to model surgery allocation problems (e.g., Denton et al., 2010; Min and Yih, 2010; Batun et al., 2011; Gul et al., 2015; Kamran et al., 2018; Najjarbashi and Lim, 2019). For SP models, probability distributions are assumed to be known exactly in advance, which are generally estimated by the historical data. Methodologically speaking, this might give rise to two important practical issues in practice. First, the resolution of SP-based models can constitute a real computational challenge especially when the outcome space is continuous, necessitating the use of sample average approximation (SAA) schemes (Kleywegt et al., 2002). Second, it is usually impossible for the decision-makers to exactly know the true distribution of random variables. Instead, it is more common to only have very limited historical observations in practice.

Fortunately, these difficulties can in some sense be alleviated by using the RO paradigm (e.g., Bertsimas and Sim, 2004; Ben-Tal et al., 2009), where the uncertain parameters are modeled as belonging to uncertainty sets without assuming any distributional information. We refer the interested readers to a recent survey by

Gorissen et al. (2015) for more advances about RO. Therefore, there is also a line of studies about ORs allocation and scheduling using the RO tool to address the uncertainty (Addis et al., 2015), e.g., surgery duration. Denton et al. (2010) propose a two-stage robust formulation for the daily decisions of ORs opening and surgeries assignment with uncertain surgical durations. Their numerical experiments show that the robust method is computationally much faster than the stochastic method on average. Following Denton et al. (2010), Rath et al. (2017) propose a novel RO model with the decisions of simultaneous allocating and sequencing of surgeries, where the uncertainty set of surgical duration is constituted by the nominal value and maximum deviation. Addis et al. (2014) study a robust surgical case assignment problem when taking into account the variability on patient surgery duration. Holte and Mannino (2013) model both the uncertain and the cyclic allocation problems as an adjustable robust scheduling problem which can be solved by a row and column generation algorithm. Neyshabouri and Berg (2017) propose a two-stage RO model to address the existing uncertainty in surgery duration and length-of-stay in the surgical intensive care unit. More recently, Breuer et al. (2020) develop a robust model that combines staffing and scheduling decisions to minimize the impact of foreseeable variation in surgery durations, staff availability, and urgent or emergency arrivals. Bandi and Gupta (2020) study the ORs staffing and scheduling problem, and propose a new criterion (called "robust competitive ratio") for designing online algorithms.

However, the optimal solutions of classical RO models are always over-conservative. To bridge the gap between the conservatism of RO and the requirement of exact distribution in stochastic programming, distributionally robust optimization (e.g., Delage and Ye, 2010; Wiesemann et al., 2014) can seek the solution that performs best under the worst-case distribution within an ambiguity set that contains a family of probability distributions. We refer the interested readers for more details to a recent review by Rahimian and Mehrotra (2019). Such solutions are both robust to estimation error and converge to the true optimum as more distribution information is obtained. Therefore, there are also many distributionally robust models for ORs assignment and surgery allocation when the distribution of surgery duration is unknown but resides in an ambiguity set (e.g., Wang et al., 2019; Shehadeh and Padman, 2021). More recently, distributionally robust chance-constrained models are widely developed in literature while addressing that the probability of overtime for ORs is no more than a targeted risk level (e.g., Wang et al., 2017; Zhang et al., 2018a,b, 2020; Wang et al., 2021a,b; Wang and Mehrotra, 2021).

Finally, it is worth noting that, besides the uncertain service duration, our study is also related to the literature that addresses the no-show behaviours, which is also widely studied for the outpatient appointment scheduling problem. We refer the interested readers to Gupta and Denton (2008); Ahmadi-Javid et al. (2017); Dantas et al. (2018) for a more comprehensive study of appointment scheduling problems with no-show behaviours. Besides considering patient no-show, our study focuses on the telemedicine assignment problem, which further enables us to incorporate the no-show behaviour of the telemedical doctors, using a "budgeted" uncertainty set.

2.2 Literature on Telemedicine Operations Management

The existing studies on telemedicine mainly focuses on the technology (e.g., Baker and Stanley, 2018; Bahl et al., 2020), feasibility analysis of telemedicine (e.g., Jetty et al., 2018; Sun et al., 2020), as well as analysis on prevention and control of COVID-19 (e.g., Hollander and Carr, 2020; Jnr, 2020). The introduction of telemedicine is effective in reducing patients waiting time, improving social welfare, preventing unnecessary hospital access, and saving costs for the health system, especially in rural areas that lack medical resources (Zanaboni et al., 2009). More recently, the researchers show its great potential as an emerging technology with rich opportunities to model the healthcare operations management problems and further improve the quality of medical services (Dai and Tayur, 2020; KC et al., 2020).

From an operations management perspective, this streamline of research seems to be very rare. We briefly summarize the existing studies in Table 1. More specifically, Qiao et al. (2020) establish a queuing simulation system to search for the most reasonable resource allocation combination of telemedical doctors, while Wang et al. (2020) adopt a mixed duopoly game to obtain the optimal price as well as capacity decisions of the non-profit general hospital and the for-profit telemedicine firm. In order to improve triage decisions in telemedical physician triage, Saghafian et al. (2018) develop a novel optimization model of agent knowledge and deploy it in a partially observable Markov decision process (MDP) model to describe the optimal policy for deciding which cases (patients) to refer to the second level for further evaluation. Similarly, **?** also propose a MDP

Table 1: A summary of the existing studies on telemedicine operations management from OR/MS perspective in terms of problem descriptions, modeling techniques and solution methods.

Existing Study	Problem Description	Modeling Technique	Solution Method
Erdogan et al. (2018)	optimal appointments of telemedicine patients considering random service duration and patient no-show behaviour	two-stage stochastic linear program	GUROBI solver
?	assessing the overall costs and benefits of teletriage in health-care demand management	Markov decision process model	analytical solution
Saghafian et al. (2018)	the optimal policy for deciding which cases (patients) should be assigned to the telemedical physician for further evaluation	Markov decision process model	analytical solution
Wang et al. (2020)	optimal price and capacity decisions of non-profit general hospital and the for-profit telemedicine firm	mixed duopoly game model	analytical solution
Qiao et al. (2020)	the optimization of teleconsultation resources allocation	queuing model	discrete-event simulation
Rajan et al. (2019)	investigating the effect of telemedicine on chronic care	utility model using queue system	analytical solution
Ji et al. (2020)	telemedicine assignment with uncertain service duration and no-show behaviour of the doctors	two-stage chance-constrained model	enumeration C&CG
Our work	telemedicine assignment with uncertain service duration and no-show behaviours for both patients and telemedical doctors	two-stage robust optimization model	nested C&CG

model to determine in which cases teletriage is efficient and effective. Rajan et al. (2019) investigate the effect of telemedicine on chronic care and derive the analytical solution using a queue system based utility model.

The most relevant work to our study are Erdogan et al. (2018) and Ji et al. (2020). Specifically, Erdogan et al. (2018) present a two-stage stochastic linear program to derive optimal scheduling of telemedicine patients by considering cleaning of procedure devices and patient no-show behaviour. They employ a set of finite scenarios to capture the uncertain service duration and patient no-show behaviour. More recently, Ji et al. (2020) propose a two-stage chance-constrained model to study the telemedicine assignment between the patients and telemedical specialists by considering the doctors no-show behaviour and uncertain service duration. They develop an enumeration-based C&CG method to solve the resulting problem. By contrast, our work presents a two-stage RO model to address the telemedicine assignment problem, and employ a "budgeted" uncertainty set to capture the no-show behaviour of the telemedical doctors as well as a finite number scenarios to capture the random service duration and patient no-show. Moreover, we also design a more efficient C&CG method to solve the problem of realistic sizes.

3 Model Formulation

In this section, we first present all the notations for our model in Section 3.1, and we then propose a two-stage robust telemedicine assignment model with three different sources of uncertainty in Section 3.2. We are interested in exploring the optimal robust telemedicine assignment in a highly uncertain environment, especially when some telemedical doctors fail to show up due to the emergency events or requests.

Table 2: Notations

Sets	
$\mathcal{I}$	the set of patients, $i \in \mathcal{I}$
$\mathcal{J}$	the set of telemedical doctors, $j \in \mathcal{J}$
Ω	the set of scenarios, $\omega \in \Omega$
Parameters	
T	the length for a given time block
c_{ij}	assignment cost for telemedical doctor j who is assigned to serve patient i
h_j	working cost for telemedical doctor j in a time block of length T
K	the maximum number of assigned telemedical doctors
b_j	unit cost of overtime for the telemedical doctor j
r_i	penalty cost for patient i who can not be assigned at a time block
ρ	the weight for the total cost of normal case (without telemedical doctor no-show) and the worst case when telemedical doctor no-show is considered, $\rho \in [0, 1]$
d_i^ω	the service duration for patient i under scenario ω
v_j	the maximum tolerable overtime for the doctor j
p_ω	the probability that scenario ω occurrs
Decision Variables	
x_{ij}^ω	binary variables, $x_{ij} = 1$ if patient i is assigned to telemedical doctor j under scenario ω, and otherwise 0
y_j	binary variables, $y_j = 1$ if doctor j is assigned to provide telemedicine service, and otherwise 0
q_{ij}^ω	binary variables, $q_{ij}^\omega = 1$ if patient i is re-assigned to telemedical doctor j under scenario ω, and otherwise 0
u_i^ω	binary variables, $u_i^\omega = 1$ if patient i is unassigned under scenario ω, and otherwise 0
o_j^ω	continuous variables, the overtime of telemedical doctor j under scenario ω

3.1 Notations

We aim to explore the optimal assignment decisions between the telemedical doctors and patients for a given time block. During the time block with a fixed length, a set of patients are assigned to a set of telemedical doctors to obtain the telemedicine services. We describe all the notations that are used throughout the rest of the paper in Table 2. Like Soltani et al. (2019), Zacharias and Pinedo (2017), and Zheng et al. (2015), we assume multiple-provider systems with identical providers. There is no difference among different providers in terms of the quality of care and the service time. Thus the random service durations d_i^ω are server-independent. For our study, the assignment cost c_{ij} between patient i and doctor j refers to the related costs that are involved in preparing, operating, and managing the telemedicine assignment by the medical center (e.g., NTCC) in practice, which seems to be commonly used in the classical healthcare assignment literature (e.g., Zhang et al., 2018a; Wang et al., 2021a; Min and Yih, 2010; Neyshabouri and Berg, 2017). We also note that the assignment cost c_{ij} might also be dependent on the priority score that is given to a patient, different specialties, the knowledge of the doctors, and so on.

3.2 Two-Stage Robust Optimization Model

As we know, telemedical doctors are one of the most important healthcare resources in the operations of providing telemedicine services. Despite the rapid development of telemedicine in recent years, telemedical resources (e.g., doctors) are still very scarce for most developing countries (e.g., China). However, in practice some of the telemedical doctors fail to show up before performing the telemedicine service (e.g., the remote consultation), because of an emergency event (e.g., urgent surgery) or a scheduling conflict. Based on our investigation in our collaborative healthcare providers, such cases regularly occur in the real world. Given the scarce resources, it seems to be not practical to find another doctor to fully replace the no-show telemedical doctor's job. This further motivates us to propose a novel optimization model to explore the optimal telemedicine assignment after considering the no-show behaviour of the telemedical doctors. In our study, it is reasonable to assume that if a doctor fails to show up due to some unexpected cases, he/she will not provide the telemedical service during a given time block (e.g., 3 hours).

Leveraging the recent advances in robust optimization, we propose a two-stage robust optimization model, which minimizes the weighted total cost of the normal case and the worst case. Specifically, in the first stage we want to explore the optimal assignment decisions of telemedical doctors and patients in the normal case when considering uncertain service duration and patient no-show. Once the realizations of the no-show behaviour of the telemedical doctors are realized, we need to re-schedule the assignment of the patients over the worst-case scenarios in the second stage, in which the overtime of the telemedical doctors is allowed in order to serve as many patients.

We assume the uncertain service duration $\boldsymbol{d}$ with a distribution that is finitely known with N historical scenarios, namely, $\{\boldsymbol{d}_\omega\}_{\omega\in\Omega}$ where $\Omega = \{1, 2, \cdots, N\}$. Each scenario with probability p_ω, such that $\sum_{\omega\in\Omega} p_\omega = 1$. In order to capture the uncertainty of patient no-show, we introduce δ_i^ω, a random vector of indicators for the patient who shows up i ($\delta_i^\omega = 1$) and who does not ($\delta_i^\omega = 0$) under scenario ω, which occurs with probability $1 - \mathrm{p}_{\mathrm{no}}$, and where p_{no} denotes the probability of no-show. One can easily realize that the service duration for the patients who are no-shows can be considered as zero under scenario ω, given that it can be represented as $\delta_i^\omega d_i^\omega$.

As we know that it is nearly impossible to exactly describe the distribution of the telemedical doctor no-show in practice, even for the limited distributional information. Therefore, we employ a cardinality constraint set (also called a "budgeted" uncertainty set with binary variables) to capture the no-show behaviour of the telemedical doctors for our two-stage robust optimization model. Specifically, our uncertainty set is explicitly represented as

$$\mathcal{Z} := \{\boldsymbol{z} \in \{0,1\}^{|\mathcal{J}|},\ \sum_{j\in\mathcal{J}} z_j \le \Gamma\},$$

where $z_j = 1$ if telemedical doctor j does not show up, and $z_j = 0$ otherwise. Parameter Γ can be considered as the budget of uncertainty for telemedical doctor no-show, which controls the maximum number of the telemedical doctors with no-show behaviour. It is easy to know that Γ is integer-valued within the interval $[0, K]$. The uncertainty set adjusts the robustness against the conservation level of the solutions by sizing Γ. Note that, if $\Gamma = 0$, all the telemedical doctors show up, i.e., $z_j = 0$ for all $j \in \mathcal{J}$ (see Ramark 1 below). On the other hand, if $\Gamma = K$, it indicates that all K assigned the telemedical doctors in the normal case do not show up, which however rarely to happens in practice.

Therefore, the two-stage robust optimization model can be represented as follows:

$$\begin{aligned}
[\text{2RO}]:\ \underset{\boldsymbol{x},\boldsymbol{y}}{\text{minimize}}\ & \rho\left(\sum_{\omega\in\Omega} p_\omega \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} c_{ij} x_{ij}^\omega + \sum_{j\in\mathcal{J}} h_j y_j\right) \\
& + (1-\rho)\mathcal{S}(\boldsymbol{y}) && && \text{(1a)} \\
\text{subject to}\ & x_{ij}^\omega \le y_j && \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega && \text{(1b)} \\
& \sum_{j\in\mathcal{J}} x_{ij}^\omega = \delta_i^\omega && \forall i \in \mathcal{I}, \omega \in \Omega && \text{(1c)} \\
& \sum_{j\in\mathcal{J}} y_j \le K && && \text{(1d)} \\
& \sum_{i\in\mathcal{I}} d_i^\omega x_{ij}^\omega \le T y_j && \forall j \in \mathcal{J}, \omega \in \Omega && \text{(1e)} \\
& x_{ij}^\omega \in \{0,1\}, y_j \in \{0,1\} && \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega, && \text{(1f)}
\end{aligned}$$

where

$$\mathcal{S}(\boldsymbol{y}) = \underset{\boldsymbol{z}\in\mathcal{Z}}{\text{maximize}}\ \underset{\boldsymbol{q},\boldsymbol{u},\boldsymbol{o}}{\text{minimize}} \sum_{\omega\in\Omega} p_\omega \left(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} c_{ij} q_{ij}^\omega(\boldsymbol{z}) + \sum_{j\in\mathcal{J}} b_j o_j^\omega(\boldsymbol{z})\right)$$

$$+\sum_{\omega\in\Omega} p_\omega \sum_{i\in\mathcal{I}} r_i u_i^\omega(\boldsymbol{z}) \tag{1g}$$

$$\text{subject to} \sum_{j\in\mathcal{J}} q_{ij}^\omega(\boldsymbol{z}) + u_i^\omega(\boldsymbol{z}) = \delta_i^\omega \qquad \forall i\in\mathcal{I}, \omega\in\Omega \tag{1h}$$

$$\sum_{i\in\mathcal{I}} d_i^\omega q_{ij}^\omega(\boldsymbol{z}) \le T y_j(1-z_j) + o_j^\omega(\boldsymbol{z}) \qquad \forall j\in\mathcal{J}, \omega\in\Omega \tag{1i}$$

$$o_j^\omega(\boldsymbol{z}) \le v_j y_j(1-z_j) \qquad \forall j\in\mathcal{J}, \omega\in\Omega \tag{1j}$$

$$q_{ij}^\omega(\boldsymbol{z}) \in \{0,1\}, u_i^\omega(\boldsymbol{z}) \in \{0,1\}, o_j^\omega(\boldsymbol{z}) \ge 0 \qquad \forall i\in\mathcal{I}, j\in\mathcal{J}, \omega\in\Omega. \tag{1k}$$

The objective function (1a) is to minimize the weighted sum of the total expected cost for the normal case and worst case when the no-show behaviour of the telemedical doctors is realized, including the total expected assignment cost and the working cost of the telemedical doctors in the first stage, as well as the total expected re-assignment cost and penalty cost incurred by the unserved patients in the second stage (i.e., $\mathcal{S}(\boldsymbol{y})$). Parameter ρ indicates the decision-makers risk preference towards the total expected re-assignment cost in the worst case. The smaller ρ is, the more risk-averse and conservative the model is. Constraints (1b) and (1c) restrict that all the patients who show up must be assigned under scenario ω and only the telemedical doctor who is assigned to work can provide the telemedical services. Constraint (1d) restricts that the number of assigned telemedical doctor is no more than K ($K \le |\mathcal{J}|$). Constraint (1e) ensures that the total working time for telemedical doctor j can not exceed the fixed length of each block. Constraint (1f) enforces all the variables to be binaries.

For the second-stage problem $\mathcal{S}(\boldsymbol{y})$, given the assigned telemedical doctors (i.e., y_j), the objective function (1g) is to minimize the total expected cost in the worst case (i.e., including the expected re-assignment cost, overtime cost, and penalty cost for the unserved patients) when telemedical doctor no-shows are considered, in order to determine the optimal decisions of the re-assignment, overtime, and the unassigned patients. Constraint (1h) shows that patient i would be re-assigned to the telemedical doctor j who shows up under scenario ω, and otherwise will be unserved due to the shortage of the telemedical doctors. Here we include a large penalty cost for the unserved patients in (1g). Constraint (1i) calculates the overtime of the telemedical doctor j, which is based on the fixed length T of each block, the realizations of service duration, as well as no-show behaviours of the patients and telemedical doctors. Constraint (1j) indicates that a telemedical doctor j can provide the telemedicine service during the overtime if and only if he/she does show up or is assigned in the first-stage problem (namely, $y_j = 1$ and $z_j = 0$), and the overtime of telemedical doctor j should not exceed v_j. Finally, constraint (1k) ensures the non-negativity of $\boldsymbol{o}(\boldsymbol{z})$ and integer on $\boldsymbol{q}(\boldsymbol{z})$ and $\boldsymbol{u}(\boldsymbol{z})$ in the second-stage problem.

Remark 1. *For our model (1), we remark that, if $\Gamma = 0$, it implies that all the telemedical doctors do show up (i.e., $z_j = 0, \forall j \in \mathcal{J}$). Then all patients can be served in the first stage. For such a case, there is no need to re-schedule patients in the second stage. Therefore, ρ is set to 0 if $\Gamma = 0$.*

Remark 2. *Note that, the second-stage decisions $q_{ij}^\omega(\boldsymbol{z})$,$u_i^\omega(\boldsymbol{z})$, $o_j^\omega(\boldsymbol{z})$ are made, once the uncertainty of the telemedical doctor no-show is realized. For the ease of exposition, we use a simple set of alternative notation (i.e., q_{ij}^ω, u_i^ω, and o_j^ω) for the rest of the paper.*

In order to illustrate the no-show behaviour of the telemedical doctors in our optimization model, we give the following example under a very simple setting while we assume that all the patients will show up. Example 1 highlights the importance and necessity to take the no-show behaviour of the telemedical doctors into account to hedge against the high penalty cost.

Example 1. *As is illustrated in Figure 1, suppose that one plans to assign four patients to two telemedical doctors. The assignment cost of patients are symmetric, namely, $c_{11} = c_{12} = c_{21} = c_{22} = c_{31} = c_{32} = 2$. The overtime cost is set to $b_j = 2$, the working cost h_j is 20, and the penalty cost r_i for the unserved patients to 60. We set the weight $\rho = 0.5$, the length of time block $T = 20$, and we also assume the telemedical doctors can work overtime unlimitedly. We consider two empirical scenarios with equal probability. For the 1st scenario, the service durations for the patients are $d_1^1 = 10$, $d_2^1 = 7$, $d_3^1 = 8$, and $d_4^1 = 8$, while for the 2nd scenario, the service durations are $d_1^2 = 9$, $d_2^2 = 7$, $d_3^2 = 6$, and $d_4^2 = 8$. As is assumed without considering the no-show behaviour of doctors, one of the optimal assignment decisions is as shown in Figure 1(a) and* *the total cost is 48. Nevertheless, suppose $\Gamma = 1$, which means that at most a telemedical doctor fails to show up, thus two patients have to suffer from the risk of being unserved due to the lack of doctors, which leads to a high penalty cost (i.e.* 120*). In this regard, under the worst-case scenario, we re-assign all patients to one telemedical doctor and the total cost is* 52 *in Figure 1(b). Although the total cost is increased by 8.33%, we could mitigate the high risk of having a very expensive worst-case cost by re-assigning patients when observing the no-show behaviour of the telemedical doctors.*

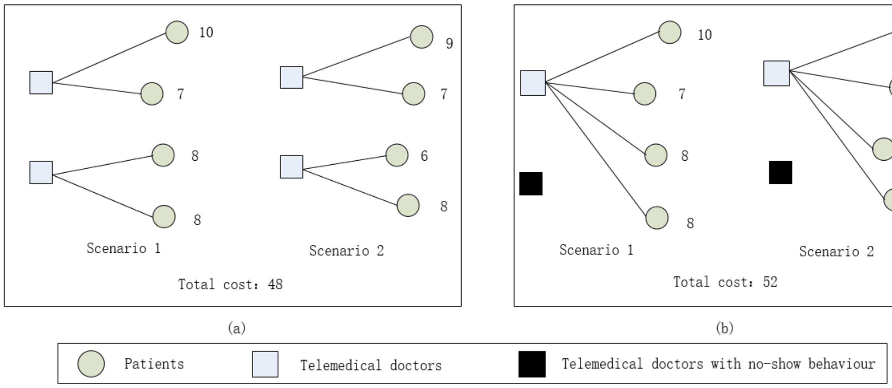


Figure 1: A simple illustrative example

We note that our model (1) can be considered as a tri-level optimization problem with mixed-integer variables, which specifically takes the form of $\min_{\boldsymbol{x},\boldsymbol{y}}$-$\sup_{\boldsymbol{z}\in\mathcal{Z}}$-$\min_{\boldsymbol{q},\boldsymbol{u},\boldsymbol{o}}$. For any given assignment decisions ($\boldsymbol{x}$, $\boldsymbol{y}$) generated by the first-stage problem, the second-stage problem seems to be always feasible, given the fact that we penalize the unassigned patients in the objective function. Thus the relatively complete recourse condition is satisfied for our problem. Nevertheless, it is still computationally challenging for the problems of a more realistic size. The strong duality can not be directly applied to the recourse problem, due to a max-min optimization problem with mixed-integer variables (i.e. $\boldsymbol{q} \in \{0, 1\}$, $\boldsymbol{u} \in \{0, 1\}$). This restriction also prevents us using the standard decomposition methods (e.g., Benders decomposition (Benders, 1962; Rahmaniani et al., 2017; Peng et al., 2020), L-shaped method (Laporte and Louveaux, 1993; Kim and Mehrotra, 2015), classical C&CG method (Zeng and Zhao, 2013)) to solve our model (1). Finally, the state-of-art solvers (e.g., CPLEX, GUROBI) can not be directly used to solve model (1).

We emphasize that Ji et al. (2020) propose an enumeration-based C&CG algorithm to solve their two-stage chance-constrained SP model with mixed-integer decisions, which seems to be used to solve our model (1). However, it appears to be computationally challenging for large-size problems. This further motivates us to develop an efficient solution scheme for solving model (1) of realistic sizes within a reasonable time, which will be discussed in Section 4.

4 A Nested Column-and-Constraint Generation Solution Scheme

While we know that two-stage robust optimization problems are generally computationally intractable and NP-hard (Ben-Tal et al., 2004), a method known as C&CG, firstly introduced by (Zhao and Zeng, 2012; Zeng and Zhao, 2013), has achieved good numerical performance under mild conditions and has been recently applied in various applications in the past. The main idea of the C&CG is similar to Benders decomposition approach (Benders, 1962), which is also implemented in a master-subproblem framework. Since the uncertainty set is polyhedral with a finite number of extreme points, C&CG converges to the true optimal value within a finite number of iterations.

However, as we discussed before, our proposed two-stage robust model is involved with binary recourse variables $\mathbf{q}$ and $\mathbf{u}$, which makes the problem non-convex and the strong duality theorem infeasible. In this regard, the classical C&CG method can not be directly employed to solve our two-stage robust model. By exploiting the special structure of our two-stage robust optimization model (1), the recourse problem can also be treated as a two-stage max-min optimization problem. Therefore, in the following we propose an adapted version of C&CG method (called nested C&CG) to solve our two-stage robust formulation.

In this section, we discuss how to use the nested C&CG to solve our two-stage robust telemedicine assignment problem efficiently, in which the whole problem can be decomposed into the outer level problem in Section 4.1 and inner level problem in Section 4.2, both of which can be solved by C&CG algorithm. Specifically, the outer level problem determines the assignment by solving the first-stage problem with the worst-case scenarios that are identified by the sub-problem. The inner level problem will try to iteratively solve the second-stage problem to find the worst-case scenarios for the given assignment policy (i.e., $\boldsymbol{y}$) from the outer level problem. In order to speed up our algorithm, we also propose a set of symmetry-breaking constraints in Section 4.3.

4.1 Outer Level C&CG Solution Framework

The outer level of our nested C&CG algorithmic framework actually is a standard C&CG implementation procedure that consists of a master-subproblem iterative process. The main idea is that the two-stage robust model comprises a Master Problem (denoted by MP) and a Sub-Problem (denoted by SP) that takes the form of max-min optimization problem with binary decision variables. The MP and SP are presented by model (2) and model (3), respectively.

We now briefly show how the outer level C&CG solution algorithm works. The algorithm first solves the MP (2) to obtain the optimal solutions of $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, as well as the optimal objective value that provides a lower bound. Then given the $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, one needs to solve the SP (3) to obtain the optimal objective value that returns an upper bound, and the worst-case scenarios $\boldsymbol{z}$. After obtaining the worst-case scenarios and upper bound from SP, the algorithm will terminate if either the sub-optimality gap is no more than the predefined tolerance or the solution time reaches the time limit, and otherwise, a set of extra variables q_{ij}^{ω}, u_i^{ω}, o_j^{ω} are generated and a number of constraints (2b)-(2f) is added to the MP in order to obtain a better solution and further improve the lower bound. Since the relatively complete recourse holds for our problem, so we do not generate the feasibility cuts. Let LB and UB represent the lower bound and upper bound of the outer level solution algorithm, respectively. Let m be the running index of iterations, and ℓ represent the running index of the set of extreme points (worst-case scenarios) that are derived by the SP . We use $\boldsymbol{q}^l$, $\boldsymbol{u}^l$, and $\boldsymbol{o}^l$ to represent the new variables that are associated with the ℓ-th scenario ($\ell \in \{1, \cdots, m\}$), and $\boldsymbol{z}^l$ to represent the worst-case scenarios of the telemedical doctors under the ℓ-th scenario (i.e., show-up or no-show). Finally, the procedures of our outer level C&CG solution framework are summarized in Algorithm 1. Given that the extreme points of the feasible region are finite, Algorithm 1 will converge within a finite number of iterations.

$$
\begin{aligned}
[\mathbf{MP}]: \quad \underset{\boldsymbol{x},\boldsymbol{y}}{\text{minimize}} \quad & \rho \sum_{\omega \in \Omega} p_{\omega} \left(\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} c_{ij} x_{ij}^{\omega} + \sum_{j \in \mathcal{J}} h_j y_j \right) \\
& + (1-\rho)\eta && && \text{(2a)} \\
\text{subject to} \quad & \text{(1b)} - \text{(1d)} \\
& \sum_{j \in \mathcal{J}} q_{ij}^{\omega\ell} + u_i^{\omega\ell} = \delta_i^{\omega} && \forall i \in \mathcal{I}, \omega \in \Omega, \ell \le m && \text{(2b)} \\
& \sum_{i \in \mathcal{I}} d_i^{\omega} q_{ij}^{\omega\ell} \le T y_j (1 - \hat{z}_j) + o_j^{\omega\ell} && \forall j \in \mathcal{J}, \omega \in \Omega, \ell \le m && \text{(2c)} \\
& o_j^{\omega\ell} \le v_j y_j (1 - \hat{z}_j) && \forall j \in \mathcal{J}, \omega \in \Omega, \ell \le m && \text{(2d)} \\
& \eta \ge \sum_{\omega \in \Omega} p_{\omega} \left(\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} c_{ij} q_{ij}^{\omega\ell} + \sum_{j \in \mathcal{J}} b_j o_j^{\omega\ell} \right) \\
& \quad + \sum_{\omega \in \Omega} p_{\omega} \sum_{i \in \mathcal{I}} r_i u_i^{\omega\ell} && \forall \ell \le m && \text{(2e)} \\
& x_{ij}^{\omega} \in \{0,1\}, y_j \in \{0,1\}, \eta \in \mathbb{R} && \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega \\
& q_{ij}^{\omega} \in \{0,1\}, u_i^{\omega} \in \{0,1\}, o_j^{\omega} \ge 0 && && \text{(2f)}
\end{aligned}
$$

$$
\begin{aligned}
[\mathbf{SP}]: \ \underset{\boldsymbol{z}\in\mathcal{Z}}{\text{maximize}}\,\underset{\boldsymbol{q},\boldsymbol{u},\boldsymbol{o}}{\text{minimize}}\ & \sum_{\omega\in\Omega} p_\omega\left(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} c_{ij}q_{ij}^\omega + \sum_{j\in\mathcal{J}} b_j o_j^\omega + \sum_{i\in\mathcal{I}} r_i u_i^\omega\right) && \text{(3a)}\\
\text{subject to}\ & \sum_{j\in\mathcal{J}} q_{ij}^\omega + u_i^\omega = \delta_i^\omega && \forall i\in\mathcal{I},\omega\in\Omega \quad \text{(3b)}\\
& \sum_{i\in\mathcal{I}} d_i^\omega q_{ij}^\omega \le T\hat{y}_j(1-z_j) + o_j^\omega && \forall j\in\mathcal{J},\omega\in\Omega \quad \text{(3c)}\\
& o_j^\omega \le v_j\hat{y}_j(1-z_j) && \forall j\in\mathcal{J},\omega\in\Omega \quad \text{(3d)}\\
& q_{ij}^\omega\in\{0,1\}, u_i^\omega\in\{0,1\} && \forall i\in\mathcal{I}, j\in\mathcal{J},\omega\in\Omega \quad \text{(3e)}\\
& o_j^\omega \ge 0 && \forall j\in\mathcal{J},\omega\in\Omega \quad \text{(3f)}
\end{aligned}
$$

Algorithm 1 The Outer Level Implementation of the Nested C&CG Algorithmic Framework

```
1: Initialize A tolerance $\epsilon \ge 0$ and maximum run time stoptime.
2: Initialize $m = 0, \mathrm{LB} = -\infty, \mathrm{UB} = +\infty$.
3: while (runtime $\le$ stoptime and $|\frac{\mathrm{UB}-\mathrm{LB}}{\mathrm{UB}}| > \epsilon$) do
4:     Solve the MP (2).
5:     Record optimal solution $(\boldsymbol{x}^m, \boldsymbol{y}^m, \eta^m)$ and optimal objective $lobj^m$.
6:     Update LB := $lobj^m$.
7:     Fix $\boldsymbol{y} := \boldsymbol{y}^m$, and solve the SP (3).
8:     Obtain the worst-case cost $\mathcal{Q}^m$ with the selected scenario $\boldsymbol{z}^m$.
9:     Update UB: $=\min\left\{\mathrm{UB}, \rho\min_{\boldsymbol{x},\boldsymbol{y}}\sum_{\omega\in\Omega}p_\omega\left(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}c_{ij}x_{ij}^\omega + \sum_{j\in\mathcal{J}}h_jy_j\right) + (1-\rho)\mathcal{Q}^m\right\}$.
10:    Create new integer variables $\boldsymbol{q}^{m+1}$, $\boldsymbol{u}^{m+1}$ and add constraints (2b)-(2f) to the MP .
11:    Set $m := m + 1$.
12: end while
13: return UB and corresponding optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ for which $obj^* = \mathrm{UB}$.
```

4.2 Inner Level C&CG Solution Framework

In this section, we show how to solve the inner level problem that takes the form of a *max-min* mixed-integer linear program. A key step is to identify the worst-case scenarios. Based on the structure of the inner level problem (i.e., SP), we can rewrite the SP as a tri-level equivalent formulation that takes the form of *max-min-min*, as follows:

$$
\underset{\boldsymbol{z}\in\mathcal{Z}}{\text{maximize}}\,\underset{\boldsymbol{q},\boldsymbol{u}\in\mathcal{Q}}{\text{minimize}}\ \sum_{\omega\in\Omega} p_\omega\left(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} c_{ij}q_{ij}^\omega + \sum_{i\in\mathcal{I}} r_i u_i^\omega\right) + \underset{\boldsymbol{o}\in\mathcal{O}(\boldsymbol{y},\boldsymbol{q},\boldsymbol{u},\boldsymbol{d},\boldsymbol{z})}{\text{minimize}} \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}} p_\omega b_j o_j^\omega,
$$

where $\mathcal{Q} := \{(\boldsymbol{q},\boldsymbol{u})\,|\,(3\text{b}),(3\text{e})\}$ and $\mathcal{O}(\boldsymbol{y},\boldsymbol{q},\boldsymbol{u},\boldsymbol{d},\boldsymbol{z}) := \{\boldsymbol{o}\,|\,(3\text{c}),(3\text{d}),(3\text{f})\}$. Moreover, we observe that it is derived by separating the integer variables (i.e., $\boldsymbol{q}$, $\boldsymbol{u}$) and continuous variables (i.e., $\boldsymbol{o}$). This motivates us to employ a C&CG method to solve the inner level problem. If the binary decision variables are given, one can reformulate the subproblem (3) as a maximization problem by using KKT condition or strong duality.

Based on these thoughts, the SP can also be implemented by C&CG within a master-subproblem framework. Specifically, one can first solve the master problem (denoted by $\mathrm{MP_S}$) to obtain the optimal solutions of $\boldsymbol{q}$ and $\boldsymbol{u}$ which also returns the upper bound of the inner level problem, given the *max-min* structure. Then for the given binary variables $\boldsymbol{q}$ and $\boldsymbol{u}$, one can solve the $\mathrm{MP_S}$ to identify the worst scenarios and pass them to the subproblem (denoted by $\mathrm{SP_S}$). Then one can further solve the $\mathrm{SP_S}$ and obtain the lower bound of the inner level problem. This process is repeatedly implemented until the convergence. Let m' be the index of iteration, $\mathrm{LB_S}$ and $\mathrm{UB_S}$ represent the upper bound and upper bound of the inner level problem.

Here we let ℓ' represent the running index of the set of extreme points (worst-case scenarios) that is derived by the $\mathrm{SP_S}$. We summarize the detailed procedures of the inner level C&CG implementation in Algorithm 2. In the following, we present two methods (namely, the KKT condition in Section 4.2.1 and the strong duality in Section 4.2.2) to solve the inner level problem. Finally, we summarize an overview of the proposed algorithmic solution framework in Figure 2, where Gap1 and Gap2 denote the relative optimality gap of the outer level and inner level problem, respectively.

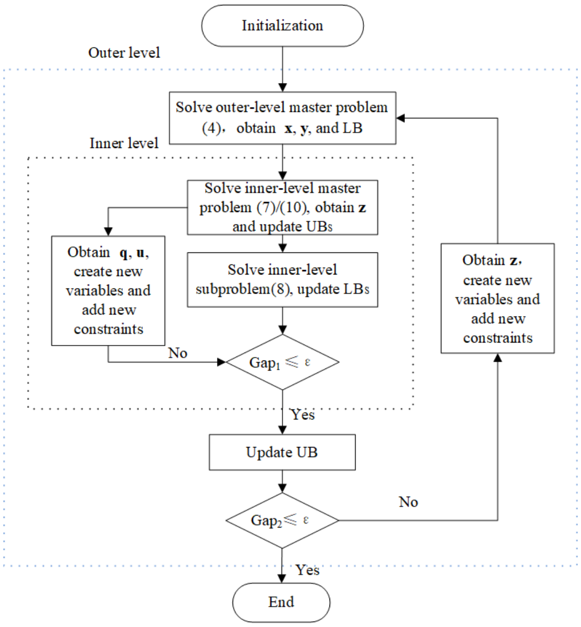

Initialization
Outer level
Solve outer-level master problem (4), obtain **x**, **y**, and LB
Inner level
Solve inner-level master problem (7)/(10), obtain **z** and update UBs
Obtain **q**, **u**, create new variables and add new constraints
Solve inner-level subproblem(8), update LBs
No
$\mathrm{Gap}_1 \leqslant \varepsilon$
Yes
Update UB
$\mathrm{Gap}_2 \leqslant \varepsilon$
No
Obtain **z**, create new variables and add new constraints
Yes
End

Figure 2: The framework of our nested C&CG solution scheme

4.2.1 KKT Condition based Inner Level Method

In this section, we use the KKT condition to derive the $\mathrm{MP_S}$ if the first-stage decision $\hat{\boldsymbol{y}}$ is given. For the given values of $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{u}}$, let α_j^ω and β_j^ω be the dual variables in terms of constraints (3c) and (3d), respectively. Then, for the ℓ'-th iteration of our nested C&CG algorithm, we rewrite the objective function (3a) as its epigraph form (i.e., constraint (4b)), then we derive the KKT condition, which gives rise to the following maximization-based MILP problem (4).

$$\text{maximize} \;\; \rho\,\theta \tag{4a}$$

$$\begin{aligned}\text{subject to} \;\; \theta \le & \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} p_\omega c_{ij}\hat{q}_{ij}^\omega + \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}} p_\omega r_i \hat{u}_i^\omega \\ & + \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}} p_\omega b_j o_j^{\omega\ell'} \qquad \forall \ell' \le m'\end{aligned} \tag{4b}$$

$$\alpha_j^{\omega\ell'} - \beta_j^{\omega\ell'} \le p_\omega b_j \qquad \forall j\in\mathcal{J}, \omega\in\Omega, \ell'\le m' \tag{4c}$$

$$o_j^{\omega\ell'} \ge \sum_{i\in\mathcal{I}} d_i^\omega \hat{q}_{ij}^\omega - T\hat{y}_j(1-z_j^{\ell'}) \qquad \forall j\in\mathcal{J}, \omega\in\Omega, \ell'\le m' \tag{4d}$$

$$o_j^{\omega\ell'} \le v_j\hat{y}_j(1-z_j^{\ell'}) \qquad \forall j\in\mathcal{J}, \omega\in\Omega, \ell'\le m' \tag{4e}$$

$$o_j^{\omega\ell'}(p_\omega b_j - \alpha_j^{\omega\ell'} + \beta_j^{\omega\ell'}) = 0 \qquad \forall j\in\mathcal{J}, \omega\in\Omega, \ell'\le m' \tag{4f}$$

$$\alpha_j^{\omega\ell'}\left(o_j^{\omega\ell'} - \sum_{i\in\mathcal{I}} d_i^\omega \hat{q}_{ij}^\omega + T\hat{y}_j(1-z_j^{\ell'})\right) = 0 \qquad \forall j\in\mathcal{J}, \omega\in\Omega, \ell'\le m' \tag{4g}$$

$$\beta_j^{\omega\ell'}(v_j\hat{y}_j(1-z_j^{\ell'}) - o_j^{\omega\ell'}) = 0 \qquad \forall j\in\mathcal{J}, \omega\in\Omega, \ell'\le m' \tag{4h}$$

$$\sum_{j\in\mathcal{J}} z_j^{\ell'} \le \Gamma \tag{4i}$$

$$o_j^\omega, \alpha_j^\omega, \beta_j^\omega, \theta \ge 0; z_j \in \{0,1\} \qquad \forall i \in \mathcal{I}, \omega \in \Omega, j \in \mathcal{J}. \tag{4j}$$

Since constraints (4f)-(4h) have bilinear terms, we use the big-M method to linearize them. We introduce the binary variables $\text{bin}_{j\omega}^1$, $\text{bin}_{j\omega}^2$, and $\text{bin}_{j\omega}^3$ with respect to constraints (4f)-(4h), and obtain the upper bounds of following set of the constraints:

$$\begin{aligned}
&\alpha_j^\omega \le M_\alpha && \forall j \in \mathcal{J}, \omega \in \Omega\\
&\beta_j^\omega \le M_\beta && \forall j \in \mathcal{J}, \omega \in \Omega\\
&p_\omega b_j - \alpha_j^\omega + \beta_j^\omega \le p_\omega b_j + M_\beta && \forall j \in \mathcal{J}, \omega \in \Omega\\
&o_j^\omega - \sum_{i\in\mathcal{I}} d_i^\omega \hat{q}_{ij}^\omega + T\hat{y}_j(1-z_j) \le v_j^\omega + T && \forall j \in \mathcal{J}, \omega \in \Omega\\
&v_j \hat{y}_j (1-z_j) - o_j^\omega \le v_j && \forall j \in \mathcal{J}, \omega \in \Omega.
\end{aligned}$$

We finally obtain the following MILP-based formulation of the MP_S at the ℓ'-th iteration.

$$\begin{aligned}
[\textbf{MP}_\text{S}]: \ \max \ & \rho\,\theta && \tag{5a}\\
\text{s.t.} \ & \theta \le \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} p_\omega c_{ij} \hat{q}_{ij}^\omega + \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}} p_\omega r_i \hat{u}_i^\omega \nonumber\\
& \quad + \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}} p_\omega b_j o_j^{\omega\ell'} && \forall \ell' \le m' \tag{5b}\\
& \alpha_j^{\omega\ell'} - \beta_j^{\omega\ell'} \le p_\omega b_j && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5c}\\
& o_j^{\omega\ell'} \ge \sum_{i\in\mathcal{I}} d_i^\omega \hat{q}_{ij}^\omega - T\hat{y}_j(1-z_j^{\ell'}) && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5d}\\
& o_j^{\omega\ell'} \le v_j \hat{y}_j (1-z_j^{\ell'}) && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5e}\\
& o_j^{\omega\ell'} \le v_j(1-\text{bin}_{j\omega\ell'}^1) && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5f}\\
& p_\omega b_j - \alpha_j^{\omega\ell'} + \beta_j^{\omega\ell'} \le (b_j + M_\beta)\text{bin}_{j\omega\ell'}^1 && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5g}\\
& \alpha_j^{\omega\ell'} \le M_\alpha(1-\text{bin}_{j\omega\ell'}^2) && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5h}\\
& o_j^{\omega\ell'} - \sum_{i\in\mathcal{I}} d_i^\omega \hat{q}_{ij}^\omega + T\hat{y}_j(1-z_j^{\ell'}) \le (v_j + T)\text{bin}_{j\omega\ell'}^2 && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5i}\\
& \beta_j^{\omega\ell'} \le M_\beta(1-\text{bin}_{j\omega\ell'}^3) && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5j}\\
& v_j \hat{y}_j (1-z_j^{\ell'}) - o_j^{\omega\ell'} \le v_j \text{bin}_{j\omega\ell'}^3 && \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{5k}\\
& \sum_{j\in\mathcal{J}} z_j^{\ell'} \le \Gamma && \tag{5l}\\
& o_j^\omega, \alpha_j^\omega, \beta_j^\omega, \theta \ge 0; z_j \in \{0,1\} && \forall j \in \mathcal{J}, \omega \in \Omega \tag{5m}\\
& \text{bin}_{j\omega}^1, \text{bin}_{j\omega}^2, \text{bin}_{j\omega}^3 \in \{0,1\} && \forall j \in \mathcal{J}, \omega \in \Omega \tag{5n}
\end{aligned}$$

Given the worst-case scenarios that are derived from MP_S, the SP_S can be formulated as follows:

$$\begin{aligned}
[\textbf{SP}_\text{S}]: \ \underset{\boldsymbol{q},\boldsymbol{u},\boldsymbol{o}}{\text{minimize}} \ & \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} p_\omega c_{ij} q_{ij}^\omega + \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}} p_\omega r_i u_i^\omega + \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}} p_\omega b_j o_j^\omega && \tag{6a}\\
\text{subject to} \ & \sum_{j\in\mathcal{J}} q_{ij}^\omega + u_i^\omega = \delta_i^\omega && \forall i \in \mathcal{I}, \omega \in \Omega \tag{6b}\\
& \sum_{i\in\mathcal{I}} d_i^\omega q_{ij}^\omega \le T\hat{y}_j(1-\hat{z}_j) + o_j^\omega && \forall j \in \mathcal{J}, \omega \in \Omega \tag{6c}
\end{aligned}$$

$$o_j^{\omega} \le v_j \hat{y}_j (1 - \hat{z}_j) \qquad \forall j \in \mathcal{J}, \omega \in \Omega \tag{6d}$$

$$o_j^{\omega} \ge 0; q_{ij}^{\omega}, u_i^{\omega} \in \{0,1\} \qquad \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega. \tag{6e}$$

Algorithm 2 The Inner Level C&CG Implementation

```
**Initialize** $m' = 0$, $\mathrm{LB_S} = -\infty, \mathrm{UB_S} = +\infty$ for all $\omega$..
**while** ( $|\frac{\mathrm{UB_S} - \mathrm{LB_S}}{\mathrm{UB_S}}| > \epsilon$) **do**
    Solve the $\mathrm{MP_S}$ (5)
    Record optimal solution $\boldsymbol{z}^{m'}$, and optimal objective $Uobj^{m'}$.
    Update $\mathrm{UB_S} := Uobj^{m'}$.
    Fix $\hat{\boldsymbol{z}} := \boldsymbol{z}^{m'}$, and solve $\mathrm{SP_S}$ (6).
    Obtain optimal solution $\boldsymbol{q}^{m'}, \boldsymbol{u}^{m'}, \boldsymbol{o}^{m'}$.
    Update $\mathrm{LB_S} := \max\left\{\mathrm{LB_S}, \min_{\boldsymbol{q},\boldsymbol{u},\boldsymbol{o}} \sum_{\omega\in\Omega} p_{\omega}\left(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} c_{ij} q_{ij}^{\omega} + \sum_{j\in\mathcal{J}} b_j o_j^{\omega} + \sum_{i\in\mathcal{I}} r_i u_i^{\omega}\right)\right\}$.
    Create integer variables $\boldsymbol{q}^{m'+1}$, $\boldsymbol{u}^{m'+1}$ and add constraints (5b)-(5k) to the $\mathrm{MP_S}$.
    Set $m' := m' + 1$.
**end while**
**return** $\boldsymbol{z^m} := \hat{\boldsymbol{z}}$
```

4.2.2 Strong Duality based Inner Level Method

In this section, we employ strong duality approach to derive the $\mathrm{MP_S}$. Let $\alpha_j^{\omega}, \beta_j^{\omega}$ be the dual variables in terms of constraints (3c) and (3d), respectively. Then, for the ℓ'-th iteration, we still rewrite the objective function (3a) as its epigraph form (i.e., constraint (7b)), then we apply the strong duality for the innermost minimization over $\boldsymbol{o}$, which gives rise to the following maximization-based MILP problem (7). We remark that the $\mathrm{SP_S}$ is the same as problem (6).

$$\text{maximize} \;\; \rho\,\theta \tag{7a}$$

$$\begin{aligned} \text{subject to} \;\; \theta \le & \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} p_{\omega} c_{ij} \hat{q}_{ij}^{\omega} + \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}} p_{\omega} r_i \hat{u}_i^{\omega} \\ & + \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}} d_i^{\omega} q_{ij}^{\omega} \alpha_j^{\omega} - \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}} T\hat{y}_j(1 - z_j^{\ell'}) \\ & - \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}} v_j \hat{y}_j (1 - z_j^{\ell'}) \beta_j^{\omega} \end{aligned} \tag{7b}$$

$$\alpha_j^{\omega\ell'} - \beta_j^{\omega\ell'} \le p_{\omega} b_j \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \le m' \tag{7c}$$

$$\sum_{j\in\mathcal{J}} z_j^{\ell'} \le \Gamma \tag{7d}$$

$$\alpha_j^{\omega}, \beta_j^{\omega} \ge 0; z_j \in \{0,1\} \qquad \forall j \in \mathcal{J}, \omega \in \Omega \tag{7e}$$

Similarly to problem (4), we still use the big-M approach to linearize the bilinear terms in constraint (7b) by introducing binary variables $\mathrm{bin}_{j\omega}^4$ and $\mathrm{bin}_{j\omega}^5$, which finally gives rise to the following MILP-based master problem (8).

$$[\mathbf{MP}_{\mathrm{S}}]: \; \max \;\; \rho\,\theta \tag{8a}$$

$$\begin{aligned} \text{s.t.} \;\; \theta \le & \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} p_{\omega} c_{ij} \hat{q}_{ij}^{\omega} + p_{\omega} r_i \hat{u}_i^{\omega} \\ & + \sum_{\omega\in\Omega}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} d_i^{\omega} \hat{q}_{ij}^{\omega} \alpha_j^{\omega\ell'} - \\ & \sum_{\omega\in\Omega}\sum_{j\in\mathcal{J}} \Big(T\hat{y}_j(\alpha_j^{\omega\ell'} - \mathrm{bin}_{j\omega\ell'}^4) + v_j \hat{y}_j (\beta_j^{\omega\ell'} - \mathrm{bin}_{j\omega\ell'}^5) \Big) \qquad \forall \ell' \le m' \end{aligned} \tag{8b}$$

$$\alpha_j^{\omega\ell'} - \beta_j^{\omega\ell'} \leq p_\omega b_j \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \leq m' \tag{8c}$$

$$\sum_{j \in \mathcal{J}} z_j^{\ell'} \leq \Gamma \tag{8d}$$

$$\text{bin}^4_{j\omega\ell'} \leq \alpha_j^{\omega\ell'} \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \leq m' \tag{8e}$$

$$\text{bin}^4_{j\omega\ell'} \leq M_\alpha z_j^{\ell'} \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \leq m' \tag{8f}$$

$$\text{bin}^4_{j\omega\ell'} \geq \alpha_j^{\omega\ell'} - M_\alpha(1 - z_j^{\ell'}) \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \leq m' \tag{8g}$$

$$\text{bin}^5_{j\omega\ell'} \leq \beta_j^{\omega\ell'} \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \leq m' \tag{8h}$$

$$\text{bin}^5_{j\omega\ell'} \leq M_\beta z_j^{\ell'} \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \leq m' \tag{8i}$$

$$\text{bin}^5_{j\omega\ell'} \geq \beta_j^{\omega\ell'} - M_\beta(1 - z_j^{\ell'}) \qquad \forall j \in \mathcal{J}, \omega \in \Omega, \ell' \leq m' \tag{8j}$$

$$\alpha_j^\omega, \beta_j^\omega \geq 0; \text{bin}^4_{j\omega}, \text{bin}^5_{j\omega} \in \{0,1\}; z_j \in \{0,1\} \qquad \forall j \in \mathcal{J}, \omega \in \Omega \tag{8k}$$

4.3 Symmetry-breaking Constraints

Without loss of generality, we assume that the cost parameters are homogeneous. This is a common assumption in the healthcare operations management literature. Given this assumption, we add the following set of symmetry-breaking constraints (Denton et al., 2010) to the MP (2).

$$y_j \geq y_{j+1} \qquad \forall j \in \mathcal{J} \tag{9}$$

Since we assume that the cost of the assignment, telemedical doctors' working, and overtime are homogeneous, which means complete symmetry for the telemedical doctors. Therefore, for any solution, an equivalent solution can be derived by exchanging the sets of patients assigned to any pair of telemedical doctors. To break the symmetry and limit the number of solutions, we add the constraints $y_1 \geq y_2 \geq \ldots \geq y_{|\mathcal{J}|}$. This could in turn help us to quickly identify a good solution of MP and further generate a better upper bound that is given by the inner level problem.

5 Numerical Study

In this section, we conduct an extensive study to verify the performance of our model and the proposed solution algorithms. We first describe the implementation details in Section 5.1, then report the computational performance in Section 5.2 and the robustness and sensitivity analysis in Section 5.3. Section 5.4 presents how the no-show behaviours affect the performance. Finally, Section 5.5 compares our model with its two-stage SP counterpart.

5.1 Implementation Details

We consider a set of randomly generated instances of different sizes. For the sake of brevity, we use $|\mathcal{I}|$-$|\mathcal{J}|$-K to represent each class of problem size. We consider the problem size with 30-10-3 in Section 5.2.1 and with 100-10-10 for the rest. Note that each class of problems has 5 randomly generated instances, which are generated as follows. We consider Ω empirical scenarios to capture the uncertain service duration and the patient no-show, and we use $\Omega \in \{30, 50, 80\}$ in Section 5.2.1 and $\Omega \in \{50, 100, 200\}$ in Section 5.2.2. For each scenario ω, the service duration d_i^ω of patient i is randomly and independently generated from the uniform distribution [5,10], with probability $p_\omega = 1/|\Omega|$. Regarding the no-show behaviours of the patients, we consider the probability of patient no-show p_{no}, which means the binary vector δ_i^ω of the patient i who does not show up under scenario ω follows the binomial distribution with a mean of $1 - \text{p}_{\text{no}}$. The budget of uncertainty Γ takes a set of integer values $\Gamma \in \{0, 1, \ldots, K\}$. Finally, we set the the assignment cost c_{ij} to 2, the overtime cost b_j to 0.3, the working cost h_j to 20, and the penalty cost r_i to 60. We consider the fixed-length of the block $T = 100$ minutes, maximum length of the overtime $v_j = 50$ minutes and the weight $\rho \in \{0.2, 0.5, 0.8\}$.

All the algorithms are implemented in MATLAB 2019a and YALMIP package (`https://yalmip.github.io/`) using CPLEX 12.71 as the solver with the default setting on a Desktop machine with Intel(R) Xeon(R)

3.30 GHz processor and 64 GB RAM in a Windows 64-bit system. The algorithms are run until either an optimality gap below 1% or a time limit of 7200 seconds is reached. All the reported performance measurements are averaged over five instances.

5.2 Computational Performance

In this section, we present the computational results of the proposed algorithms in terms of different problem sizes (e.g., $|\mathcal{I}|$-$|\mathcal{J}|$-K, Ω) and different parameters (e.g., Γ, p_{no}, ρ). More specifically, Section 5.2.1 shows the algorithmic performance comparisons of three variants of our algorithm for small-size problems, and Section 5.2.2 further presents the numerical efficiency of our improved nested C&CG algorithm for large-size instances. For each class of problem sizes, we report the average CPU solution time (Time, in seconds), the proportion of unsolved instances (prop) over five instances, and the average number of iterations (# of Iter), respectively.

5.2.1 Computational Performance of Different Algorithms

In this section, we consider the problem size with $\Omega \in \{30, 50, 80\}$ and 30-10-3. For other parameters, ρ is assumed to 0.5 and p_{no} is assumed to 0.1. We evaluate the computational performance using the following three variants: **KKT** refers to the nested C&CG with KKT conditions in Section 4.2.1; **Dual** refers to the nested C&CG with strong duality in Section 4.2.2 and symmetry-breaking constraints (9) in Section 4.3; **Improve** refers to the nested C&CG with KKT conditions in Section 4.2.1 and symmetry-breaking constraints (9) in Section 4.3.

Table 3 reports the numerical efficiency of the above three variants. From the table we can observe that, as expected, for all three algorithms the computational difficulty increases as the number of scenarios Ω is increasing from 30 to 80. Given Ω and Γ, we can see that, the improved method outperforms the best among all three methods and can solve all the instances optimally within 2 minutes on average, while the KKT method is the most inefficient one and needs the largest number of iterations on average (e.g., the outer level with 7.2 iterations on average), which can only solve a half of all instances (i.e. 30 instances over 60 in total) optimally within 2 hours. Moreover, the dual method could solve 97% of all instances optimally within about an hour on average. Here we can conclude that the symmetry-breaking constraints (9) plays a very important role in improving the numerical efficiency of our proposed nested C&CG algorithm, which however greatly reduces the number of integer variables in the outer master problem and makes our algorithm quickly find a good feasible solution of the assignment for telemedical doctors to speed up the convergence. In addition, the dual method seems to have fewer constraints, which results in a better computational efficiency than the KKT method in terms of the average solution time (e.g., 2508.5 vs 5753.2). Although the dual method converges slowly than the KKT method for each inner level iteration, the KKT method needs more iterations in the outer level problem. We remark that we also tried the dual method without the symmetry-breaking constraints (9). However, the numerical efficiency performs very bad, and thus we omit it in the table. For any given Ω and each method, we also learn from the table that the average solution time increases as Γ increases from 0 to 3. This could be explained by the fact that the proposed model becomes more conservative and robust when Γ becomes larger, which in some sense makes the computations more expensive. In summary, the improved method performs the best, which motivates us to further test its peformance for solving large-sized problem in the next section.

5.2.2 Computational Performance of Large-scaled Problems using Improved C&CG Algorithm

In order to further explore the algorithmic efficiency for the problems of more realistic sizes, this section reports the computational performance for the problem with 100-10-10 over a large number of scenarios (i.e., $\Omega \in \{50, 100, 200\}$), using the improved C&CG method.

From Table 4, we can learn that we can solve 97% of the instances optimally within 4516.6 seconds on average (over 360 instances in total). As we increase the number of the patients and telemedical doctors, the average solution time significantly increases, because the size of the model becomes bigger. Moreover, the "budgeted" robustness Γ and the number of the empirical scenarios Ω have a big impact on the computational performance.

Table 3: Computational performance for three variants of our algorithm for 30 patients, 10 telemedical doctors over $\Omega \in \{30, 50, 80\}$ and $\Gamma \in \{0, 1, 2, 3\}$, where we report the average solution time (Time, in seconds), proportion of unsolved instances within 2 hours (prop), and average number of iterations for inner level (inner) and outer level (outer) C&CG.

Ω	Γ	Dual				KKT				Improve			
		Time	prop	inner	outer	Time	prop	inner	outer	Time	prop	inner	outer
30	0	89.0	0.0	1.0	1.0	10.1	0.0	1.0	1.0	8.3	0.0	1.0	1.0
	1	3398.2	0.4	3.8	5.4	706.1	0.0	2.0	9.0	26.3	0.0	2.0	2.0
	2	1437.9	0.0	5.8	2.0	11657.7*	1.0	1.8	12.4	29.4	0.0	2.0	2.0
	3	580.7	0.0	2.0	2.0	10589.4*	1.0	1.8	11.6	27.7	0.0	2.0	2.0
50	0	327.4	0.0	1.0	1.0	53.1	0.0	1.0	1.0	13.5	0.0	1.0	1.0
	1	2189.8	0.0	4.0	2.0	2484.2	0.0	2.0	9.0	48.5	0.0	2.0	2.0
	2	3760.1	0.0	6.0	2.0	9659.3*	1.0	1.8	9.6	52.9	0.0	2.0	2.0
	3	1467.8	0.0	2.0	2.0	9923.5*	1.0	1.7	7.6	50.6	0.0	2.0	2.0
80	0	798.8	0.0	1.0	1.0	98.7	0.0	1.0	1.0	24.0	0.0	1.0	1.0
	1	5148.3	0.0	4.0	2.0	5738.0	0.0	2.0	9.0	86.2	0.0	2.0	2.0
	2	7429.2*	0.0	3.0	2.0	8748.0	1.0	1.7	7.8	92.1	0.0	2.0	2.0
	3	3475.3	0.0	2.0	2.0	9369.9*	1.0	1.7	7.6	88.9	0.0	2.0	2.0
Avg		2508.5	0.03	3.0	2.0	5753.2	0.50	1.6	7.2	45.7	0.0	1.8	1.8

"*" in the column of "Time" means that at least one of the five instances can not be solved optimally within the time limit.

As is shown in the previous section, it is clear that once again increasing robustness (i.e., Γ), the computational cost becomes higher for any given parameters setting of Ω, p_{no}, and ρ. Moreover, as expected the average solution time greatly increases when we increase the number of empirical scenarios from 50 to 200. If we have a closer analysis on the nested C&CG method, we note that MP_S and SP_S are easier to solve, and the bottleneck is to solve MP, which initially is a large integer program with a large number of binary variables and constraints over each iteration. Fortunately, as shown in Table 4, it only takes a small number of iterations to reach the convergence. Finally, the numerical results show that p_{no} and ρ have a slight impact on the average solution time.

5.3 Robustness and Sensitivity Analysis

In this section, we explore how the three different sources of uncertainty and the parameters affect the balances between the robustness and cost budget, i.e., the weight ρ that balances the expected cost of the first-stage and second-stage problem, the "budgeted" robustness parameter Γ, and the probability of patient no-show p_{no}. As described before, our analysis in this section is based on 100 patients and 10 telemedical doctors with 100 empirical scenarios. More specifically, Section 5.3.1 presents how Γ affects the cost configuration while Section 5.3.2 performs the sensitivity analysis on the weight ρ. Finally, Section 5.3.3 shows a trade-off analysis between the expected total cost and assignment policy.

5.3.1 Robustness Analysis of Cost Configuration

In this section, we further explore how the robustness parameter Γ affects the cost configuration (i.e., first-stage cost, second-stage cost, and total cost) for $\rho = \{0.2, 0.5, 0.8\}$. We set the probability of patient no-show to 0.1. We presents the average expected total cost (TC), first-stage cost (FSC), second-stage cost (SSC) with respect to $\Gamma \in [0, 10]$ for $\rho = 0.2$ in Figure 3(a), $\rho = 0.5$ in Figure 3(b), and $\rho = 0.8$ in Figure 3(c), respectively, while Figure 3(d) presents the TC as a function of Γ for $\rho \in \{0.2, 0.5, 0.8\}$. As we can see from the figure when ρ is relatively small (e.g., $\rho = 0.2$), FSC seems to be nearly unchanged with a relatively small expected cost (except when $\Gamma = 0$), while TC and SSC are increasing as Γ increases, especially $\Gamma \in [4, 10]$. In addition, for any given $\Gamma \in [4, 10]$, the average expected total cost increases with the increase of ρ. More interestingly, there is a very slight difference between SSC and TC. This is because, for the given Γ, FSC is very small when ρ is relatively small. When ρ is large (e.g., $\rho \in \{0.5, 0.8\}$), we can see the very similar trends.

Table 4: Computational performance of the improved C&CG for 100 patients, 10 telemedical doctors over $\Omega \in \{50, 100, 200\}$, $\Gamma \in \{0, 2, 8, 10\}$, $\rho \in \{0.2, 0.5, 0.8\}$, and $\mathrm{p}_{\mathrm{no}} \in \{0.1, 0.3\}$, where we report the average solution time (Time, in seconds), proportion of unsolved instances within 2 hours (prop), and average number of iterations (# of Iter).

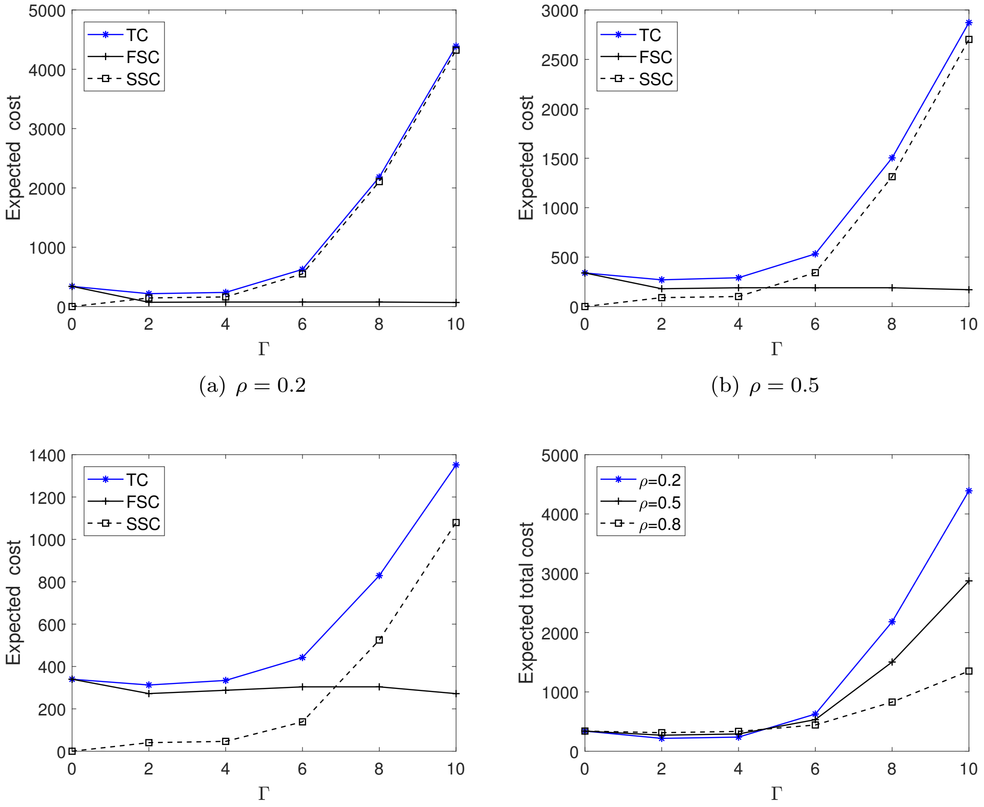
Ω	ρ	p_{no}	Gamma=2			Gamma=4			Gamma=8			Gamma=10		
			time	prop	# of iter	time	prop	# of iter	time	prop	# of iter	time	prop	# of iter
50	0.2	0.1	401.6	0	2.0	278.6	0	2.0	260.3	0	2.0	428.0	0	3.0
		0.3	339.9	0	2.0	308.3	0	2.0	456.9	0	2.6	618.1	0	3.0
	0.5	0.1	425.0	0	2.0	308.0	0	2.0	298.0	0	2.0	480.3	0	3.0
		0.3	448.9	0	2.0	498.2	0	2.0	426.8	0	2.0	650.5	0	3.0
	0.8	0.1	380.0	0	2.0	325.0	0	2.0	310.7	0	2.0	498.2	0	3.0
		0.3	426.9	0	2.0	513.6	0	2.0	665.7	0	2.8	705.8	0	3.0
Avg			403.7	0.0	2.0	371.9	0.0	2.0	403.1	0.0	2.2	563.5	0.0	3.0
100	0.2	0.1	1713.1	0	2.0	1499.1	0	2.0	1130.9	0	2.0	1754.8	0	3.0
		0.3	1119.5	0	2.0	1190.6	0	2.0	945.0	0	2.0	1658.3	0	3.0
	0.5	0.1	1628.5	0	2.0	1406.0	0	2.0	1118.0	0	2.0	1713.7	0	3.0
		0.3	1201.8	0	2.0	1438.9	0	2.0	1043.0	0	2.0	1696.9	0	3.0
	0.8	0.1	1309.1	0	2.0	1456.1	0	2.0	1124.7	0	2.0	1711.3	0	3.0
		0.3	1158.9	0	2.0	2390.5	0	2.0	1268.7	0	2.0	1867.9	0	3.0
Avg			1355.2	0.0	2.0	1563.5	0.0	2.0	1105.1	0.0	2.0	1733.8	0.0	3.0
200	0.2	0.1	5155.7	0	2.0	3651.4	0	2.0	2929.4	0	2.0	4825.5	0	3.0
		0.3	3284.6	0	2.0	3260.0	0	2.0	2369.1	0	2.0	4976.5	0	3.0
	0.5	0.1	3845.6	0	2.0	4432.3	0	2.0	3270.1	0	2.0	5026.4	0	3.0
		0.3	2710.7	0	2.0	2317.5	0	2.0	2747.4	0	2.0	4127.2	0	3.0
	0.8	0.1	5338.0	0	2.0	5022.5	0	2.0	3203.9	0	2.0	5155.9	0	3.0
		0.3	3643.2	0	2.0	3941.4	0	2.0	4516.6	0.2	2.4	4232.1	0	3.0
Avg			3996.2	0	2.0	3770.9	0.0	2.0	3172.8	0.03	2.1	4723.9	0.0	3.0

However, the distance between TC and SSC seems to be bigger than that with $\rho = 0.2$, because FSC becomes larger when more weight is imposed on the normal case of telemedicine assignment. We also observe from Figure 3(d) that TC under different ρ crosses and the ranking changes after $\Gamma \geq 4$. This could be explained by the fact that when Γ is small (e.g., $\Gamma \leq 4$), nearly all the patients could be served by the telemedical doctors who do show up, which results in a relatively low TC when ρ is small. However, if we continue to increase Γ, TC becomes higher. This is caused by the fact that more patients might not be served, which in turn results in a higher SSC (i.e., including the overtime cost and penalty cost for the unserved patients).

5.3.2 Sensitivity Analysis on Weight ρ

In order to further explore the cost configuration, in this section, we conduct a sensitivity analysis on weight parameter ρ. We present the TC, FSC, and SSC as a function of $\rho \in [0, 1]$ for different $\Gamma \in \{1, 5, 7\}$, which are shown in Figures 4 (a), (b), and (c) respectively, while Figure 4(d) presents TC as a function of ρ for different $\Gamma \in \{1, 5, 7\}$. We do not consider the patient no-show (i.e., $\mathrm{p}_{\mathrm{no}} = 0$), in order to clearly see how ρ affects the cost configuration.

From Figure 4(a) and Figure 4(b), we can clearly see that TC is increasing as ρ increases, and TC is a simple linear combination of FSC and SSC, when Γ is small (i.e., $\Gamma = 1$ and $\Gamma = 5$). More interesting, if we further increase Γ (e.g., $\Gamma = 7$ in Figure 4(c)), it seems to be that TC is linearly decreasing as ρ increases. This is because SSC decreases very fast when Γ is large, which makes TC be decreasing. Besides, one could further confirm these observations in Figure 4(d). Finally, we conclude that the weight ρ plays a key role in the trade-off between FSC and SSC, especially when facing the uncertainty with the no-show behaviour of the telemedical doctors. Our analyses shed light on the fact that the decision-maker is less conservative with a large ρ and unwilling to configure the system in a way such that less recourse operation costs be incurred under a situation with telemedical doctor no-show.



(a) $\rho = 0.2$

(b) $\rho = 0.5$

(c) $\rho = 0.8$

(d) TC with $\rho \in \{0.2, 0.5, 0.8\}$

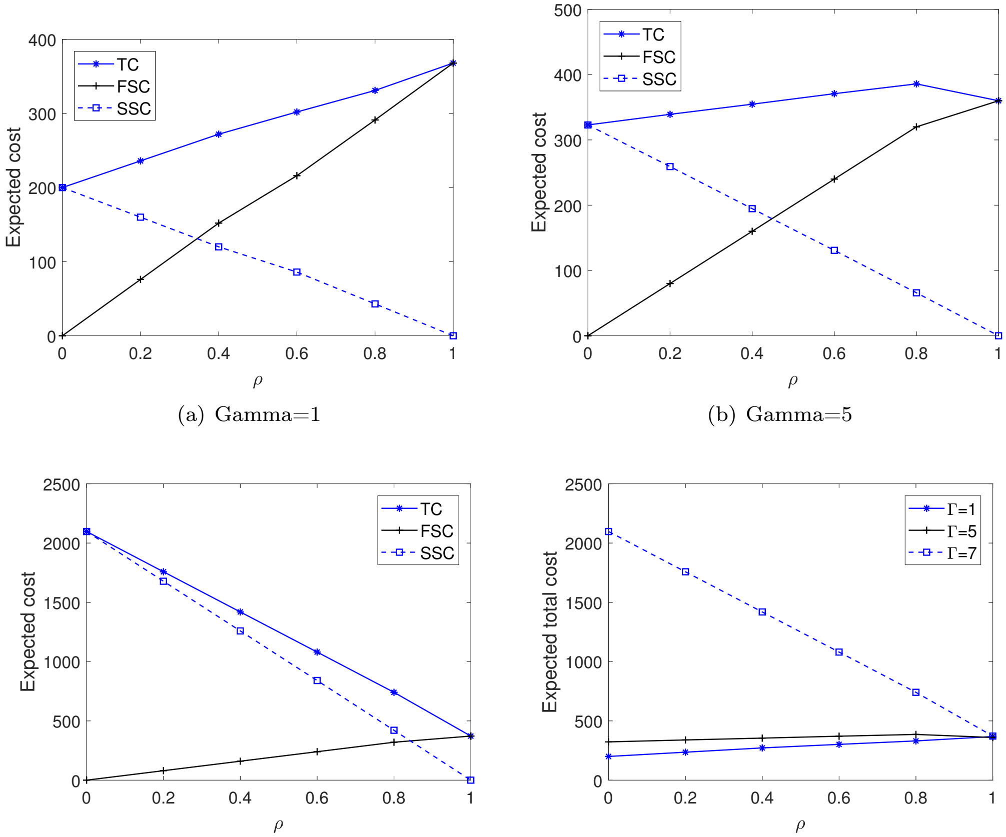
Figure 3: The average expected cost of first-stage (FSC), second-stage (SSC), and total cost (TC) for 100 patients and 10 telemedical doctors as a function of Γ over $\Omega = 100$, $\rho = \{0.2, 0.8\}$, and $\mathrm{p_{no}} = 0.1$.

5.3.3 Trade-Off Analysis between Expected Total Cost and Assignment Policy

In this section, we conduct a trade-off analysis between the expected total cost and the number of unassigned patients under $\Omega = 100$, $\rho = 0.8$, $\mathrm{p_{no}} \in \{0.1, 0.3, 0.5\}$ and $\Gamma \in \{0, 2, 4, 6, 8, 10\}$, which is specifically presented in Figure 5. The tickmarks under different $\mathrm{p_{no}}$ represent different Γ levels. From Figure 5, we can observe that the trade-off frontier between the expected total cost and the number of unassigned patients under $\mathrm{p_{no}} = 0.5$ dominates these with $\mathrm{p_{no}} = 0.3$ and $\mathrm{p_{no}} = 0.1$ in order. This is due to the fact that the probability of overtime being lower when $\mathrm{p_{no}}$ is relatively large. Thus, given a very limited number of telemedical doctors who show up, they could serve as many patients as possible while achieving a lower expected total cost. This can be further confirmed in Section 5.4. As a complement, Table 5 further reports the expected total cost and the number of unassigned patients under different $\mathrm{p_{no}}$ and Γ. We can oberve that, for a given $\mathrm{p_{no}}$, when Γ increases, the number of unassigned patients increases and the expected total cost is increasing (except when $\Gamma = 0$, see Remark 1), which behaves the same as Figure 3(c). On the other hand, for a given Γ, the same observations are derived as those in Figure 5.

5.4 Analysis on the Impact of the No-Show Behaviours

In this section, we aim to explore that how the no-show behaviours of patients (i.e., $\mathrm{p_{no}}$) and telemedical doctors (i.e., Γ) affect the average expected total cost. We set the weight ρ to 0.5. Figure 6(a) presents the average expected total cost as a function of Γ for $\mathrm{p_{no}} \in \{0, 0.1, 0.3\}$, while Figure 6(b) presents the average expected total cost as a function of $\mathrm{p_{no}}$ for $\Gamma \in \{1, 3, 5\}$. As we can clearly see from Figure 6(a) that, for a fixed $\mathrm{p_{no}}$ (e.g., $\mathrm{p_{no}} = 0$ means that only the telemedical doctor no-show is considered), the average



(a) Gamma=1

(b) Gamma=5

(c) Gamma=7

(d) TC with $\Gamma \in \{1, 5, 7\}$

Figure 4: The average expected cost as a function of ρ over $\Omega = 100$, $\Gamma \in \{1, 5, 7\}$, and $\mathrm{p_{no}} = 0$.

expected total cost is non-decreasing as Γ increases from 2 to 10. In particular, the average expected total cost drastically increases when $\Gamma \geq 4$. This could be based on the fact that the workload of telemedical doctors is extremely heavy and more and more patients might not be served at the price of the shortage of available telemedical doctors, when at least 4 telemedical doctors might fail to show up due to the emergency requests. In terms of the patient no-show, we can learn from Figure 6(b) that, for any given Γ (e.g., $\Gamma = 0$ means that only the patient no-show is considered), the average expected total cost decreases as $\mathrm{p_{no}}$ becomes larger. This is because it is easy to serve all the patients who show up by the available telemedical doctors, without the overtime and the penalty for the unserved patients. Moreover, we observe that the no-show of the telemedical doctors has a bigger impact on the expected total cost than the patient no-show, given that the former generally leads to a larger expected total cost, which is clearly shown in Figure 6. This might imply that the optimal choice of the hyper-parameters in this study consists in $\Gamma \in (0, 4]$ and a relatively small $\mathrm{p_{no}}$, which says that, under current setting, we might allow a limited number of telemedical doctors to be no-shows (due to emergency events) and can serve as many patients as possible while paying a relatively cheaper expected total cost. This is in line with the observations derived in Section 5.3.3. Finally, we derive that the total expected cost decreases as patient no-show rates increases, and the total expected cost will significantly increase when the number of telemedicine doctors failing to show up exceeds a certain proportion (e.g., 60% in our numerical study).

5.5 A Comparison of Robust and Stochastic Solutions

To further show the quality of our robust solutions, in this section we compare our two-stage robust optimization (2RO) model with a benchmark model (i.e., two-stage stochastic programming (2SP) model, which is

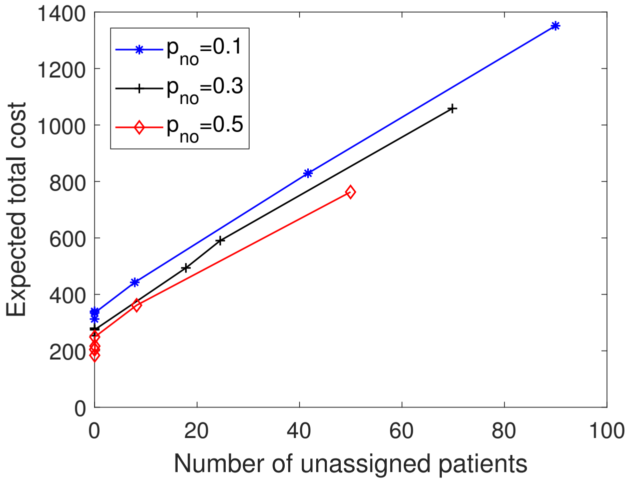


Figure 5: The trade-off between average expected total cost and the number of unassigned patients for 100 patients and 10 telemedical doctors under $\Omega = 100$, $\rho = 0.8$, and $\mathrm{p_{no}} \in \{0.1, 0.3, 0.5\}$.

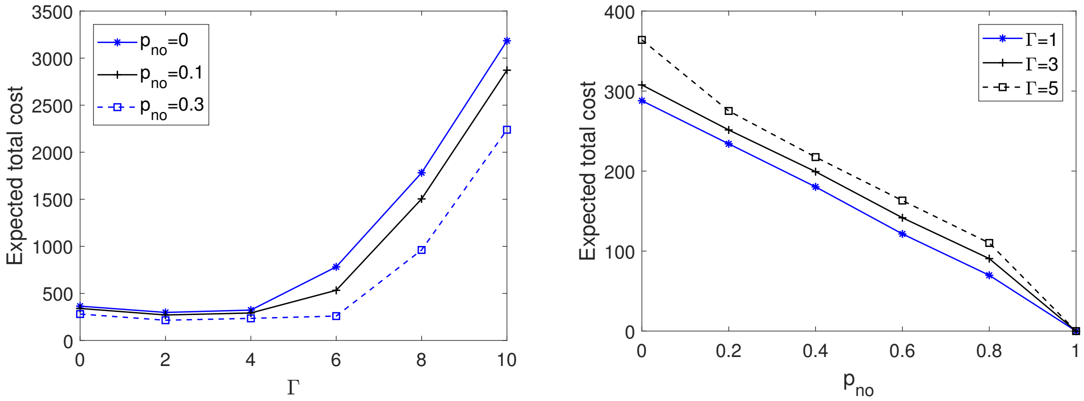
Table 5: The expected total cost and the number of unassigned patients for 100 patients and 10 telemedical doctors under $\Omega = 100$, $\rho = 0.8$, $\mathrm{p_{no}} \in \{0.1, 0.3, 0.5\}$ and $\Gamma \in \{0, 2, 4, 6, 8, 10\}$.

	Γ	0	2	4	6	8	10
$\mathrm{p_{no}}$ =0.1	Expected total cost	339.9	312.7	334.3	442.5	829.0	1351.6
	Number of unassigned patients	0.0	0.0	0.0	7.8	41.7	90.0
$\mathrm{p_{no}}$=0.3	Expected total cost	280.1	253.8	274.6	493.7	590.4	1058.5
	Number of unassigned patients	0.0	0.0	0.0	17.8	24.5	69.8
$\mathrm{p_{no}}$ =0.5	Expected total cost	204.8	184.1	216.8	249.0	361.4	762.6
	Number of unassigned patients	0.0	0.0	0.1	0.0	8.2	50.0

presented in A). For both models, we consider 100 patients and 10 telemedical doctors with $\rho = 0.5$, $\Omega = 50$ and $\mathrm{p_{no}} \in \{0.1, 0.3\}$. For the 2SP model, the key point is how to capture the no-show of the telemedical doctors, in order to make the comparison to be fair. Specifically, we assume that the probability that the at most Γ $(1 \leq \Gamma \leq 9)$ telemedical doctors were no-shows is $1 - \rho$, i.e., 0.5 in this study, which roughly matches the weight of normal situation in the 2RO model. We also assume that each telemedical doctor has the same no-show probability (denoted by γ) and their no-shows are independent of each other. We use the random variable X to denote the event that there are Γ telemedical doctors with no-show behaviour. Thus, the random variable X follows a Binomial Distribution, i.e., $X \sim B(|\mathcal{J}|, \gamma)$. For example, if $\Gamma = 2$, one can easily calculate the no-show probability of a telemedical doctor is 0.26. In doing so, the no-show probability of the telemedical doctors varies with Γ accordingly.

We propose two measurements for the total cost (i.e., the expected total cost (ETC) and the worst-case total cost (WTC), respectively) to evaluate the performance of the robust solutions and stochastic solutions. For 2SP model, the WTC is calculated by fixing its optimal assignments of telemedical doctors (i.e., $\hat{\boldsymbol{y}}_{\mathrm{SP}}$) to 2RO model, while for 2RO model, the ETC is obtained by fixing its optimal assignment of telemedical doctors (i.e., $\hat{\boldsymbol{y}}_{\mathrm{RO}}$) to 2SP model.

Table 6 and Table 7 present the comparisons with 2SP model in terms of the ETC and WTC under $\mathrm{p_{no}} \in \{0.1, 0.3\}$ and $1 \leq \Gamma \leq 9$. In order to evaluate the performance, we define $\Delta_{\mathrm{ETC}} = (\mathrm{ETC_{SP}} - \mathrm{ETC_{RO}})/\mathrm{ETC_{RO}}$ and $\Delta_{\mathrm{WTC}} = (\mathrm{WTC_{SP}} - \mathrm{WTC_{RO}})/\mathrm{WTC_{RO}}$ to show their relative differences. As we can see from these tables, for both models the expected total cost increases as Γ increases from 1 to 9 and $\mathrm{p_{no}}$ decreases from 0.3 to 0.1. This is because the larger the no-show probability of patients and the more telemedical doctors who do not show up, the less unserved patients and the lower penalty cost. For $\mathrm{p_{no}} = 0.3$, we observe from Table 7 that an 2SP solution might have a little bit less ETC (with negative Δ_{ETC}) for most of Γ values,



(a) $\mathrm{p_{no}} \in \{0, 0.1, 0.3\}$

(b) $\Gamma \in \{1, 3, 5\}$

Figure 6: The average expected total cost as a function of $\mathrm{p_{no}}$ and Γ.

while its WTC could be significantly more (with positive Δ_{WTC}), when compared with an 2RO solution. A similar observation could also be derived from Table 6. Furthermore, we see that the difference in ETC for these two models is not drastic, which implies that, 2SP and 2RO models are comparable, even when Γ is relatively large, and that our 2RO model is not overly conservative. In terms of the WTC in both tables, we can observe that an 2SP solution always results in a higher total cost than our 2RO solution for both $\mathrm{p_{no}} = 0.1$ and $\mathrm{p_{no}} = 0.3$, given that all Δ_{WTC} are positive, which could save at most 15.32% of the total cost under $\mathrm{p_{no}} = 0.1$ and 22.6% under $\mathrm{p_{no}} = 0.3$, respectively. This might imply that, instead of relying on accurate probabilistic information for the 2SP model, our 2RO model seems to provide a relatively cheaper modeling alternative that requires much less information support under a worst-case situation.

Table 6: The comparisons with the 2SP model in terms of the WTC and ETC with $\mathrm{p_{no}} = 0.1$.

	Γ	1	2	3	4	5	6	7	8	9
ETC	2SP	268.0	286.7	339.9	428.7	636.9	929.1	1287.2	1764.1	2380.3
	2RO	270.3	303.0	347.7	411.0	611.1	852.3	1194.9	1807.5	2433.4
	Δ_{ETC}	-0.85%	-5.39%	-2.23%	4.29%	4.22%	9.01%	7.72%	-2.40%	-2.18%
WTC	2SP	270.6	280.6	290.2	300.6	316.6	550.8	1004.1	1506.6	2435.5
	2RO	259.8	269.6	280.6	292.0	306.4	536.1	991.9	1506.1	2112.0
	Δ_{WTC}	4.14%	4.07%	3.41%	2.94%	3.31%	2.73%	1.23%	0.03%	**15.32%**

Table 7: The comparisons with the 2SP model in terms of the WTC and ETC with $\mathrm{p_{no}} = 0.3$.

	Γ	1	2	3	4	5	6	7	8	9
ETC	2SP	209.4	218.6	238.3	277.6	386.7	495.6	799.5	1266.2	1778.3
	2RO	155.4	245.8	251.3	308.8	357.9	521.1	870.7	1296.1	1776.3
	Δ_{ETC}	34.80%	-11.08%	-5.14%	-10.10%	8.04%	-4.88%	-8.18%	-2.31%	0.11%
WTC	2SP	240.8	246.4	259.4	268.7	275.6	290.3	516.1	982.8	1859.4
	2RO	206.9	214.4	225.8	236.5	246.0	257.1	492.8	955.5	1516.2
	Δ_{WTC}	16.39%	14.90%	14.88%	13.61%	12.02%	12.91%	4.73%	2.86%	**22.64%**

6 Concluding Remarks

Telemedicine plays an important role in health care delivery during the COVID-19 pandemic to reduce the transmission of the virus while minimizing people's exposure to the public and in-person visits. Therefore, it is crucial for healthcare managers to make an assignment plan for the patients and telemedical doctors when providing telemedicine services, especially in a highly uncertain environment. Motivated by this background, we present the first comprehensive study of a telemedicine assignment problem when three different sources of uncertainty are incorporated. Methodologically speaking, we propose a novel two-stage robust model to address the assignment plan for the normal case that all the telemedical doctors show up and the reassignment plan for the worst case with telemedical doctor no-show, once some of the telemedical doctors fail to show up due to the emergency events. The proposed model is a hard problem since the recourse problem is a max-min-based MILP, which prevents us from using the state-of-art decomposition schemes in the literature. Therefore, we propose an efficient nested C&CG solution scheme that decomposes the whole model into the outer level and inner level problems. The results have confirmed its efficiency for the problems of realistic sizes. We can solve the problems with 100 patients, 10 telemedical doctors, and 200 empirical scenarios within two hours.

On the empirical side, this paper demonstrated in detail how the hyper-parameters affect the balances between cost management and the coverage level under a highly uncertain context. We derive some interesting findings and insights from our analyses, which might make our study more attractive for healthcare practitioners. Specifically, when the patient no-show rate is high, we might allow a limited number of telemedical doctors who show up to serve as many patients as possible while producing a relatively cheaper expected total cost. We also observe that higher patient no-show rates generally result in a decrease in the expected total cost, and the expected total cost will increase significantly when the number of telemedicine doctors who fail to show up exceeds a relatively high proportion (e.g., 60% in our numerical study). Our comparison with the 2SP model implies that our 2RO model is not overly conservative and seems to provide a relatively cheaper modeling alternative that requires much less information support when hedging against three different sources of uncertainty under a worst-case situation.

Finally, we believe that our work still leaves space for some directions regarding: 1) address the case that service duration $\boldsymbol{d}$ is influenced by the assignment decisions, where $\boldsymbol{d}$ is captured by a decision-dependent uncertainty set (e.g., Nohadani and Sharma, 2018); 2) incorporate scheduling and sequencing decisions with downstream resources constraints and multiple performance criteria (e.g., revenue and the number of patients) in the model.

Acknowledgments

The authors would like to thank the associate editor and anonymous reviewers for their useful and insightful comments that have significantly improved the paper. This work was partially supported by Natural Science Foundation of China [Grants 71972012 and 71964034]. The authors thank all these grants support very much.

References

Addis, B., Carello, G., Grosso, A., Lanzarone, E., Mattia, S., Tànfani, E., 2015. Handling uncertainty in health care management using the cardinality-constrained approach: Advantages and remarks. Operations Research for Health Care 4, 1–4.

Addis, B., Carello, G., Tànfani, E., 2014. A robust optimization approach for the operating room planning problem with uncertain surgery duration. In: Proceedings of the International Conference on Health Care Systems Engineering. Springer, pp. 175–189.

Ahmadi-Javid, A., Jalali, Z., Klassen, K. J., 2017. Outpatient appointment systems in healthcare: A review of optimization studies. European Journal of Operational Research 258 (1), 3–34.

An, Y., Zeng, B., Zhang, Y., Zhao, L., 2014. Reliable p-median facility location problem: two-stage robust models and algorithms. Transportation Research Part B: Methodological 64, 54–72.

Bahl, S., Singh, R. P., Javaid, M., Khan, I. H., Vaishya, R., Suman, R., 2020. Telemedicine technologies for confronting covid-19 pandemic: a review. Journal of Industrial Integration and Management 5 (4).

Baker, J., Stanley, A., 2018. Telemedicine technology: a review of services, equipment, and other aspects. Current Allergy and Asthma Reports 18 (11), 1–8.

Bandi, C., Gupta, D., 2020. Operating room staffing and scheduling. Manufacturing & Service Operations Management 22 (5), 958–974.

Batun, S., Denton, B. T., Huschka, T. R., Schaefer, A. J., 2011. Operating room pooling and parallel surgery processing under uncertainty. INFORMS Journal on Computing 23 (2), 220–237.

Ben-Tal, A., El Ghaoui, L., Nemirovski, A., 2009. Robust optimization. Princeton university press.

Ben-Tal, A., Goryashko, A., Guslitzer, E., Nemirovski, A., 2004. Adjustable robust solutions of uncertain linear programs. Mathematical Programming 99 (2), 351–376.

Benders, J., 1962. Partitioning procedures for solving mixed-variables programing problems. Numerische Mathematik 4.

Bertsimas, D., Sim, M., 2004. The price of robustness. Operations Research 52 (1), 35–53.

Birge, J. R., Louveaux, F., 2011. Introduction to stochastic programming. Springer Science & Business Media.

Breuer, D. J., Lahrichi, N., Clark, D. E., Benneyan, J. C., 2020. Robust combined operating room planning and personnel scheduling under uncertainty. Operations Research for Health Care 27, 100276.

Cakici, O., Mills, A., 2020. On the role of teletriage in healthcare demand management. Manufacturing & Service Operations Management. Kelley School of Business Research Paper (16-71).

Cardoen, B., Demeulemeester, E., Beliën, J., 2010. Operating room planning and scheduling: A classification scheme. International Journal of Health Management and Information 1 (1), 71–83.

Dai, T., Tayur, S., 2020. Om forum—healthcare operations management: A snapshot of emerging research. Manufacturing & Service Operations Management 22 (5), 869–887.

Dantas, L. F., Fleck, J. L., Oliveira, F. L. C., Hamacher, S., 2018. No-shows in appointment scheduling–a systematic literature review. Health Policy 122 (4), 412–421.

Delage, E., Ye, Y., 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research 58 (3), 595–612.

Denton, B. T., Miller, A. J., Balasubramanian, H. J., Huschka, T. R., 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. Operations Research 58 (4-part-1), 802–816.

Erdogan, S. A., Krupski, T. L., Lobo, J. M., 2018. Optimization of telemedicine appointments in rural areas. Service Science 10 (3), 261–276.

Ferrand, Y. B., Magazine, M. J., Rao, U. S., 2014. Managing operating room efficiency and responsiveness for emergency and elective surgeries—a literature survey. IIE Transactions on Healthcare Systems Engineering 4 (1), 49–64.

Gorissen, B. L., Yanıkoğlu, İ., den Hertog, D., 2015. A practical guide to robust optimization. Omega 53, 124–137.

Gul, S., Denton, B. T., Fowler, J. W., 2015. A progressive hedging approach for surgery planning under uncertainty. INFORMS Journal on Computing 27 (4), 755–772.

Gupta, D., Denton, B., 2008. Appointment scheduling in health care: Challenges and opportunities. IIE Transactions 40 (9), 800–819.

Hollander, J. E., Carr, B. G., 2020. Virtually perfect? telemedicine for covid-19. New England Journal of Medicine 382 (18), 1679–1681.

Holte, M., Mannino, C., 2013. The implementor/adversary algorithm for the cyclic and robust scheduling problem in health-care. European Journal of Operational Research 226 (3), 551–559.

Jetty, A., Moore, M. A., Coffman, M., Petterson, S., Bazemore, A., 2018. Rural family physicians are twice as likely to use telehealth as urban family physicians. Telemedicine and e-Health 24 (4), 268–276.

Ji, M., Li, J., Peng, C., 2020. Two-stage chance-constrained telemedicine assignment model with no-show behavior and uncertain service duration. In: AI and Analytics for Public Health, Proceedings of INFORMS International Conference on Service Science. Springer, pp. xx–xx (in press).

Jnr, B. A., 2020. Use of telemedicine and virtual care for remote treatment in response to covid-19 pandemic. Journal of Medical Systems 44 (7), 1–9.

Kamran, M. A., Karimi, B., Dellaert, N., 2018. Uncertainty in advance scheduling problem in operating room planning. Computers & Industrial Engineering 126, 252–268.

KC, D. S., Scholtes, S., Terwiesch, C., 2020. Empirical research in healthcare operations: past research, present understanding, and future opportunities. Manufacturing & Service Operations Management 22 (1), 73–83.

Kim, K., Mehrotra, S., 2015. A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. Operations Research 63 (6), 1431–1451.

Kleywegt, A. J., Shapiro, A., Homem-de Mello, T., 2002. The sample average approximation method for stochastic discrete optimization. SIAM Journal on Optimization 12 (2), 479–502.

Laporte, G., Louveaux, F. V., 1993. The integer l-shaped method for stochastic integer programs with complete recourse. Operations Research Letters 13 (3), 133–142.

Manshadi, S. D., Khodayar, M. E., 2017. Risk-averse generation maintenance scheduling with microgrid aggregators. IEEE Transactions on Smart Grid 9 (6), 6470–6479.

Matthews, L. R., Gounaris, C. E., Kevrekidis, I. G., 2019. Designing networks with resiliency to edge failures using two-stage robust optimization. European Journal of Operational Research 279 (3), 704–720.

Min, D., Yih, Y., 2010. Scheduling elective surgery under uncertainty and downstream capacity constraints. European Journal of Operational Research 206 (3), 642–652.

Myers, M. R., 2003. Telemedicine: an emerging health care technology. Health Care Management 22 (3), 219–223.

Najjarbashi, A., Lim, G. J., 2019. A variability reduction method for the operating room scheduling problem under uncertainty using cvar. Operations Research for Health Care 20, 25–32.

Neyshabouri, S., Berg, B. P., 2017. Two-stage robust optimization approach to elective surgery and downstream capacity planning. European Journal of Operational Research 260 (1), 21–40.

Nohadani, O., Sharma, K., 2018. Optimization under decision-dependent uncertainty. SIAM Journal on Optimization 28 (2), 1773–1795.

Peng, C., Delage, E., Li, J., 2020. Probabilistic envelope constrained multiperiod stochastic emergency medical services location model and decomposition scheme. Transportation Science 54 (6), 1471–1494.

Qiao, Y., Ran, L., Li, J., 2020. Optimization of teleconsultation using discrete-event simulation from a data-driven perspective. Telemedicine and e-Health 26 (1), 112–123.

Rahimian, H., Mehrotra, S., 2019. Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659.

Rahmaniani, R., Crainic, T. G., Gendreau, M., Rei, W., 2017. The benders decomposition algorithm: A literature review. European Journal of Operational Research 259 (3), 801–817.

Rajan, B., Tezcan, T., Seidmann, A., 2019. Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. Management Science 65 (3), 1236–1267.

Rath, S., Rajaram, K., Mahajan, A., 2017. Integrated anesthesiologist and room scheduling for surgeries: Methodology and application. Operations Research 65 (6), 1460–1478.

Saghafian, S., Hopp, W. J., Iravani, S. M., Cheng, Y., Diermeier, D., 2018. Workload management in telemedical physician triage and other knowledge-based service systems. Management Science 64 (11), 5180–5197.

Saif, A., Delage, E., 2021. Data-driven distributionally robust capacitated facility location problem. European Journal of Operational Research 291 (3), 995–1007.

Shehadeh, K. S., Padman, R., 2021. A distributionally robust optimization approach for stochastic elective surgery scheduling with limited intensive care unit capacity. European Journal of Operational Research 290 (3), 901–913.

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

Sun, S., Lu, S. F., Rui, H., 2020. Does telemedicine reduce emergency room congestion? evidence from new york state. Information Systems Research 31 (3), 972–986.

Wang, S., Li, J., Mehrotra, S., 2021a. Chance-constrained bin packing problem with an application to operating room planning. INFORMS Journal on Computing 33 (4), 1661–1677.

Wang, S., Li, J., Mehrotra, S., 2021b. A solution approach to distributionally joint robust chance-constrained assignment problems. INFORMS Journal on Optimization (forthcoming), pre-print via `http://www.optimization-online.org/DB_FILE/2019/05/7207.pdf`.

Wang, S., Li, J., Peng, C., 2017. Distributionally robust chance-constrained program surgery planning with downstream resource. In: 2017 International Conference on Service Systems and Service Management. IEEE, pp. 1–6.

Wang, S., Mehrotra, S., 2021. Robust concave utility maximization over a chance constraint. working paper, `http://www.optimization-online.org/DB_FILE/2021/12/8716.pdf`.

Wang, X., Zhang, Z., Yang, L., Zhao, J., 2020. Price and capacity decisions in a telemedicine service system under government subsidy policy. International Journal of Production Research, 1–14.

Wang, Y., Zhang, Y., Tang, J., 2019. A distributionally robust optimization approach for surgery block allocation. European Journal of Operational Research 273 (2), 740–753.

Wang, Z., Qi, M., 2020. Robust service network design under demand uncertainty. Transportation Science 54 (3), 676–689.

Wiesemann, W., Kuhn, D., Sim, M., 2014. Distributionally robust convex optimization. Operations Research 62 (6), 1358–1376.

Xie, S., Hu, Z., Wang, J., 2020. Two-stage robust optimization for expansion planning of active distribution systems coupled with urban transportation networks. Applied Energy 261, 114412.

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

Zanaboni, P., Scalvini, S., Bernocchi, P., Borghi, G., Tridico, C., Masella, C., 2009. Teleconsultation service to improve healthcare in rural areas: acceptance, organizational impact and appropriateness. BMC Health Services Research 9 (1), 238.

Zeng, B., Zhao, L., 2013. Solving two-stage robust optimization problems using a column-and-constraint generation method. Operations Research Letters 41 (5), 457–461.

Zhai, Y., Wang, Y., Zhang, M., Gittell, J. H., Jiang, S., Chen, B., Cui, F., He, X., Zhao, J., Wang, X., 2020. From isolation to coordination: how can telemedicine help combat the covid-19 outbreak? MedRxiv.

Zhang, Y., Jiang, R., Shen, S., 2018a. Ambiguous chance-constrained binary programs under mean-covariance information. SIAM Journal on Optimization 28 (4), 2922–2944.

Zhang, Y., Shen, S., Erdogan, S. A., 2018b. Solving 0–1 semidefinite programs for distributionally robust allocation of surgery blocks. Optimization Letters 12 (7), 1503–1521.

Zhang, Z., Denton, B. T., Xie, X., 2020. Branch and price for chance-constrained bin packing. INFORMS Journal on Computing.

Zhao, L., Zeng, B., 2012. An exact algorithm for two-stage robust optimization with mixed integer recourse problems. submitted, available on Optimization Online.

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

Zhu, S., Fan, W., Yang, S., Pei, J., Pardalos, P. M., 2019. Operating room planning and surgical case scheduling: a review of literature. Journal of Combinatorial Optimization 37 (3), 757–805.

Appendix A A Two-Stage Stochastic Programming Model

In this appendix, we present a two-stage stochastic programming (denoted by 2SP) counterpart for the telemedicine assignment problem when incorporating three different sources of uncertainty.

For the 2SP model, the way how we capture the no-show behaviour of the telemedical doctors is the main difference between the 2RO and 2SP models. In order to propose a comparable benchmark model, it should be noted that the no-show behaviour of the telemedical doctors is also dependent on the "budgeted" robustness parameter Γ that is used in the 2RO model. Here we still introduce a set of random binary variables z_j^ω to capture the no-show behaviour of the telemedical doctors, and we allow that at most Γ $(1 \leq \Gamma \leq 9)$ telemedical doctor no-shows. We assume that each telemedical doctor has the same no-show probability (denoted by γ) and their no-shows are independent of each other. Specifically, $z_j^\omega = 1$ indicates that telemedical doctor j will be no-show under scenario ω, and otherwise $z_j^\omega = 0$. We keep the same notations that are used in our paper. Since we note that the 2SP model has a similar two-stage optimization structure and shares a set of constraints, here we omit the description of notations and constraints for the sake of simplicity. Then, the 2SP model is presented as follows:

$$
\begin{aligned}
\textbf{[2SP]}: \quad \underset{\boldsymbol{x},\boldsymbol{y}}{\text{minimize}} \;& \rho \sum_{\omega\in\Omega} p_\omega \left(\sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} c_{ij} x_{ij}^\omega + \sum_{j\in\mathcal{J}} h_j y_j \right) + (1-\rho)\mathcal{S}(\boldsymbol{y}) && \\
\text{subject to} \;& x_{ij}^\omega \leq y_j && \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega \\
& \sum_{j\in\mathcal{J}} x_{ij}^\omega = \delta_i^\omega && \forall i \in \mathcal{I}, \omega \in \Omega \\
& \sum_{i\in\mathcal{I}} d_i^\omega x_{ij}^\omega \leq T y_j && \forall j \in \mathcal{J}, \omega \in \Omega \\
& \sum_{j\in\mathcal{J}} y_j \leq K && \\
& x_{ij}^\omega \in \{0,1\}, y_j \in \{0,1\} && \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega \\
\textbf{where} \;\; \mathcal{S}(\boldsymbol{y}) = \underset{\boldsymbol{q},\boldsymbol{o},\boldsymbol{u}}{\text{minimize}} \;& \sum_{\omega\in\Omega} p_\omega \left(\sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} c_{ij} q_{ij}^\omega + \sum_{j\in\mathcal{J}} b_j o_j^\omega + \sum_{i\in\mathcal{I}} r_i u_i^\omega \right) && \\
\text{subject to} \;& \sum_{j\in\mathcal{J}} q_{ij}^\omega + u_i^\omega = \delta_i^\omega && \forall i \in \mathcal{I}, \omega \in \Omega \\
& \sum_{i\in\mathcal{I}} d_i^\omega q_{ij}^\omega \leq T y_j (1 - z_j^\omega) + o_j^\omega && \forall j \in \mathcal{J}, \omega \in \Omega \\
& o_j^\omega \leq v_j y_j (1 - z_j^\omega) && \forall j \in \mathcal{J}, \omega \in \Omega \\
& q_{ij}^\omega \in \{0,1\}, u_i^\omega \in \{0,1\}, o_j^\omega \geq 0 && \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega,
\end{aligned}
$$

where the objective function minimizes the expected weighted total cost of the cases with/without incorporating the no-show behaviour of the telemedical doctors.