IISE Transactions

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

Continuity-skill-restricted scheduling and routing problem: Formulation, optimization and implications

Mingda Liu, Yanlu Zhao & Xiaolei Xie

To cite this article: Mingda Liu, Yanlu Zhao & Xiaolei Xie (2023): Continuity-skill-restricted scheduling and routing problem: Formulation, optimization and implications, IISE Transactions, DOI: 10.1080/24725854.2023.2215843

To link to this article: https://doi.org/10.1080/24725854.2023.2215843

IISE TRANSACTIONS
https://doi.org/10.1080/24725854.2023.2215843

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Continuity-skill-restricted scheduling and routing problem: Formulation, optimization and implications

Mingda Liu[a], Yanlu Zhao[b], and Xiaolei Xie[a]

[a]Department of Industrial Engineering, Tsinghua University, Beijing, China; [b]Durham University Business School, Durham University, Durham, UK

ABSTRACT
As the aging population grows, the demand for long-term continuously Attended Home Healthcare (AHH) services has increased significantly in recent years. AHH services are beneficial since they not only alleviate the pressure on hospital resources, but also provide more convenient care for patients. However, how to reasonably assign patients to doctors and arrange their visiting sequences is still a challenging task due to various complex factors such as heterogeneous doctors, skill-matching requirements, continuity of care, and uncertain travel and service times. Motivated by a practical problem faced by an AHH service provider, we investigate a deterministic continuity-skill-restricted scheduling and routing problem (CSRP) and its stochastic variant (SCSRP) to address these operational challenges. The problem is formulated as a heterogeneous site-dependent and consistent vehicle routing problem with time windows. However, there is not a compact model and a practically implementable exact algorithm in the literature to solve such a complicated problem. To fill this gap, we propose a branch-price-and-cut algorithm to solve the CSRP and a discrete-approximation-method adaption for the SCSRP. Extensive numerical experiments and a real case study verify the effectiveness and efficiency of the proposed algorithms and provide managerial insights for AHH service providers to achieve better performance.

ARTICLE HISTORY
Received 25 April 2022
Accepted 11 May 2023

KEYWORDS
Attended home healthcare; heterogeneous site-dependent VRPTW; continuity restriction; skill-matching; branch-price-and-cut

1. Introduction

According to the World Health Organization estimation, 22% of the global population will be over 60 years in age by 2050 (World Health Organization, 2014). As the aging population is medically vulnerable, the demand for healthcare resources is rapidly increasing (Gupta and Denton, 2008). In recent years, an effective measure to relieve hospital congestion is emerging as home medical care or convalescent care for patients, which has encouraged service areas such as Attended Home Healthcare (AHH). AHH services not only reduce avoidable hospital readmission rates, but also are typically less expensive, more convenient, and just as effective as the care provided in a hospital or nursing facility. As a result, the AHH services expansion is on a fast lane and perceived as a new opportunity for healthcare excellence. Against this backdrop, AHH has gained increasing attention in academia over the past 5 years. Most existing studies investigate marketing, scheduling, and planning-related decisions, but in many developed countries these decisions are still made manually; see Fikar and Hirsch (2017) for a comprehensive review. Due to the increasing demand for AHH services, it is imperative and worthwhile to provide effective solutions to enable patients to access AHH services in a timely and convenient manner.

This research was motivated by a real-world scheduling problem at *Pinetree*, a Chinese AHH service provider based in Beijing. Pinetree (http://www.qskh.cn) was founded in 2004 and provides primary home medical care and assistance to patients. The company manages more than 1000 part-time doctors and has scheduled their visits to more than 600,000 home-based patients for AHH services. Each week, given the appointments and time windows required for the visits, the schedulers take full responsibility for manually determining the routing and scheduling for doctors based on personal past experience and improvisation. To address this challenging problem and also resolve certain limitations within existing research (as reviewed in Section 1.1), we investigate a general AHH scheduling and routing problem that includes the following important features:

- First, the doctors are *depot-dependent* (different home/-depot addresses) and *pattern-dependent* (different availability/days to provide service). In this regard, most existing AHH studies assume that only a single depot is available for doctors to leave/return every day and the doctors are always available over the whole planning horizon (Gamst and Jensen, 2012; Trautsamwieser and

CONTACT Xiaolei Xie xxie@tsinghua.edu.cn
Supplemental data for this article is available online at https://doi.org/10.1080/24725854.2023.2215843

Hirsch, 2014). However, for convenience, part-time and even full-time doctors usually leave and return to their homes directly and have their own diverse availability to provide AHH services every week (Mosquera *et al.*, 2019). In this article, departing from the existing studies, different depots and availability for individual doctors are considered to reflect more practical operations. Different from the traditional Multi-Depot Vehicle Routing Problem (MDVRP) (Rabbouch *et al.*, 2018), the assignment of doctors to depots is achieved here implicitly and does not introduce additional decision variables.

- The problem is further complicated by *skill-matching* requirements between doctors and patients, which indicate that the assigned doctors must satisfy the medical demand and various additional requirements of patients, such as gender, age, and task types. The traditional skill VRP stipulates that customers' requirements are fulfilled by vehicles with higher skill levels (Cappanera *et al.*, 2011) or compatible pattern (Shao *et al.*, 2012). However, in AHH practice, the patient requirements (tasks) are standardized and each doctor is only expert at a subset of these tasks. Given the potential mismatch between demand and supply, the scheduler has to allocate scarce resources (i.e., doctors with specific scarce skills) to the most demanding patients, thereby improving the overall patient coverage. The problem is converted into a Site-Dependent Vehicle Routing Problem (SDVRP) to cover the skill-matching requirements (Baldacci *et al.*, 2010), but we extend the traditional SDVRP from one-dimensional determinant (skill) to two-dimensional factors (skill and availability) when determining the compatibility of a doctor.
- Another intricate factor is *continuity*, which stipulates that the number of doctors caring for the same patient within the prescribed time horizon should be limited within a cohort to ensure consistently high quality and satisfaction when providing service, and avoid imbalanced workload across doctors. This factor has been considered as consistency constraints in the literature, which require that the nodes are consistently served by the same vehicle (Lian *et al.*, 2016; Song *et al.*, 2020). However, we relax the consistency requirement from a particular vehicle to a subset of vehicles to reflect more practical operations, bringing more flexibility in the scheduling decisions but together with extra challenges. Zhao and Alfandari (2020) investigated the possibility to visit each node by a subset of vehicles, however, under the situation of one common depot and thus is different from the heterogeneous depots setting here. To this end, we refer to the consistent vehicle routing problem (ConVRP) and extend it from a consistent singleton to a consistent subset.

To better illustrate the above features, we present a realistic case in Figure 1 with five patients and four doctors within 2 days, where each patient and doctor has its own address, skills and availability (heterogeneity). The data triplet next to each patient displays the individual information, including time window, skill-matching requirements, service duration, arrival time, and service start time. The gray node denotes that the patient/doctor is unavailable and can not be visited/assigned on that day. The optimal routes are shown in connected lines in which each patient can receive AHH service at a prescribed time slot provided by her familiar doctors. Specifically, patient 4 is assigned to doctor 2 on day 2 rather than to a geographically closer doctor 3, due to the skill-matching requirements. In addition, this patient could be assigned to doctor 0 on day 1 but re-assigned to doctor 2 on day 2, since her continuity requirement is relaxed to 2, allowing additional flexibility in scheduling visits given doctor 0 is unavailable on day 2. By considering these features in the AHH problem, we aim to describe a consistent scheduling and routing problem for doctors that takes into account their heterogeneous characteristics. We therefore define this problem as the *Continuity-Skill-Restricted scheduling and routing Problem (CSRP)*. Clearly, the following three distinct features of the CSRP render it rather challenging to be addressed exactly:

1. First and foremost, the complexity is compounded by the inconsistent availability among doctors and patients across the planning horizon, indicating that simply repeating the daily agendas is impractical. The solution procedure should differentiate networks and generate multiple graphs with respect to different combinations of doctor and time (i.e., day). In other words, several heterogeneous SDVRP are addressed on different days, which is certainly more difficult than the classic VRP.
2. Designing exact solution methods including the continuity constraints is challenging, as these constraints come into effect over multiple periods and multiple depot-dependent graphs to ensure the number of doctors visiting each patient is less than a threshold. In this regard, it is more challenging than the general ConVRP, as the latter only describes a special case of our problem by forcing the continuity parameter to equal one in the CSRP.
3. Last but not least, the interaction between the site-dependency and continuity, for example, allocating the doctors with specific skills (which can be perceived as scarce resources) to the most demanding patients at an appropriate slot, leads to foreseeably extra computational iterations when searching for optimal solutions.

Therefore, in terms of modeling perspective, the CSRP can be affiliated as a *heterogeneous site-dependent consistent vehicle routing problem with time windows (HSDConVRPTW)*, where doctors are represented as vehicles and patients are described as nodes. For ease of explanation, we might abuse these terminologies when it is not misleading in the following sections. As far as we know, there is not yet a compact model and an exact solution approach in the literature that can handle such a complicated HSDConVRPTW with the aforementioned practical features. In light of these observations, this article is dedicated to developing a high-performance exact algorithm to search

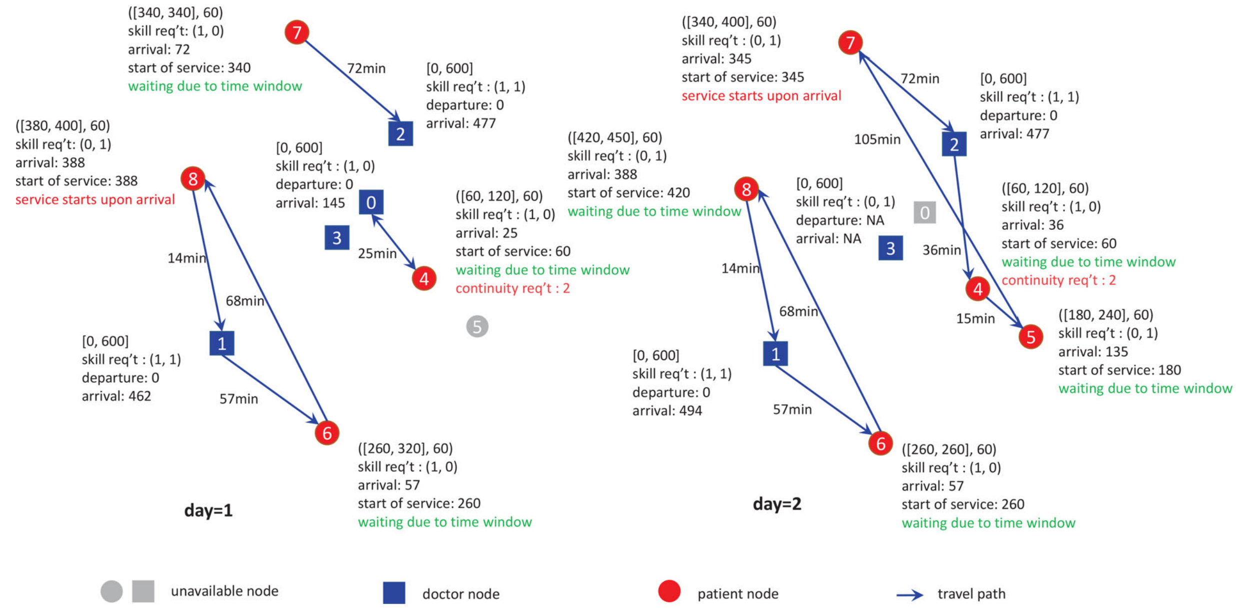

Figure 1. A 2-day scheduling example in Pinetree.

for solutions of proven optimality. In the end, by dropping certain AHH-specific features, the CSRP can be adapted to address traditional VRPTW problems. As a result, the proposed model and algorithms can be extended to tackle similar scheduling and routing problems as well as VRPTW applications such as operating room planning and scheduling (Naderi *et al.*, 2021), waste collection (Beliën *et al.*, 2014), school bus routing and scheduling (Park and Kim, 2010) and petroleum products distribution (Ronen, 1995).

In the following, we briefly outline the related works on AHH research and their solution approaches in the literature to clearly position the contributions of this article.

1.1. Related literature

The AHH scheduling and routing problem has been extensively studied in recent years. We refer interested readers to Fikar and Hirsch (2017) for a comprehensive overview. In this context, continuity of care (consistency) is an important determinant of service quality and patient satisfaction (Kovacs *et al.*, 2015a). Among others, most studies on the ConVRP focus on vehicle consistency and visiting time consistency, where the vehicle and visiting time are determined for each customer (Kovacs *et al.*, 2015b; Goeke *et al.*, 2019). Kovacs *et al.* (2015b) generalized the vehicle consistency constraint by considering a limited set of vehicles, which is a widely accepted practice in AHH problems, to create a long–term consistent schedule in the caregiver–patient assignments by introducing an additional decision variable (Cappanera and Scutellà, 2015; Cappanera *et al.*, 2018). Moreover, caregivers are rarely homogeneous in practice, and most studies assume that caregivers are heterogeneous and differ by demographic characteristics (e.g., residence and address) or by professional characteristics (e.g., skill and gender) (Cappanera *et al.*, 2011; Cappanera and Scutellà, 2015; Grenouilleau *et al.*, 2019; Hashemi Doulabi *et al.*, 2020). The most similar work to us is Stavropoulou (2022), which considered heterogeneous vehicle fleets that arose in practical ConVRP environments. However, to the best of our knowledge, HSDConVRPTW has not yet been considered in the literature. We also refer the reader to some other interesting studies related to AHH, such as periodic fixed appointment scheduling (Bennett and Erera, 2011), the multigraph representation (Bard *et al.*, 2014), and studies on fixed and overtime costs (Naderi *et al.*, 2023).

There are a large number of studies that have designed solution approaches for AHH scheduling and routing problems. In this regard, Cappanera and Scutellà (2015) proposed an integrated pattern-based heuristic to address weekly assignments, scheduling and routing decisions with continuity constraints, where the pattern specifies a possible schedule for skilled visits and provides the key insight to address skill-matching requirements. Grenouilleau *et al.* (2019) investigated a practical weekly routing and scheduling problem in home health care and designed a set partitioning heuristic to incorporate most of the AHH service requirements. However, most previous studies focused on efficient heuristics, such as the scenario-based approach in Demirbilek *et al.* (2019), or the adaptive large neighborhood search metaheuristic in Yang *et al.* (2021), to efficiently find acceptable solutions for large-scale cases. In addition, there are some studies reporting that the Logic-Based Benders Decomposition (LBBD) method may have been applied to address AHH variations by focusing on maximizing the number of patients visited or determining the optimal number of caregivers, taking into account overtime costs (Heching *et al.*, 2019; Grenouilleau *et al.*, 2020). However, requiring a tailored design of master and subproblems based on compatible decomposition structures, the LBBD method has not been developed to effectively address complex problems like the HSDConVRPTW here. Consequently, in this regard, there is still no existing exact solution approach capable of solving such problem for practical requirements.

Table 1. Position of this article in the literature regarding AHH problem characteristics.

	Time horizon		*Depot*							*Solution method*	
Literature	*Single*	*Multiple*	*Single*	*Multiple*	*TW*	*SK*	*WT*	*UC*	*C*	*Exact*	*Heuristic*
Cordeau *et al.* (1997)		✓		✓			H				✓
Cappanera and Scutellà (2015)		✓	✓			H	H		H		✓
Cappanera *et al.* (2018)		✓	✓			H	H	RO	H		✓
Liu *et al.* (2019)	✓			✓	H			CC		BP	
Heching *et al.* (2019)		✓		✓	H	H	H			LBBD	
Grenouilleau *et al.* (2019)		✓		✓	H	S	S		EP		✓
Hashemi Doulabi *et al.* (2020)	✓		✓		S	H	S	SO		L-shaped	
Sauré *et al.* (2020)	✓		✓			H	H	SO			✓
Yang *et al.* (2021)		✓	✓		S		S	SO	SO		✓
Zhan *et al.* (2021)	✓		✓		S			SA		L-shaped	
Naderi *et al.* (2023)		✓	✓			H	S	RO		LBBD	
This article		✓		✓	H	H	H	CC	H	(DAM-)BPC	

TW, Time Windows (arrive on time); SK, Skill-matching Requirements; WT, Working Time Regulation; UC, Uncertainty; C, Continuity Constraint
H, Hard constraints (i.e., forced to satisfy); S: Soft constraints (i.e., violated with penalty); EP: Exogenous Parameters
RO, Robust Optimization; SO, Stochastic Optimization; CC, Chance-Constrained; SA, Scenarios Approximation

In recent decades, the Branch-Price-and-Cut (BPC) method has been widely applied in VRP and its variants to achieve optimality (Archetti *et al.*, 2011; Zhen *et al.*, 2023). We recommend readers to Costa *et al.* (2019) for a comprehensive discussion of this method. Focusing on AHH problems, the BPC method is able to incorporate all the AHH service features into pricing subproblems and solve them efficiently through dynamic programming. In fact, several studies have applied the BPC method to solve the scheduling and routing problem in the AHH industry. For example, Trautsamwieser and Hirsch (2014) considered homogeneous caregivers for weekly schedules and developed BPC algorithms for single-depot VRP, and Liu *et al.* (2019) considered heterogeneous caregivers in terms of single-day planning, which is a special case of our work. In this article, we propose a dedicated BPC algorithm to solve the HSDConVRPTW. In particular, we consider continuity constraints in the master problem, which can be handled as robust cuts by incorporating corresponding pricing graphs in terms of different days and doctors, hence maintaining the structure of the pricing problem. Moreover, our BPC method can incorporate chance constraints to deal with uncertain travel and service times by extending the resource and path inequalities from the deterministic to the robust optimization context and generating robust routes on the premise of on-time service and working time regulation (Liu *et al.*, 2019; Munari *et al.*, 2019). For brevity, a summary of the relevant literature is presented in Table 1.

1.2. Contributions

Facing the above challenges in AHH scheduling and routing decisions, this article aims to answer the following questions: (i) How should visits be scheduled and solved in advance to maximize visit effectiveness and minimize travel time? (ii) How are the performances (e.g., the efficiency of the solution approach and the effectiveness of the routing results) affected by the design parameters and how can the values of these parameters be determined? To answer these questions, the main contributions of this article are summarized as follows:

- We formally model the CSRP as a mixed-integer program and present a compact, arc-based formulation with a polynomial number of constraints and variables, which is quite inclusive of practical characteristics such as skill-matching requirements, continuity of care, working time regulation and doctor heterogeneity. Mathematically, the CSRP compact formulation is categorized as an HSDConVRPTW, which captures most of the required practical features as hard constraints. These practice-required features allow our model to characterize more general AHH operations and variations. As far as we know, there is no evidence of a compact model in the literature that has handled all the above practical features as hard constraints, and our HSDConVRPTW model aims to fill this gap.
- Since the compact formulation is NP-hard, commercial solvers cannot solve it optimally for medium- and large-scale instances in real environments. We therefore present a set-partitioning reformulation and propose a dedicated BPC algorithm to solve the problem. We develop a greedy algorithm to generate the initial route pool and a tabu search heuristic pricing algorithm to efficiently obtain potential routes with a negative reduced cost. Among other advances, we apply the acceleration approaches including bi-direction and ng-route to increase computational efficiency and speed up the convergence of column generation. Finally, since the subset row cuts, which are generated as Chvatal–Gomory cuts to eliminate infeasible solutions, are non-robust cuts, we also redesign the dominance rule accordingly. As the literature review indicated, our tailored BPC algorithm is the first exact method to solve the AHH problem with continuity constraints, or mathematically speaking, an HSDConVRPTW. Specifically, we measure the impact of continuity constraints across different temporal and geographical horizons. We demonstrate that, by handling pricing graphs in terms of different days and doctors, the induced continuity constraints in the master problem are robust cuts and will not change the structure of the pricing problem.
- Given the *uncertain nature* of travel and service times, optimal solutions for deterministic CSRP usually fall short in practice to ensure the promised on-time service and the working time regulation. In other words, doctors

probably arrive later than the pre-determined appointments with patients and complete all assigned work later than the working time regulation. Aligning with more practical requirements, we further investigate a *Stochastic Continuity-Skill-Restricted scheduling and routing Problem (SCSRP)*. More precisely, chance constraints are formulated for each patient and doctor to impose that the probability of lateness is less than a prescribed criterion. To reduce the risk of lateness, the method developed for the deterministic problem is further extended by integrating the BPC algorithm with a Discrete Approximation Method (DAM-BPC), which renders flexibility for the set partitioning model in dealing with on-time service and completion constraints against uncertain travel and service times in pricing subproblems. To this end, we update the resource extension functions and dominance rule in the column generation procedure by incorporating the uncertainty features. The DAM-BPC framework presents the potential of integrating the BPC method with advanced stochastic or distributionally robust optimization methods where the column generation approach is adopted to tackle problems that involve uncertain elements. For instance, the effectiveness of the DAM method can be enhanced with the service fulfillment risk index (SRI) which is evaluated under specific ambiguity set close to practical operational contexts (Zhang *et al.*, 2021).

- We perform extensive numerical studies on several benchmark data sets in the literature and a real-world dataset to demonstrate the effectiveness of our approach and analyze the impact of various underlying parameters. The results show that the proposed algorithms clearly outperform the commercial solver CPLEX and other existing methods (e.g., BP, L-Shaped and LBBD), and are significantly efficient in handling practical scale instances. The results also provide important managerial insights to improve operational performance. After implementing our scheduling strategy, the collaborating company reported a significant improvement in their home visit operations. The average daily home visits increased by almost 20%, from 2.76 to 3.35, and the average travel time per visit was reduced by at least 30.6 minutes (from 113.9 minutes to 83.3 minutes). Furthermore, they were able to reduce the number of employed doctors from 145 to 119, resulting in a 17% reduction in employment costs and an estimated monthly cost savings of CNY 117,000 ($17,035.2). These results demonstrate the efficacy and cost-effectiveness of our method when compared with the manual scheduling scheme previously used in the company. In addition, patient satisfaction has been improved as patients can receive care from their trusted doctors. Our analysis shows that it is worthwhile to calibrate the continuity constraint in scheduling to balance the trade-off between patient satisfaction and scheduling flexibility.

The remainder of this article is organized as follows. First, in Section 2, we give a brief problem description and an arc-based compact formulation for the CSRP. Section 3 describes the BPC algorithm, including the set-partitioning reformulation and the pricing subproblem. In Section 4, we introduce the SCSRP and address it with a DAM-BPC. The computational results and managerial insights are presented in Section 5, followed by the concluding remarks and future research in Section 6.

2. Problem description and formulation

In this section, we formulate the CSRP formally as the following Mixed-Integer Programming (MIP) model. Let T be the planning time horizon. The CSRP is defined on a complete directed graph, $G = (V, A)$, where V is the set of vertices and A is the set of edges. We partition set V as $V = I \cup J$, where $I = \{0, ..., |I| - 1\}$ is the node set of doctor depots and $J = \{|I|, ..., |I| + |J| - 1\}$ is the node set of patients. The edge set A is defined as $\{(i,j) : i \in V, j \in V\}$ by removing arc set $\{(i,j) : i \in I, j \in I\}$. The deterministic travel times on arcs $(i,j) \in A$ are denoted as t_{ij}, and the deterministic service times at patients $j \in J$ are denoted as t_j. Let e_j^d (i.e., l_j^d) denote the earliest (i.e., the latest) service (appointment) start time for patient $j \in J$ on day $d \in T$, and let B denote a sufficiently large constant. According to the working time regulation, let ξ represent the maximum workload criterion in a day, which is the upper bound for the total service and travel times of doctors. For modeling convenience, we also declare an additional index set $O = \{0, ..., |I| - 1\}$ to represent the doctors. The doctor index in O is aligned with the depots in I, so we might slightly abuse either of them to represent doctors for notational convenience when it is not misleading.

We now describe the elements of availability and skill-matching requirements. First, we define a binary parameter P_v^d that specifies the availability of node v on day d. On the one hand, if $v \in I$, then $P_v^d = 1$ means that doctor v is able to provide services on day d; otherwise, the doctor cannot visit patients on that day. On the other hand, if $v \in J$, then $P_v^d = 1$ means that patient v needs to be served on day d; otherwise, the patient does not require the service. Moreover, the heterogeneity of doctors is also reflected in the requirements of patients. We define a set M to represent the service requirements of patients, while M should also be the corresponding skill set of doctors. With the set M, let another binary parameter K_{om} denote whether doctor $o \in O$ owns skill $m \in M$. If so, $K_{om} = 1$; otherwise, $K_{om} = 0$. Similarly, we use a binary parameter $D_{jm} = 1$ to represent service requirement $m \in M$ of patient $j \in J$. Finally, we define a continuity parameter ω, indicating that the number of different doctors serving the same patient within the time horizon should not exceed ω. Note that, ω can be patient-differentiated. However, as it does not significantly affect the model applicability, we keep the same ω for all patients for modeling convenience. With these notations, the decision variables of the arc-based model are defined as follows:

$$x_{ij}^{od} = \begin{cases} 1, & \text{if doctor } o \text{ traverses arc } (i,j) \text{ on day } d, (i,j) \in A, o \in O, d \in T; \\ 0, & \text{otherwise.} \end{cases}$$
$$u_{oj} = \begin{cases} 1, & \text{if doctor } o \text{ serves patient } j \text{ at least once within } T, o \in O, d \in T; \\ 0, & \text{otherwise.} \end{cases}$$
$$S_i^d = \text{the actual service start time of node } i \text{ on day } d, i \in V, d \in T.$$

The objective of the CSRP is to minimize the total travel time for all doctors within the planning time horizon:

$$\min \quad \sum_{d \in T} \sum_{o \in O} \sum_{(i,j) \in A} t_{ij} x_{ij}^{od} \tag{1}$$

To satisfy the above requirements and obtain exact solutions, the constraints are described as:

$$\sum_{o \in O} \sum_{j \in J} x_{ij}^{od} \leq P_i^d, \forall d \in T, i \in I \tag{2}$$

$$\sum_{o \in O} \sum_{(i,j) \in A} x_{ij}^{od} = P_j^d, \forall d \in T, j \in J \tag{3}$$

$$\sum_{o \in O} \sum_{j \in J} x_{ij}^{od} = \sum_{o \in O} \sum_{k \in J} x_{ki}^{od}, \forall d \in T, i \in I \tag{4}$$

$$\sum_{(i,j) \in A} x_{ij}^{od} = \sum_{(j,k) \in A} x_{jk}^{od}, \forall d \in T, o \in O, j \in J \tag{5}$$

$$S_j^d + t_j + t_{ji} \leq \xi + B(1 - x_{ji}^{od}), \forall d \in T, o \in O, i \in I, j \in J \tag{6}$$

$$S_i^d + t_i + t_{ij} \leq S_j^d + B(1 - x_{ij}^{od}), \forall d \in T, o \in O, i \in V, j \in J, i \neq j \tag{7}$$

$$e_j^d \leq S_j^d \leq l_j^d, \forall d \in T, j \in J \tag{8}$$

$$S_i^d = 0, \forall d \in T, i \in I \tag{9}$$

$$\sum_{(i,j) \in A} x_{ij}^{od} - K_{om} + D_{jm} \leq 1, \forall d \in T, o \in O, j \in J, m \in M \tag{10}$$

$$\sum_{o \in O} u_{oj} \leq \omega, \forall j \in J \tag{11}$$

$$x_{ij}^{od} \leq u_{oj}, \forall (i,j) \in A, d \in T, o \in O, j \in J \tag{12}$$

$$u_{oj} \leq \sum_{d \in T} \sum_{(i,j) \in A} x_{ij}^{od}, \forall o \in O, j \in J \tag{13}$$

$$x_{ij}^{od} \in \{0, 1\}, \forall d \in T, o \in O, (i,j) \in A \tag{14}$$

$$u_{oj} \in \{0, 1\}, \forall o \in O, j \in J \tag{15}$$

$$S_i^d \geq 0, \forall d \in T, i \in V \tag{16}$$

These constraints can be described in three aspects: route-defining constraints (2)-(9), skill-matching constraints (10) and continuity constraints (11)-(13). Specifically, constraints (2)-(3) guarantee the availability of doctors and patients, and ensure that demanding patients are visited. Constraints (4)-(5) refer to the flow conservation requirements on depots and patient nodes. Constraints (6) ensure that working time regulation is fulfilled. Constraints (7)-(9) specify that the start time for serving patient j on day d should be no earlier than e_j^d and no later than l_j^d. Constraints (10) impose skill-matching on each visit and enforce that all skill requirements are satisfied. Constraints (11) guarantee that at most ω doctors are assigned to a patient within the time horizon. In addition, mimicking the Emergency Medical Systems for ambulances usage constraints such as "at least 90% of the emergency calls must be answered within 9 minutes", we also provide and test alternative service level requirement constraints by imposing that at least a prescribed percentage of patients receiving service with the cohort size of doctors being less than ω in Appendix A. Constraints (12)-(13) are linking constraints between routing and assignment variables. Constraints (14)-(16) define the variable characteristics.

In summary, the above formulation (1)-(16) contains at least $O(|T||I||A|)$ variables and $O(|T||I||J|^2)$ constraints. This model becomes intractable for commercial solvers as the instance scale increases (as shown in Section 5). Therefore, we develop a BPC algorithm that can obtain optimal solutions for large instances in the following sections.

Algorithm 1: BPC algorithm for the CSRP

Input: Data sets with pattern information and distance matrix T, I, J, O, M, P, D; Parameters ω, ξ.
Output: Sol_{best} is the optimal solution or best-known upper bound.
Initialization: Original graph $G = (V, A)$; Best incumbent value $Sol_{best} \leftarrow \infty$; Last LP value $Sol_{last} \leftarrow \infty$; Branch-and-bound search tree $ST \leftarrow \emptyset$.
Generate column set $\tilde{R}$ with the CWS-H algorithm & Solve the RMP & Obtain current solution Sol_{curt}.
while $\exists \min_{r\in\tilde{R}_i^d}\{\bar{C}_{i,r}^d\} < 0 (\forall i \in I, d \in T)$ & $(Sol_{last} - Sol_{curt}) > 0$ **do**
    $Sol_{last} \leftarrow Sol_{curt}$.
    Update $\{\bar{C}_{i,r}^d\}_{r\in\tilde{R}_i^d} (\forall i \in I, d \in T)$ according to Algorithm 1 in Appendix C.
    $\tilde{R} = \tilde{R} \bigcup_{i\in I, d\in T} \tilde{R}_i^d$ & Resolve Sol_{curt}.
Add the current RMP to ST.
while $ST \neq \emptyset$ **do**
    Get the branching node with the smallest Sol_{curt} value.
    for $d \in T$ **do**
        for $i \in I$ **do**
            Generate a graph for depot i on day d: $G_i^d \leftarrow (V_i^d, A_i^d)$.
            Compute the reduced cost for $\bar{C}_{i,r}^d \leftarrow$ ESPPRC(G_i^d), where $r \in \tilde{R}_i^d$.
    while *The first iteration* || *3-SRCs are added in the last iteration* **do**
        if $\exists \min_{r\in\tilde{R}_i^d}\{\bar{C}_{i,r}^d\} < 0 (\forall i \in I, d \in T)$ **then**
            Update column set $\tilde{R} = \tilde{R} \bigcup_{i\in I, d\in T} \tilde{R}_i^d$ with negative reduced cost.
            Given $\tilde{R}$, solve the RMP & Obtain current solution Sol_{curt}.
        else
            for $d \in T$ **do**
                Separate the 3-SRCs and add the violated cuts to the RMP and resolve.
    Every five iterations, run the MIP-H algorithm & Update Sol_{best}.
    if $Sol_{curt} \leq Sol_{best}$ **then**
        if *Integer solution is found* **then**
            Update $Sol_{best} = Sol_{curt}$ & Remove the branching node from ST.
        else
            if *total flow of an arc in one day is fractional* **then**
                Branching on total arc flow & Add two branching nodes to ST.
            else if *node-visited in one graph is fractional* **then**
                Branching on node & Add two branching nodes to ST.
            else if *arc-visited in one graph is fractional* **then**
                Branching on arc & Add two branching nodes to ST.
    else
        Remove the branching node from ST.
return Sol_{best}

3. Solution approach

In this section, we present a BPC algorithm for solving the CSRP. BPC is a combinatorial optimization method for solving MIP with many variables (Costa *et al.*, 2019), where the restricted linear relaxation of the Master Problem (MP) is solved by the column generation approach at each node of the branch-and-bound search tree. In the remainder of this section, we first reformulate the compact model into an extended MP, and then discuss the solution process and two acceleration strategies for the subproblem. Later, we present the separation inequalities, warm-up procedures, and branching strategies. The overall procedure for organizing these components and implementing the BPC approach is summarized in Algorithm 1.

3.1. MP

Let $\tilde{R}$ denote the set of feasible routes. The corresponding travel time for route $r \in \tilde{R}$ is denoted as w_r, which is the sum of the travel times connecting all traversed nodes within the route. We define a binary parameter a_{ir} to indicate whether node i is visited ($a_{ir} = 1$) in route r or not ($a_{ir} = 0$). The remaining notations are defined as the same in the arc-based model. We introduce the following two decision variables:

$$z_r^d = \begin{cases} 1, & \text{if route } r \text{ is used on day } d, d \in T, r \in \tilde{R}; \\ 0, & \text{otherwise.} \end{cases}$$

u_{ij} = non-negative continuous value if patient node j is assigned to doctor node $i, i \in I, j \in J$.

Then, the formulation (1)-(16) can be reformulated as the MP:

$$\min \sum_{d \in T} \sum_{r \in \tilde{R}} w_r z_r^d \tag{17}$$

subject to

$$\sum_{r \in \tilde{R}_d} a_{jr} z_r^d \geq P_j^d, \forall d \in T, \forall j \in J \tag{18}$$

$$\sum_{r \in \tilde{R}_d} a_{ir} z_r^d \leq P_i^d, \forall d \in T, \forall i \in I \tag{19}$$

$$\sum_{r \in \tilde{R}_d} a_{ir} a_{jr} z_r^d \leq u_{ij}, \forall d \in T, \forall i \in I, \forall j \in J \tag{20}$$

$$\sum_{i \in I} u_{ij} \leq \omega, \forall j \in J \tag{21}$$

$$z_r^d \in \{0, 1\}, \forall d \in T, \forall r \in \tilde{R} \tag{22}$$

$$u_{i,j} \geq 0, \forall i \in I, \forall j \in J \tag{23}$$

The objective function in (17) minimizes the total travel time in the selected routes. Constraints (18)-(19) guarantee the availability of doctors and patients and ensure that patients with needs are visited. Constraints (20) link together routing variables and continuity variables. Constraints (21) guarantee that at most ω doctors are assigned to a patient within the time horizon. Note that, the advantages of current reformulation are two fold: On the one hand, it can process the skill-matching constraints implicitly in the following pricing subproblem; On the other hand, the continuity constraints (20)-(21) are robust and they will not destroy the subproblem structure in the iterative solving procedure, which will be further explained in Section 3.3.

3.2. Column generation scheme

Column generation follows the following principle, which iteratively solves the linear relaxation and search problems for routes (columns) that are not in the Restricted MP (RMP). We call $\overline{\text{MP}}$ the linear relaxation of MP. The solution to $\overline{\text{MP}}$ is a lower bound. We restrict $\overline{\text{MP}}$ to a subset of routes $\hat{R} \subset \tilde{R}$, namely, RMP. In the column generation step, we iteratively search for routes $r \in \tilde{R}$ with negative reduced costs that could be added to the RMP to reduce the Linear Programming (LP) objective value. If there are no routes with negative reduced costs, then the LP solution of the RMP in the last iteration is optimal and also the best solution for $\overline{\text{MP}}$; otherwise, the found routes with negative reduced costs are added to the RMP, and the iteration process continues until no improvement is found. Finally, all desired routes with negative reduced costs are added to the RMP, and we obtain the optimal solution.

3.3. Pricing subproblem

The LP solution to the RMP provides dual variables $\alpha_j^d, \alpha_i^d, \beta_{ij}^d$, and ϵ_j associated with constraints (18)-(21). In this model, α_j^d is non-negative, while α_i^d and β_{ij}^d are nonpositive. The reduced cost of route $r \in \tilde{R}^d$ on day d is:

$$\bar{C}_r^d = w_r - \sum_{j \in J} a_{jr} \alpha_j^d - \sum_{i \in I} a_{ir} \alpha_i^d - \sum_{i \in I} \sum_{j \in J} a_{ir} a_{jr} \beta_{ij}^d \tag{24}$$

If route $r \in \tilde{R}_i^d$, where $\tilde{R}_i^d$ is a column set that collects routes generated in a graph with doctor node i as the depot, then the reduced cost can be further simplified as:

$$\begin{aligned} \bar{C}_{i,r}^d &= w_r - \sum_{j \in J} a_{jr} \alpha_j^d - \alpha_i^d - \sum_{j \in J} a_{jr} \beta_{ij}^d \\ &= \sum_{(j,k) \in A, j \neq i} b_{jk}^r \bar{c}_{jk}^r - \alpha_i^d (\text{where } \bar{c}_{jk}^r = d_{jk} - \alpha_j^d - \beta_{ij}^d, \beta_{ii}^d = 0) \end{aligned} \tag{25}$$

In addition, we make a small change with respect to the notification of graph $G = (V, A)$. Previously, in Section 2, we denote the service time at node $j \in J$ as t_j and the travel time on arc $(j, k) \in A$ as t_{jk}. Now we define a new arc-based time d_{jk} on arc $(j, k) \in A$, which includes both the travel time to connect node j to k and the service time spent at node j. According to (25), we can conveniently obtain the modified cost $\bar{c}_{jk}^r$ on node j by changing the values of its outgoing arcs to generate support graph $G_i^d (i \in I, d \in T)$ with values only on arcs. The binary parameter b_{jk}^r indicates whether arc (j, k) is selected in route $r \in \tilde{R}_i^d$, which is generated through the dynamic programming method (label setting) introduced later. Note that by incorporating multiple graphs $G_i^d (\forall i \in I, d \in T)$, the dual variable β_{ij}^d in constraints (25) is subtracted from the values of its outgoing arcs (j, k) in graph G_i^d. However, ϵ_j are not included in the ingredients of reduced cost $\bar{C}_{i,r}^d$. Therefore, the continuity constraints (20)-(21) have no impact on the subproblem structure. In other words, the minimization of the reduced cost $\bar{C}_{i,r}^d$ can be transformed into the search for the shortest route considering the working time regulation and time window constraints on G_i^d. Consequently, the pricing subproblem is an NP-hard Elementary Shortest Path Problem with Resource Constraints (ESPPRC). Below, we present the label setting algorithm and focus on one support graph G_i^d.

The ESPPRC on G_i^d can be solved by the label setting algorithm. That is, at each node we define a list of labels representing information about partial paths starting from doctor (depot) node i and extending (directly/indirectly) to the current node. After initializing the labels at depot i, more nodes are iteratively added to the paths by the forward extension to their descending nodes in G_i^d through Resource Extension Functions (REFs). The feasibility of each selected node to extend is determined according to several aspects of resource constraints, including working time regulation, time windows and skill-matching requirements, thereby excluding infeasible nodes from the paths.

Now, we introduce our forward label setting algorithm for ESPPRC. We define a label $L_j = (\bar{C}_j, S_j, (N_j^v)_{v\in V})$ that represents the partial path p originating from depot node i and arriving at node $j (j \in V)$. The elements in L_j are described in detail as follows:

- $\bar{C}_j$: reduced cost of path p;
- S_j: earliest service start time at node j along path p;
- N_j^v : binary value indicating whether node v has been visited ($N_j^v = 1$) or not ($N_j^v = 0$) along path p. It is also set to 1 if node v is not visited, but is unreachable from path p. Node v is said to be unreachable if $S_j + d_{jv} > l_v^d, \max\{S_j + d_{jv}, e_v^d\} + d_{vi} > \xi$ or $K_{im} < D_{vm} (\forall m \in M)$, in which case it cannot be part of any feasible extension of path p. The above three conditions indicate that the time window constraints, the maximum working time regulation, or the skill-matching requirements are violated at least once.

The initialization of the label at depot node i sets all components to zero. The extension of label $L_j = (\bar{C}_j, S_j, (N_j^v)_{v\in V})$ along arc $(j,k) \in A$ is performed respecting the following REFs:

$$\bar{C}_k = \bar{C}_j + \bar{c}_{jk}^r \tag{26}$$

$$S_k = \max\{S_j + d_{jk}, e_k^d\} \tag{27}$$

$$N_k^v = \begin{cases} N_j^v + 1, & \text{if } v{=}k; \\ \max\{N_j^v, UR_v(S_k)\}, & \text{otherwise.} \end{cases} \tag{28}$$

Let $UR_v(S_k) = 1$ denote that node v is unreachable from label L_k, which implies that at least one of the following conditions is satisfied (we assume that the triangle inequality for travel times holds): (i) $S_k + d_{kv} > l_v^d$, (ii) $\max\{S_k + d_{kv}, e_v^d\} + d_{vi} > \xi$, or (iii) $K_{im} < D_{vm} (m \in M)$. Under conditions (i-iii), for a node to be unreachable from a label, the infeasibility check to reach node v is performed by setting $N_k^v = 1$ in vector $(N_j^v)_{v\in V}$. Moreover, we explain (28) in detail as follows:

- If $v = k$, since we extend the partial path from node j to node k, it means that node v has not yet been visited, i.e., $N_j^v = 0$, and hence, $N_k^v = N_j^v + 1 = 1$;
- If $v \neq k$, there are two cases: (i) node v has not been visited in the partial path and is reachable, namely, $N_j^v = 0$ and $UR_v(S_k) = 0, N_k^v = \max(N_j^v, UR_v(S_k)) = 0$; (ii) node v was visited in the partial path to node j or node v is unreachable from node k, namely, $N_j^v = 1$ or $UR_v(S_k) = 1, N_k^v = \max(N_j^v, UR_v(S_k)) = 1$. In both cases, we have $N_k^v = \max(N_j^v, UR_v(S_k))$.

Finally, the path is feasible only if all the following conditions hold:

$$S_j + d_{jk} \leq l_k^d \tag{29}$$

$$\max\{S_j + d_{jk}, e_k^d\} + d_{ki} \leq \xi \tag{30}$$

$$K_{im} \geq D_{km}, m \in M \tag{31}$$

$$N_k^v \leq 1, v \in V \tag{32}$$

3.4. Acceleration strategies for the pricing problem

We adopt two strategies to accelerate the solving process of the pricing problem. Even though the elementary routes in the subproblem produce tighter bounds in final results, the computational effort to obtain them also grows exponentially. Therefore, our first strategy is to relax the subproblem by allowing paths containing cycles. We use *ng-route* relaxation as a compromise between elementary and non-elementary routes to efficiently obtain the desired lower bounds. Our second acceleration strategy is to use *bidirectional search* with resource-based bounds to rapidly generate feasible routes. We present the formal formulation and implementation details for the ng-route and bidirectional search approaches in Appendix B.

3.5. Subset row inequalities

To improve the quality of the lower bound obtained by the column generation scheme, we also add valid inequalities violated by the current linear relaxation solution and consider Subset-Row Cuts (SRCs) for the CSRP. The subset-row inequalities introduced by Jepsen *et al.* (2008) for the VRPTW are special cases of Chvátal–Gomory rank-1 inequalities and were later applied with success by Desaulniers *et al.* (2008) and Zhang *et al.* (2019). They can also be adapted to our problem concerning daily decisions as follows:

$$\sum_{r\in \tilde{R}^d} \left\lfloor \frac{1}{k} \sum_{j\in S} a_{jr} \right\rfloor z_r^d \leq \left\lfloor \frac{|S|}{k} \right\rfloor, \forall S \subset J, 2 \leq k \leq |S|, d \in T \tag{33}$$

where S is a subset of patients. As described in Pecin *et al.* (2017) and Costa *et al.* (2019), we focus on the 3-SRCs defined for the subsets of three customers with parameter $k = 2$. The resultant set of inequalities is given by

$$\sum_{r\in \tilde{R}_S^d} z_r^d \leq 1, \forall S \subset J, |S| = 3, d \in T \tag{34}$$

where $\tilde{R}_S^d$ is the subset of routes $\cup_{i\in I} \tilde{R}_i^d$ and covers at least two patients in S of day d. Constraints (34) ensure that at most one route will be selected in a feasible solution where the route visits at least two of the three patients in S. Let us denote the set of all 3-SRCs (i.e., constraints (34)) as a set Θ in the MP. Furthermore, for any cut $s \in \Theta$, σ_s is the associated dual variable of constraints (34). Enlarging the support graph $G_i^d (d \in T, i \in I)$, given (25) and (34), the reduced cost of a route $r \in \tilde{R}_i^d$ after adding a 3-SRC inequality s to the MP is defined as

$$\hat{C}^d_{i,r} = \bar{C}^d_{i,r} - \sigma_s \left\lfloor \frac{1}{2} \sum_{j \in S} a_{jr} \right\rfloor \quad (35)$$

where the associated dual value $\sigma_s \leq 0$ can be interpreted as the penalty upon visiting two patients in S.

For every $s \in \Theta$, a new resource $R_s(L)$ is added to the definition of label L, which counts the number of times (mod $k = 2$) that the patients in S have been visited. The REFs of resource $R_s(L_j)$ along arc (j, j') are updated as follows:

$$R_s(L_{j'}) = \begin{cases} R_s(L_j), & \text{if } j' \notin S; \\ (R_s(L_j) + 1) \text{ mod } 2, & \text{if } j' \in S. \end{cases} \quad (36)$$

Note that once the case $R_s(L_{j'}) = 0$ occurs, the value σ_s will be subtracted from $\bar{C}_{j'}$.

The 3-SRC inequalities can be separated by full enumeration (Zhang *et al.*, 2019); that is, for every subset $S \subset J$ with three patients, we check whether the inequality is violated. Note that handling the dual variables of SRCs (34) in the pricing problem will be computationally intensive. To alleviate its impact on the overall runtime, we limit the 3-SRC usage by only generating them in the first two levels of the branching tree and adding at most Q_{max} cuts simultaneously ($Q_{max} = 50$ in our tests).

Finally, to avoid enumerating all feasible paths, given that all the REFs are nondecreasing functions, we use a dominance condition to eliminate partial paths that will not be included in the final solution.

Definition 1. *(Dominance) Let* $L^1_j = (\bar{C}^1_j, S^1_j, (N^{1,v}_j)_{v \in V}, (R^1_s)_{s \in \Theta})$ *and* $L^2_j = (\bar{C}^2_j, S^2_j, (N^{2,v}_j)_{v \in V}, (R^2_s)_{s \in \Theta})$ *represent two labels associated with different routes from depot node i to node j. Then,* L^1_j *dominates* L^2_j *if and only if* $\bar{C}^1_j - \sum_{s \in \Theta_{1,2}} \sigma_s \leq \bar{C}^2_j, S^1_j \leq S^2_j, (N^{1,v}_j)_{v \in V} \leq (N^{2,v}_j)_{v \in V}$*, and* $L^1_j \neq L^2_j$*, where* $s \in \Theta_{1,2}$ *and* $\Theta_{1,2} \subseteq \Theta$ *is the subset of 3-SRC cuts for which* $R^1_s > R^2_s (i.e., R^1_s = 1, R^2_s = 0)$.

3.6. Warm-up and upper bound

The warm-up procedure is incorporated to generate a pool of initial columns. These columns are produced using the Clark–Wright Saving Heuristic (CWS-H) algorithm, which generates a set of feasible routes that satisfy the continuity constraints to visit all nodes and provide AHH services. With the initial columns, we are able to build the initial version of $\tilde{R}$ and solve the RMP. Later, the dual variables with respect to the RMP constraints are extracted to facilitate the search for more promising columns with negative reduced costs. We further develop a tabu-column-generation algorithm motivated by Archetti *et al.* (2011). Based on the basic columns with zero reduced cost in the current RMP solution, the tabu search algorithm iteratively removes or inserts nodes on each of these routes independently to find potential routes with negative reduced costs. The process is repeated until no more routes with negative reduced costs are generated or the objective value is no longer improved. The notations and implementation details for the tabu-column-generation algorithm can be found in Appendix C.

At the end of the column generation phase, if no columns with negative reduced costs and violated 3-SRCs are found, then the value of the RMP is the value of $\overline{\text{MP}}$. We run an MIP solver on the subset of columns in the current RMP by setting the route selection variables as binary. Such a procedure, called MIP-H, is executed every five iterations and can obtain an upper bound on the integer solutions. If the MIP is difficult to optimize, then we terminate the process by limiting its running time to less than 5% of the total computational time limit. The upper bound obtained is useful to prune unpromising nodes in the branch-and-bound tree.

3.7. Branching scheme

For each node in the branch-and-bound search tree, if the decision variable in the final solution of the RMP obtained after column generation is an integer and the objective value is less than the upper bound, then we update the upper bound; if the decision variable in the final solution of the RMP is not an integer and the objective value is less than the upper bound, then branching occurs in the node; otherwise, this node is discarded from the search tree. To derive integer solutions, we enforce the following types of branching decisions in the search tree: (i) on the total flow of an arc in a day; (ii) on the node visited a fractional number of times; (iii) on the total flow of an arc in graph G^d_i. Since these branching rules are standard in the literature (Desaulniers *et al.*, 2016), we defer the complete explanation to Appendix D.

The branch-and-bound search tree is explored with a best-first search strategy. In each subtree, one node is evaluated, and only the node whose objective value is less than the current upper bound is added to the search tree for future exploration. The upper bound can be updated either by the solution determined by the MIP-H based on the current column pool or by the integer solution found in the previous exploration of the subtree. The branching process is repeated until the exploration of the search tree is completed or the time limit for the calculation is reached, resulting in a final integer solution. If the branch-price process terminates within the time limit, then the final integer solution is optimal, otherwise it is an upper bound.

4. Extension: SCSRP and adapted solution approach

In the previous sections, we present the CSRP formulation with deterministic travel and service times. However, in real-world environments, these elements can be highly uncertain due to the complex and unexpected contingencies. In this section, we investigate an SCSRP by considering uncertain travel and service times. The remainder of this section is organized as follows. After a review of previous work and the problem description in Section 4.1, we provide a chance-constrained modification of the SCSRP. Section 4.2 states the iterative process of DAM to calculate the on-time service probability. In Section 4.3, we present a DAM-label setting algorithm.

4.1. SCSRP

The existing AHH studies with uncertain durations are rather limited, and the SCSRP is not suitable for most methods, due to factors such as on-time service requirements and working time regulation (Rostami *et al.*, 2021; Naderi *et al.*, 2023). More specifically, as the patients' time windows and the doctors' maximum working time in the SCSRP are fixed, uncertain travel and service times may lead to violations of on-time service and completion. Popular stochastic optimization methods such as (distributionally) robust optimization, which were developed to analyze problems with uncertainty from a worst-case (probability distribution) perspective (Bartolini *et al.*, 2021; Zhang *et al.*, 2021), are not applicable here as their *conservative* results are not the primary concern for practitioners, particularly when the stochastic distributions can be properly estimated from historical information. Therefore, we need to seek alternative ways to handle these constraints.

A more reasonable method is to utilize Sample Average Approximation (SAA), which generates a series of discrete points to approximate different scenarios (different time nodes in our settings) and represent the uncertainty regarding time windows. Furthermore, the optimality of the SAA approach can be guaranteed with a sufficiently large sample size (Dai *et al.*, 2000; Shapiro *et al.*, 2002). This technique has been applied to address similar problems. For example, Zhan *et al.* (2021) proposed an integer L-shaped method to solve the SAA version of the AHH problem with random service durations, whereas Guo *et al.* (2021) used the SAA approach to model the stochastic operating room scheduling problem and adopted LBBD cuts to obtain solutions. Florio *et al.* (2021) developed a branch-and-price algorithm to address the VRP with stochastic demands and probabilistic duration constraints by Monte Carlo sampling and statistical inference. Motivated by these studies, we attempt to combine the SAA approach with the BPC method to solve our SCSRP.

The combined algorithm embeds the DAM approach into the previous BPC algorithm to solve the SCSRP (DAM-BPC hereafter). The DAM approach inherits the essence of the SAA approach and is adapted to generate a sequence of time-related variables. Let β $(0 \leq \beta \leq 1)$ represent a predefined service level. Therefore, a path is feasible if and only if the on-time service/completion probability to visit each patient/doctor (on the path) is no less than β. Intuitively, the larger β, the higher level of service and the more stringent are the punctuality requirements. On the other hand, β also represents the risk preference of the decision maker: if β is large, then it indicates that the decision maker is conservative in securing service quality; on the contrary, if the decision maker is not conservative in securing service quality and emphasizes objective minimization, then β is small. Specifically, the *On-Time Service Probability* (OSP) at each patient node can be achieved by satisfying the following chance constraints:

$$P(S_j(\Phi) \leq l_j^d) \geq \beta, \forall j \in V, d \in T \tag{37}$$

where $S_j(\Phi)$ is a random variable that represents the service start time on node j and Φ is the set of random scenarios. The *On-Time Completion Probability* (OCP) at each doctor node can be achieved in the same manner.

To integrate with Algorithm 1, we need to consider the new chance constraints in each pricing problem while the MP remains unchanged. Specifically, in subproblem G_i^d, the pointwise chance constraints associated with the OSP/OCP of node j are described as

$$P(S_j(\Phi) \leq l_j^d) \geq \beta, \forall j \in i \cup J \tag{38}$$

4.2. Calculate the OSP/OCP

In this section, we present the details for calculating Z discrete time points of each node, where Z is called the discretized level and can help us check whether the OSP/OCP condition is satisfied along a path. In the remaining parts, we represent such an iterative calculation process for each node as $=^{DAM}$ for simplicity.

As the arrival times at each node are uncertain values in the stochastic situation, the arrival time distributions need to be evaluated to obtain feasible SCSRP solutions. Let $F(\cdot)$ denote the Cumulative Distribution Function (CDF) of the travel time or service time. Then, we can calculate the OSP/OCP as the procedure in Figure 2 with the known $F(\cdot)$. Note that, even if the CDFs are unknown, the corresponding $F(\cdot)$ can be approximated by its empirical counterpart, which is the discrete uniform distribution induced from the real-world data. For conciseness, we defer the step-by-step explanation of Figure 2 to Appendix E and illustrate the details here with a simple example: Figure 3 describes an SCSRP instance with five patients and four doctors, which inherits the settings in Figure 1. In this figure, we use the DAM procedure to generate $Z = 3$ discrete points for the time-related variables and obtain feasible routes. In particular, Figure 3 presents the optimal route solutions of the first day after implementing the DAM-label setting algorithm with parameters $Z = 3$, $\beta = 0.65$ and a uniformly distributed $F(\cdot)$. If we set the deviation parameter $\gamma = 0.3$, then the time window range is described as $[(1-\gamma)\mu, (1+\gamma)\mu]$, where μ is the corresponding average travel time and service time; thus, the three discrete time points are chosen as $\{0.8\mu, \mu, 1.2\mu\}$. To derive the optimal routes, each step of the DAM iteration alternates between service start time points and departure time points (or the departure time points and the arrival time points), thereby generating $Z^2 = 9$ new points. We first sort them, slice them into $Z = 3$ parts, and then compute the corresponding average values in each part. Finally, we compute the optimal solution by the DAM-label setting algorithm discussed in Section 4.3, where each patient/doctor should satisfy the OSP/OCP with $\beta = 0.65$.

4.3. DAM-label setting algorithm

Similar to the label definition in Section 3, an SCSRP label is denoted by $L_j = (\bar{C}_j, S_j, (N_j^v)_{v \in V})$, where j is the last

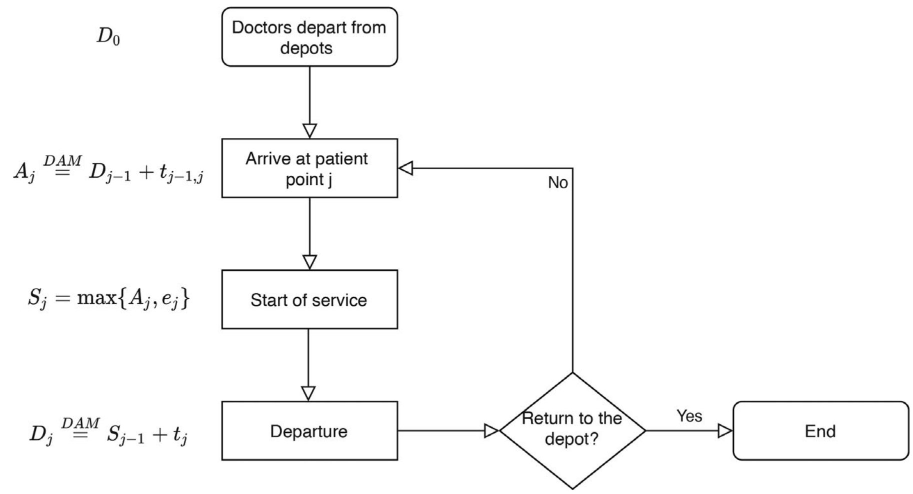

Figure 2. The iterative procedure to calculate the OSP/OCP.

visited node, $\bar{C}_j$ is the reduced cost, S_j is the service start time at patient node j and N_j^v is a binary value indicating whether node v is unreachable from the current node j. Note that S_j is a vector here including Z discrete approximation points $S_j^\tau, \tau = 1, \ldots, Z$. The extension of label $L_j = (\bar{C}_j, S_j, (N_j^v)_{v\in V})$ along arc $(j,k) \in A$ is performed respecting the following REFs:

$$\bar{C}_k = \bar{C}_j + \bar{c}_{jk}^r \tag{39}$$

$$S_k \overset{DAM}{=} \max\{S_j + d_{jk}, e_k^d\} \tag{40}$$

$$N_k^v = \begin{cases} 1 & \text{if } v = k \text{ or if } N_j^v = 1 \text{ and } v \in NG_k, \\ \max\{0, UR_v(S_k)\} & \text{otherwise.} \end{cases} \tag{41}$$

Given these REFs, (39) calculates the reduced cost of the route from patient j to k, and (40) computes the service start time distribution for patient k, where the calculation of $S_j + d_{jk}$ is based on the DAM procedure shown in Appendix E. Specifically, the DAM-label differentiates from the traditional label by inserting an additional “OSP/OCP” resource, which is checked frequently upon the moment to extend labels along arcs. Notably, when incorporating the DAM label into the label extension, we also tighten the $UR_v(S_k)$ rule. Recall that $UR_v(S_k) = 1$ indicates that node v is unreachable from label L_k. In the DAM-label setting algorithm, except for the previously mentioned feasibility rules, an additional rule is that if the probability β cannot be guaranteed under the time restriction from node k to node v, then $UR_v(S_k) = 1$.

Finally, the feasibility conditions of a path, i.e., (29)-(32), are still necessary for the DAM-label setting algorithm. However, (29)-(30) are required to be updated as follows:

$$P(S_k \le l_k^d) \ge \beta \tag{42}$$

$$P(S_k + d_{ki} \le \xi) \ge \beta \tag{43}$$

Regarding the dominance rule for this DAM-label setting algorithm, we need to compare some label elements pointwisely, which is summarised as a new *DAM-dominance* rule; that is,

Definition 2. *(DAM-dominance) Let* $L_j^1 = (\bar{C}_j^1, S_j^1, (N_j^{1,v})_{v\in V}, (R_s^1)_{s\in\Theta})$ *and* $L_j^2 = (\bar{C}_j^2, S_j^2, (N_j^{2,v})_{v\in V}, (R_s^2)_{s\in\Theta})$ *represent two SCSRP labels associated with different routes from depot node i to node j. Then, label* L_j^1 *dominates* L_j^2 *if and only if* $\bar{C}_j^1 - \sum_{s\in\Theta_{1,2}} \sigma_s \le \bar{C}_j^2, S_j^1 \le S_j^2, (N_j^{1,v})_{v\in V} \le (N_j^{2,v})_{v\in V}$, *and* $L_j^1 \ne L_j^2$, *where* $s \in \Theta_{1,2}$ *and* $\Theta_{1,2} \subseteq \Theta$ *is the subset of 3-SRC cuts for which* $R_s^1 > R_s^2$ *(i.e.,* $R_s^1 = 1, R_s^2 = 0$*). Notably,* $S_j^1 \le S_j^2$ *here means that* $S_j^{1,\tau}$ *is not greater than* $S_j^{2,\tau}$ *pointwisely for* $\tau = 1, 2, \ldots, Z$.

5. Numerical experiments

In this section, we present our computational experiments to empirically evaluate the performance of the proposed algorithms on benchmark and real-world data sets. We implement the designed algorithms in C++ and use IBM ILOG CPLEX V20.1 to solve the LP and MIP models. Our experimental environment is an AMD Ryzen Threadripper Pro 3955wx CPU. All experiments are run on a single thread. The time limit to terminate the solution process is set to 3600 seconds (denoted as T.L.).

We name the instances by the time horizon, the number of doctors, the number of patients, and the instance identifier in a group (e.g., a, b or c). For example, an instance named “7.20.67.a” refers to instance “a” with a scale of 7 days, 20 doctors, and 67 patients. For notational convenience, we also use (20, 67) to represent that there are 20 doctor nodes and 67 patient nodes in an instance. We then report the following characteristics of solutions to evaluate the computational performance: *Instances* (instance names), ω (the continuity parameter), *Opt* (the optimality indicator, where -1/0/1 represents infeasible/time-out/optimal), $t(s)$ (the computational time in seconds), *obj* (the optimal or best-known value, where “_” indicates that an upper bound with at least one integer solution cannot be derived, so does the corresponding gap), $Gap(\%)$ (the relative gap (in %) computed with respect to the upper bound and best-know values obtained in the branch-and-bound tree, (upper bound - best-known value)/upper bound). If the instance is

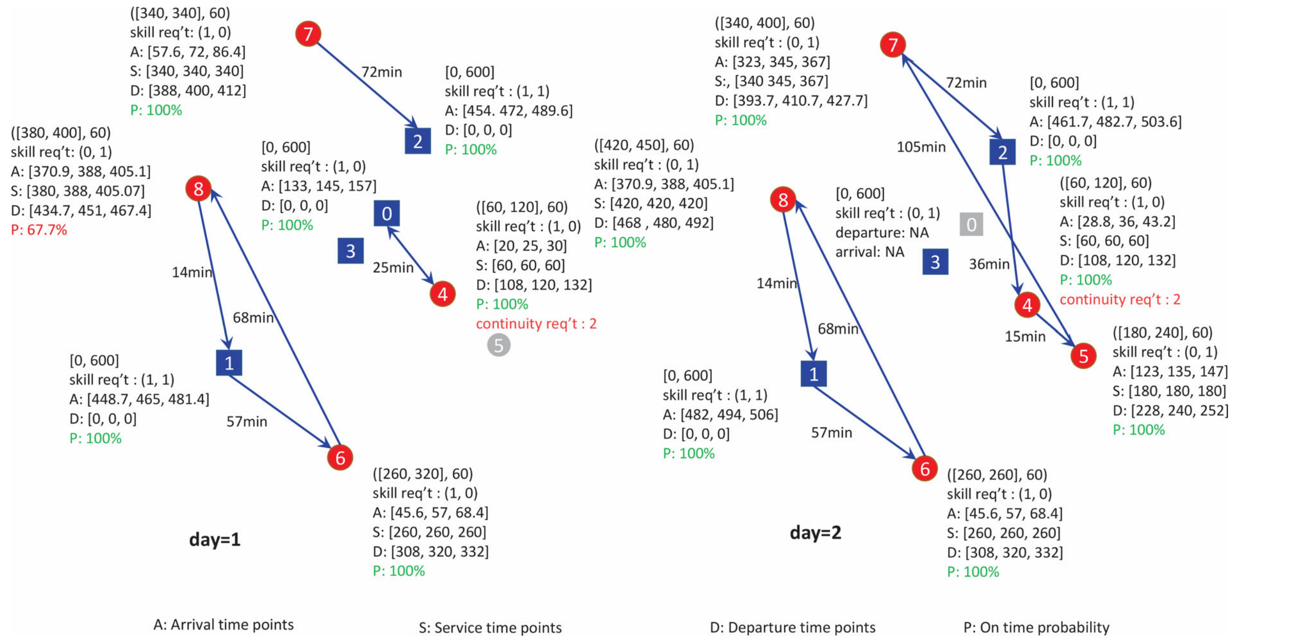

Figure 3. A simple SCSRP instance.

optimally solved, then the gap is controlled within 0.01%. *BPnodes* represents the number of nodes in the branch-and-bound search tree when the solution process terminates.

The computational tests are organized as follows. First, in Section 5.1, we test the proposed algorithm on several benchmark data sets adopted from the literature to verify its efficiency. Second, we generate a set of AHH instances originating from real operations and conduct a sensitivity analysis to evaluate the impacts of continuity constraints in Section 5.2. Comprehensive tests show that we are able to achieve high-quality solutions with considerable computational efficiency. Third, we investigate the effectiveness of our DAM-adapted solution approach to solve the SCSRP with a few benchmark instances in Section 5.3. Finally, we summarize helpful managerial insights derived from our experiments for practitioners to make better decisions in Section 5.4.

5.1. Benchmark data set tests

The proposed algorithm is tested on three benchmark data sets. As there is no data set from previous studies that are exactly the same as our settings, we make a few modifications to three related benchmark data sets in the VRP and AHH literature for computational purposes. The details of the test bed are described as follows:

1. `CordeauIns` The instances proposed by Cordeau *et al.* (1997) describe a set of MDVRPs within one day and have been widely used. We adopt 20 instances with the scale of $8 \sim 30$ doctor nodes and $48 \sim 288$ patient nodes into our test bed. Since skill-matching and continuity constraints are absent in these instances, we relax the corresponding constraints in the CSRP formulation by setting the doctor proficiency level to meet all patient requirements and the continuity parameter ω to infinity. Moreover, we extend the capacity constraints in our pricing problem so that we can handle the capacity constraints here.
2. `LiuIns` The instances proposed by Liu *et al.* (2019) describe a single-day home-caregiver scheduling and routing problem. Given the different combinations of doctor nodes and patient nodes, we consider 10 instances for each doctor–patient node combination, i.e., $(5, 30), (7, 40)$ and (9, 50) nodes, in our test bed. We reserve the availability of each node when considering the time horizon of a week; since the patterns of doctors and patients are the same every day, it is reasonable to define $\omega = 1$. The default configurations of the skill-matching constraints for these instances are the same as those in `CordeauIns`.
3. `HechingIns` The instances proposed by Heching *et al.* (2019) describe a multi-day healthcare aide scheduling and home visit problem over a given time horizon. The skill-matching constraints are relaxed, as previously mentioned. Moreover, as multi-day operational problems, instances with the same scales of nodes are different in terms of the continuity parameter ω and the availability pattern of doctors and patients. Within a week, Heching *et al.* (2019) defined only the required number of visits to each patient, but the detailed visiting pattern is a decision. However, such visiting patient patterns are pre-determined parameters in the CSRP, so we randomly generate the visiting patterns for all patients. For each $\omega = 1, .., 7$, we extract 10 instances with (20,40) and (20,50) nodes. Preliminary computational results eliminate the instances with $\omega = 1, 2$ (infeasible) and $\omega = 6, 7$ (objectives remain unchanged compared with $\omega = 5$).

For conciseness, we summarize the benchmark test bed information in Table 2 and all experimental results are presented in Tables 1, 2, and 3 in Appendix I. After analyzing the numerical results, we have the following observations:

1. The size of the problem is determined by three factors: the time horizon, the number of doctors, and the number of patients. Increasing any of these factors will slow down the solving procedure because the dynamic programming method used to solve the subproblem requires more iterations to identify beneficial routes.
2. The computational tests show that our designed method can efficiently solve most of the benchmark instances in the test bed to proven optimality. Even for unsolved instances due to time limitations, the average gap is less than 0.5%. Specifically, all `LiuIns` instances are solved to optimality within an average of 13 seconds. In addition, more than 80.0% `CordeauIns` instances are solved to optimality within the time limit, and the average exit gap is less than 0.43% for the unsolved instances. With regards to `HechingIns` instances, our approach achieves acceptable performance by successfully solving 28.3% of the instances to optimality, and the average exit gap is less than 0.61% for the unsolved instances.
3. Our designed algorithm clearly outperforms the commercial solver CPLEX. As indicated in Tables 1, 2 and 3 in Appendix I, with CPLEX, 96.6% `LiuIns` instances are solved to optimality within 261 seconds and 14464 *BPnodes* on average; Furthermore, CPLEX only solves one `Coredeaux` instance and cannot obtain feasible upper bounds for the rest of the instances and all `HechingIns` instances within the time limit. Obviously, the BPC algorithm has a better performance and is superior to CPLEX in terms of both the computational time and the number of solved instances. In addition, for a fair comparison between the BPC method and other existing methods such as BP, L-Shaped and LBBD algorithms that are widely used in healthcare optimization studies (summarized in Table 1), we also conduct additional comparative analysis in Appendix F. The performance comparison further strengthens the BPC advantages and provides more evidence to the practitioners to consider this method.
4. As the problem instance scales increase, the average computing times to solve one instance and one node in the branch-and-bound tree also increase. For example, when given the problem scales 7.20.40 and 7.20.50 in the `HechingIns` test bed, the average computing times are 2189 and 3434 seconds for each category/scale, and the average computing times to solve one node in the branch-and-bound tree are 0.56 and 1.13 seconds, respectively. Similar phenomena are observed in the `CordeauIns` and `LiuIns` tests.

Table 2. The benchmark instances description for our test bed.

Data set	*T*	$\lvert I\rvert$	$\lvert J\rvert$	ω	*Number of Instances*
`CordeauIns`	1	$8\sim30$	$48\sim288$	1	20
`LiuIns`	7	5	30	1	10
	7	7	40	1	10
	7	9	50	1	10
`HechingIns`	7	20	40	$3\sim5$	3×10
	7	20	50	$3\sim5$	3×10

Table 3. The CSRP results of real-world instances.

		BPC					*CPLEX*				
Instances	ω	*Opt*	*t(s)*	*obj*	*Gap(%)*	*BPnodes*	*Opt*	*t(s)*	*obj*	*Gap(%)*	*BP nodes*
7.20.67.a	2	0	T.L.	3767	0.08	3715	0	T.L.	–	–	71,678
7.20.67.b	2	1	93	3978	0.00	283	0	T.L.	–	–	60,335
7.20.67.c	2	1	700	4007	0.00	2369	0	T.L.	–	–	45,906
7.20.67.d	2	0	T.L.	4133	0.21	2661	1	959	4133	0.00	13,453
7.20.67.e	2	0	T.L.	4157	0.34	1891	0	T.L.	–	–	62,899
7.20.67.f	2	0	T.L.	4406	0.05	2699	0	T.L.	–	–	60,313
7.20.67.g	2	0	T.L.	3870	0.66	1845	0	T.L.	–	–	27,804
7.20.67.h	2	0	T.L.	3475	0.63	1265	0	T.L.	–	–	77,722
7.20.67.i	2	0	T.L.	3844	0.23	1823	0	T.L.	–	–	30,043
7.20.67.j	2	0	T.L.	4200	0.07	2509	1	1307	4200	0.00	10,545
7.20.67.a	3	1	1184	3764	0.00	4055	0	T.L.	–	–	125,973
7.20.67.b	3	1	57	3970	0.00	187	0	T.L.	–	–	115,917
7.20.67.c	3	1	690	4001	0.00	2967	1	912	4001	0.00	36,213
7.20.67.d	3	1	430	4123	0.00	1797	0	T.L.	–	–	68,962
7.20.67.e	3	1	290	4146	0.00	1095	0	T.L.	–	–	75,356
7.20.67.f	3	1	181	4392	0.00	687	0	T.L.	–	–	125,970
7.20.67.g	3	0	T.L.	3857	0.39	2579	1	2742	3857	0.00	59,814
7.20.67.h	3	0	T.L.	3464	0.35	2345	0	T.L.	–	–	117,025
7.20.67.i	3	1	1207	3831	0.00	4471	1	423	3831	0.00	11,406
7.20.67.j	3	1	309	4195	0.00	1275	0	T.L.	–	–	79,121
7.20.67.a	4	0	T.L.	3767	0.15	2597	0	T.L.	–	–	159,235
7.20.67.b	4	1	122	3970	0.00	425	0	T.L.	–	–	139,451
7.20.67.c	4	1	688	4001	0.00	3117	1	487	4001	0.00	23,054
7.20.67.d	4	1	162	4122	0.00	687	1	304	4122	0.00	10,516
7.20.67.e	4	1	415	4146	0.00	1687	0	T.L.	–	–	149,567
7.20.67.f	4	1	173	4392	0.00	635	0	T.L.	–	–	161,846
7.20.67.g	4	0	T.L.	3857	0.38	2659	0	T.L.	–	–	46,043
7.20.67.h	4	1	1671	3463	0.00	5319	1	3545	3463	0.00	92,285
7.20.67.i	4	1	1114	3831	0.00	4145	1	461	3831	0.00	11,641
7.20.67.j	4	1	47	4189	0.00	165	1	299	4189	0.00	10,268

5.2. Real-world data set tests

To show the effectiveness of our approach in solving practical problems, we obtained 300 AHH visiting appointment data points from November 1, 2019, to November 6, 2019. In this data set, the nodes represent doctors and patients geographically distributed within all areas in Beijing. The distance matrix is acquired with the longitudes and latitudes of nodes from Amap (https://developer.amap.com), i.e., the shortest travel time by taking a combination of metro and bus between any two nodes (the connection between any two nodes exists and the travel times satisfy the triangular inequality conditions). The planning horizon is set as a week and the time window, service time and visiting pattern for each patient are pre-determined within the time horizon.

To investigate the effect of continuity constraints on the computational efficiency and solution quality in practical environments, we conduct a sensitivity analysis to evaluate the impacts of ω. For each $\omega = 1$–4, we generate 10 instances of 20 doctors and 67 patients, which are randomly selected with respect to the five most common medical services to construct a real graph. Approximately 75% of the doctors and 45% of the patients are available per day, and the total number of requests per patient is less than four within the planning horizon. The compatibility set of patients that each doctor is capable to provide services to is obtained through data preprocessing considering skill-matching requirements. We report the overall information of the real-world instances and solution results in Table 3 and analyze the computational performance as follows:

1. The numerical results show that our designed algorithm clearly outperforms the commercial solver CPLEX with respect to the computational time and number of solved instances. The former solves 60% instances to optimality and can obtain a feasible solution on average within 1758 seconds and 0.12% gap, whereas the commercial solver only solves very few real-world instances to optimality within the time limit and even cannot obtain a feasible solution for the rest instances.
2. Regarding the impact of the continuity constraints on the computational performance, the problem becomes more challenging as the continuity parameter ω decreases and all instances become infeasible when $\omega = 1$ (thus being removed from Table 3). This is intuitive since the solution space is suppressed by the stricter continuity constraints, and more iterations are required to solve the subproblem for beneficial routes generation. In other words, the scheduling flexibility increases in ω. However, if we relax the continuity constraints by increasing ω, then it can have a negative impact on patient satisfaction as large ω values will lead to poor service experience. As a result, the continuity parameter ω should be calibrated to reach a balance between scheduling flexibility and patient satisfaction.
3. The overall computational results indicate that the BPC approach is capable of handling real-world operational route scheduling requirements. In fact, our algorithm outperforms the scheduling method currently being used by the company (i.e., mainly manual scheduling outcomes by experience), the average travel time per visit is reduced by 30.6 minutes (from 113.9 minutes to 83.3 minutes) and the average daily visits increase from 2.76 to 3.35, resulting in a $\approx 20\%$ improvement in service capability. Moreover, patient satisfaction is improved because the company invested efforts to assign familiar doctors to serve the same patients. Figure 4 describes a routing and scheduling example for one specific day solution by solving the instance "7.20.67.a" ($\omega = 3$), and we defer the rest scheduling figures to Appendix H. The red and blue icons represent doctors and patients respectively, and the lines connecting the nodes represent the scheduling and movement paths. Note that the unvisited nodes are not available on that day but can be visited in other days. Compared with the original scheme in the collaborating company, the new scheduling results are fairly acceptable by eliminating cost-expensive routes, such as situations in which doctors have to travel across the city to provide service.

5.3. Uncertain data set tests

To test the DAM-BPC method and measure the modified approach in solving the SCSRP, we further conduct sensitivity analysis on `LiuIns` instances to investigate the effects of discretized-level Z and OSP/OCP parameter β. Since the discretized-level (or the number of points) Z is crucial to control uncertainty and reduce accidents risks, we start with experiments on different configurations of $Z \in \{5, 10, 20\}$ given a baseline parameter $\beta = 0.75$. The computational results are reported in Table 4. In addition, the OSP/OCP parameter β reflects the service quality of providers, so we also present the results for different configurations of β from 0.55 to 0.95 given $Z = 10$ in Table 4 (and also Tables 4-5 in Appendix I). For each parameter configuration, the travel and service times are assumed to follow uniform distributions and the deviation parameter is set as $\gamma = 0.3$, which corresponds to the uniform distribution of travel time used by Adulyasak and Jaillet (2016). To verify the generality and applicability of the DAM-BPC method, we further conduct the following experiments: (i) On the one hand, apart from the uniform distribution, we also test several other distributions that were investigated in the literature, including the asymmetric triangular distributions and two-point distributions (Zhang *et al.*, 2021). The results and insights based on these distributions are similar to those obtained from the baseline uniform distributions, so we report their details in Tables 6-9 in Appendix I. (ii) On the other hand, to evaluate the out-of-sample performance of the proposed method, we also conduct out-of-sample tests for all the above parameter configurations and summarize the results and analysis in Appendix G. The testing results indicate that our method shows strong adaptivity and robustness to various parameter configurations as the insights derived from the out-of-sample tests do not significantly deviate from those obtained from the previous sensitivity analysis.

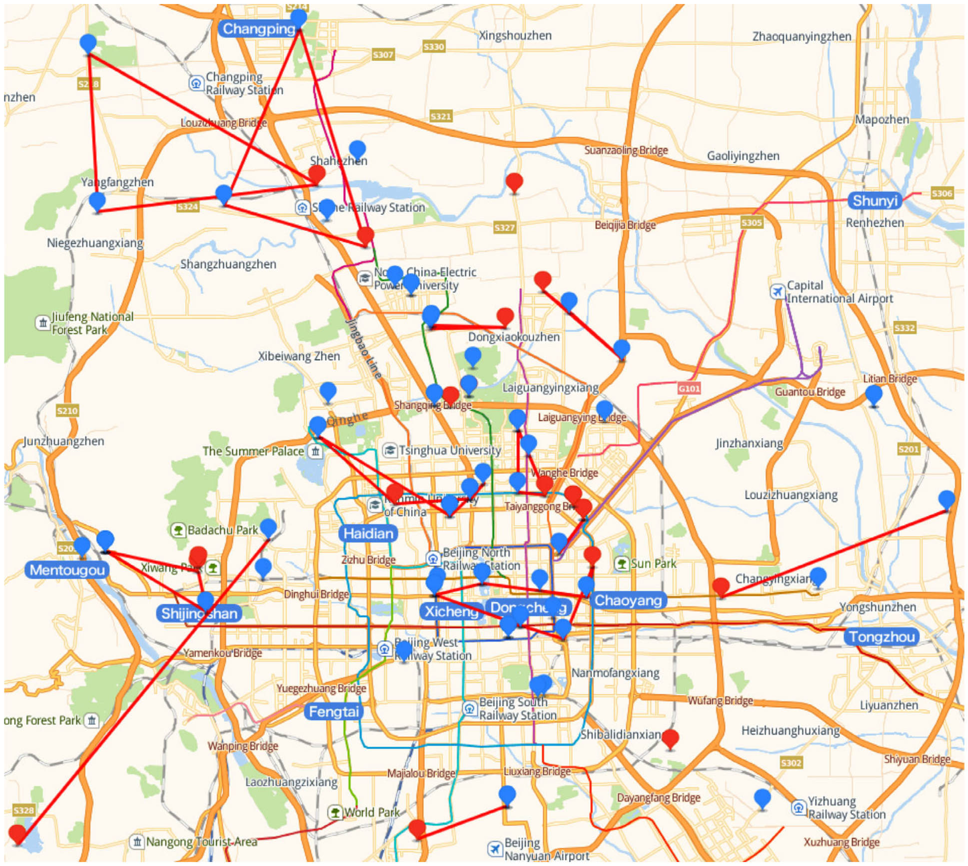

Figure 4. The optimal 1-day routing solution for the Pinetree instance "7.20.67.a" ($\omega = 3$).

With numerical experiments on SCRSP instances, we have the following observations. To address the SCSRP, the adoption of discrete points within the DAM procedure renders the problem rather difficult to solve, since it requires additional computational efforts to verify whether an extension is feasible. Notably, from $Z = 5$ to $Z = 20$, the number of infeasible instances increases and the proportion of feasible instances that can be solved to optimality decreases from 100.0% to 85.7%. Moreover, the average computing times of instances with optimal solutions are 65, 200 and 573 seconds for $Z = 5$, 10, 20 scenarios, and the average exit gaps for unsolved instances are 0.00%, 0.00% and 1.28%, respectively. Clearly, the SCSRP takes much longer time to solve than the CSRP counterpart as the algorithmic complexity increases by at least Z^2 folds, which also explains why larger Z values yield solutions of larger exit gaps. Note that, a certain number of instances might turn into infeasible as Z increases because the service starting time evaluated via DAM underestimates the mean value of the uncertain service starting time according to Jensen's inequality (Zhang *et al.*, 2021). However, we also found that the effects of tackling uncertainties by increasing Z are indeed diminishing; that is, when Z increases to a certain value, the density of the discrete time points renders the model robust enough, after which increasing Z will not improve the final solution significantly or even become worse due to the over-fitting effect. As for the impact of OSP/OCP parameter β on computational performances, from $\beta = 0.55$ to $\beta = 0.95$, the average objective values are increasing in β (i.e., from 838.36 to 866.71). Such an upward trend indicates that a higher on-time service/-completion requirement would inevitably lead to additional costs, which poses challenges for managers to find a balance between service quality and cost-efficiency.

In summary, through the previous numerical experiments, we conclude that our approaches are efficient in solving the CSRP and SCSRP and also lead to the following managerial insights.

5.4. Managerial insights

In this section, we investigate the effects of design parameters on company performance. The preceding sensitivity analysis provides decision-makers with some helpful managerial insights:

1. *Scale configuration*: We believe that careful consideration should be given to scale configuration, which is a fundamental aspect affecting the solution procedure that involves planning horizon selection and graph structure selection (the cardinality of doctors and patients). On the one hand, a shorter planning horizon, i.e., weeks, is preferred for scheduling and circulation. On the other hand, a graph structure with no more than 30 doctors and 100 patients is recommended; for which, the problem can be solved efficiently. For even larger instances, an appropriate graph partition method is thus closely required to guarantee the solutions' quality. To this end, the main functional and topological constraints in the

Table 4. The SCSRP results of instances `LiuIns`.

		$Z=5$					$Z=10$					$Z=20$				
Instances	ω	*Opt*	*t(s)*	*obj*	*Gap(%)*	*BPnodes*	*Opt*	*t(s)*	*obj*	*Gap(%)*	*BPnodes*	*Opt*	*t(s)*	*obj*	*Gap(%)*	*BPnodes*
7.5.30.a	1	1	8	733	0.00	7	1	21	738	0.00	9	1	110	733	0.00	5
7.5.30.b	1	1	3	801	0.00	1	1	11	801	0.00	1	1	64	801	0.00	1
7.5.30.c	1	1	4	689	0.00	1	1	13	689	0.00	1	1	62	689	0.00	1
7.5.30.d	1	1	5	953	0.00	1	−1	–	–	–	–	−1	–	–	–	–
7.5.30.e	1	1	21	681	0.00	35	1	38	677	0.00	9	1	225	679	0.00	19
7.5.30.f	1	−1	–	–	–	–	−1	–	–	–	–	−1	–	–	–	–
7.5.30.g	1	1	4	665	0.00	1	1	9	665	0.00	1	1	64	665	0.00	1
7.5.30.h	1	1	16	824	0.00	15	1	40	821	0.00	9	1	232	828	0.00	9
7.5.30.i	1	1	5	786	0.00	1	1	13	786	0.00	1	1	58	786	0.00	1
7.5.30.j	1	1	9	830	0.00	13	1	12	830	0.00	1	1	108	830	0.00	7
7.7.40.a	1	1	25	954	0.00	11	1	93	954	0.00	13	1	394	954	0.00	9
7.7.40.b	1	1	118	814	0.00	131	1	408	815	0.00	145	0	T.L.	835	2.95	293
7.7.40.c	1	1	47	810	0.00	25	1	86	795	0.00	3	1	380	795	0.00	3
7.7.40.d	1	1	22	755	0.00	9	1	73	759	0.00	11	1	407	759	0.00	11
7.7.40.e	1	1	46	854	0.00	29	1	132	854	0.00	25	1	699	854	0.00	29
7.7.40.f	1	1	40	812	0.00	27	1	82	816	0.00	13	1	457	816	0.00	15
7.7.40.g	1	1	13	833	0.00	3	1	40	833	0.00	3	1	227	833	0.00	3
7.7.40.h	1	1	32	806	0.00	21	1	67	803	0.00	11	1	298	803	0.00	7
7.7.40.i	1	1	212	856	0.00	183	1	164	858	0.00	19	0	T.L.	866	0.07	147
7.7.40.j	1	1	55	940	0.00	35	1	212	940	0.00	39	1	1047	942	0.00	39
7.9.50.a	1	1	82	939	0.00	15	1	264	938	0.00	15	1	1079	935	0.00	9
7.9.50.b	1	1	42	955	0.00	9	1	264	963	0.00	27	1	868	955	0.00	11
7.9.50.c	1	1	120	939	0.00	43	1	560	941	0.00	77	1	2837	939	0.00	53
7.9.50.d	1	1	45	940	0.00	11	1	88	935	0.00	5	1	490	935	0.00	5
7.9.50.e	1	1	21	905	0.00	3	1	86	905	0.00	3	1	347	905	0.00	3
7.9.50.f	1	1	249	935	0.00	163	1	305	933	0.00	39	0	T.L.	936	0.19	99
7.9.50.g	1	1	85	885	0.00	19	1	287	885	0.00	19	1	1337	885	0.00	17
7.9.50.h	1	1	39	836	0.00	5	1	627	857	0.00	93	1	713	836	0.00	5
7.9.50.i	1	1	437	1019	0.00	113	1	1408	1019	0.00	121	0	T.L.	1035	1.90	47
7.9.50.j	1	1	81	883	0.00	13	1	211	888	0.00	11	1	1250	883	0.00	13

districting problem, i.e., integrality, balancing, compactness and contiguity should be ensured. We refer interested readers to Mostafayi Darmian *et al.* (2021) for districting and partition details.

2. *Impact of care continuity*: The configuration of continuity constraints plays an important role in the optimization process. Table 3 shows that instances with larger ω values tend to be solved faster due to higher scheduling flexibility. However, the significantly higher ω values might result in low patient satisfaction, workload imbalance in doctors and more cost components such as additional coordination efforts and information systems infrastructure for the system. If the continuity restriction is calibrated appropriately, then it is possible to facilitate the company in scheduling doctors with more flexible plans without incurring downside effects. Therefore, determining a decent ω is critical to making the trade-off between resolving the CSRP efficiently and searching for high-quality solutions. In addition, motivated by the EMS for ambulances usage constraints, we also investigate alternative service level requirement constraints by imposing at least a prescribed percentage of patients receiving service with the cohort size of doctors being less than ω, which renders more managerial flexibility to implement the continuity restriction in practical AHH business.
3. *Impact of on-time service/completion requirement*: Our computational experiments show that the discretized level Z affects performance in a complicated way. The inclusion of uncertain elements degrades both the objective value and computational time. Consequently, the objective value (e.g., average cost) increases when pursuing a more robust solution. Therefore, choosing an appropriate number of discrete points is important for trading off robustness versus the corresponding cost. Values of Z that are too large, resulting in large resource costs, or too small, leading to lower robustness, should be avoided, and the value needs to be decided according to practical circumstances. For example, in our experiments, $Z=10$ is a suitable value, while the selection of β in a proper range is trivial.
4. *Labor-force management*: To improve service quality, the company should employ additional doctors according to their geographical addresses and skills, which will improve labor management and scheduling flexibility. If any particular region is resource-limited and patients in the region need to be served by doctors from remote places, then a newly employed doctor in that region will significantly reduce labor costs. Consequently, when rescheduling the new timetable, sufficient doctors are critical to producing better schedules for the next circulated horizon.
5. *Generalizable applications*: The proposed model can be successfully generalized to address similar problems under different conditions. As previously indicated, the benchmark problems from the literature are specific variants of our models. Numerical experiments show that our approaches can also obtain high-quality solutions with reasonable computational resources.
6. *Note to the practitioners*: The CSRP/SCSRP is dedicated to determining optimal schedules with certificated, contracted and full-time doctors to provide primary

healthcare services for long-term patients, whose basic information including medical conditions and care requirements (e.g., type, service dates, time windows) are obtained by attended home assessment services in advance, considering the aforementioned features and respecting the continuity restrictions. It is noted that real-time scheduling, such as assessment services or (unexpectedly) short-term patient visits, is necessary but beyond the scope of the current study. Finally, the time horizon is typically framed as a week since the service patterns of patients repeat across weeks (usually lasting for 3–6 months).

6. Conclusion

In this article, we address a complex routing and scheduling problem with skill-matching and continuity constraints (CSRP) in the AHH industry. We aim to minimize the total travel time while considering a series of comprehensive and challenging constraints extracted from practical environments. This problem is modeled as a HSDConVRPTW. To the best of our knowledge, while similar problems have been investigated in the literature, there is no clear evidence of a compact model that can characterize all the practical features in this article and any exact approaches for solving the HSDConVRPTW that is practically implementable. Motivated by this gap, we first propose an arc-based compact formulation that can adequately characterize the aforementioned constraints. The compact model is intractable and cannot be solved optimally for large-scale instances with commercial solvers. Therefore, we consider a route-based reformulation and develop a BPC approach that speeds up the solving efficiency and provides high-quality solutions in a reasonable time. In addition, we have extended our CSRP to a stochastic variant SCSRP by taking into account uncertainty on travel and service times in practical settings, and also adapted the BPC to DAM-BPC, which can handle the uncertainty in an elegant way. Finally, extensive numerical experiments on benchmark and real-world data sets show that our algorithms are efficient in solving the CSRP and SCSRP and clearly outperform both the commercial solver CPLEX and state-of-art solution approaches in literature. In summary, this article not only contributes to the literature in this stream of research but also helps AHH practitioners to make better decisions and reduce labor costs for enterprises. According to the feedback from the collaborating company, the implementation of our approach significantly improves the service capability and patient satisfaction in the company by increasing the average daily visits to 20%. In addition to the economic benefits, our approach also reduces reliance on scheduler experience and develops professional standards for operational-level decision-making. We believe that our algorithm can be a widely accepted alternative to replace manual scheduling and routing operations in many companies facing similar performance improvement targets.

Our research also has a few limitations. When considering uncertain travel and service time, we impose an assumption that the distribution of each uncertain duration is known and then apply the DAM approach to handle the SCSRP. Therefore, one of the potential venues for future studies is to address the SCSRP with unknown distributions under the umbrella of distributionally robust optimization, which is related to characterizing uncertain travel and service times with appropriate ambiguity sets and solving the problem rigorously, and contributes significantly to the stochastic vehicle routing problem area (Zhang *et al.*, 2021). Another future venue is to consider more complicated requirements, for example, synchronized visits among several doctors or overtime penalty charged induced by uncertainty (Malagodi *et al.*, 2021) and general precedence relationships among patients (Sarin *et al.*, 2014).

Acknowledgments

The authors are grateful to the editor, the associate editor, and two anonymous referees for their valuable comments which have resulted in significant improvements in this article.

Funding

This work was supported by the National Natural Science Foundation of China (No. 71872093), National Social Science Fund of China (No. 21&ZD128) and the Durham University Business School Primary Research Funding.

Notes on contributors

Mingda Liu is currently pursuing his PhD degree in the Department of Industrial Engineering, Tsinghua University, China. His research interests include modeling and optimization in healthcare operations management, and exact decomposition algorithm.

Yanlu Zhao is an assistant professor in Operations Management at Durham University Business School. His research and teaching focus on data-driven service operations and supply chain management.

Xiaolei Xie is an associate professor in the Department of Industrial Engineering, Tsinghua University, China. He received his PhD degree from the Department of Industrial and Systems Engineering, University of Wisconsin, Madison. His research interests focus on stochastic modeling, optimization, and advanced data analytics in healthcare systems.

ORCID

Mingda Liu http://orcid.org/0000-0002-0015-0848
Yanlu Zhao http://orcid.org/0000-0003-0861-3816
Xiaolei Xie http://orcid.org/0000-0002-5133-9712

Data availability statement

The data that support the findings of this study are openly available at https://doi.org/10.5281/zenodo.7928388.

References

Adulyasak, Y. and Jaillet, P. (2016) Models and algorithms for stochastic and robust vehicle routing with deadlines. *Transportation Science*, **50**(2), 608–626.

Archetti, C., Bouchard, M. and Desaulniers, G. (2011) Enhanced branch and price and cut for vehicle routing with split deliveries and time windows. *Transportation Science*, **45**(3), 285–298.

Baldacci, R., Bartolini, E., Mingozzi, A. and Roberti, R. (2010) An exact solution framework for a broad class of vehicle routing problems. *Computational Management Science*, **7**(3), 229–268.

Bard, J.F., Shao, Y., Qi, X. and Jarrah, A.I. (2014) The traveling therapist scheduling problem. *IIE Transactions*, 46(7), 683–706.

Bartolini, E., Goeke, D., Schneider, M. and Ye, M. (2021) The robust traveling salesman problem with time windows under knapsack-constrained travel time uncertainty. *Transportation Science*, **55**(2), 371–394.

Beliën, J., De Boeck, L. and Van Ackere, J. (2014) Municipal solid waste collection and management problems: A literature review. *Transportation Science*, **48**(1), 78–102.

Bennett, A.R. and Erera, A.L. (2011) Dynamic periodic fixed appointment scheduling for home health. *IIE Transactions on Healthcare Systems Engineering*, **1**(1), 6–19.

Cappanera, P., Gouveia, L. and Scutellà, M.G. (2011) The skill vehicle routing problem, in *Network Optimization*, Springer, Berlin, Heidelberg, pp. 354–364.

Cappanera, P. and Scutellà, M.G. (2015) Joint assignment, scheduling, and routing models to home care optimization: A pattern-based approach. *Transportation Science*, **49**(4), 830–852.

Cappanera, P., Scutellà, M.G., Nervi, F. and Galli, L. (2018) Demand uncertainty in robust home care optimization. *Omega*, **80**, 95–110.

Cordeau, J.-F., Gendreau, M. and Laporte, G. (1997) A tabu search heuristic for periodic and multi-depot vehicle routing problems. *Networks: An International Journal*, **30**(2), 105–119.

Costa, L., Contardo, C. and Desaulniers, G. (2019) Exact branch-price-and-cut algorithms for vehicle routing. *Transportation Science*, **53**(4), 946–985.

Dai, L., Chen, C. and Birge, J. (2000) Convergence properties of two-stage stochastic programming. *Journal of Optimization Theory and Applications*, **106**(3), 489–509.

Demirbilek, M., Branke, J. and Strauss, A.K. (2019) Home healthcare routing and scheduling of multiple nurses in a dynamic environment. *Flexible Services and Manufacturing Journal*, **33**(1), 253–280.

Desaulniers, G., Errico, F., Irnich, S. and Schneider, M. (2016) Exact algorithms for electric vehicle-routing problems with time windows. *Operations Research*, **64**(6), 1388–1405.

Desaulniers, G., Lessard, F. and Hadjar, A. (2008) Tabu search, partial elementarity, and generalized k-path inequalities for the vehicle routing problem with time windows. *Transportation Science*, **42**(3), 387–404.

Fikar, C. and Hirsch, P. (2017) Home health care routing and scheduling: A review. *Computers & Operations Research*, **77**, 86–95.

Florio, A.M., Hartl, R.F., Minner, S. and Salazar-González J.-J. (2021) A branch-and-price algorithm for the vehicle routing problem with stochastic demands and probabilistic duration constraints. *Transportation Science*, **55**(1), 122–138.

Gamst, M. and Jensen, T.S. (2012) A branch-and-price algorithm for the long-term home care scheduling problem, in *Operations Research Proceedings 2011*, Springer, Berlin, Heidelberg, pp. 483–488.

Goeke, D., Roberti, R. and Schneider, M. (2019) Exact and heuristic solution of the consistent vehicle-routing problem. *Transportation Science*, **53**(4), 1023–1042.

Grenouilleau, F., Lahrichi, N. and Rousseau, L.-M. (2020) New decomposition methods for home care scheduling with predefined visits. *Computers & Operations Research*, **115**, 104855.

Grenouilleau, F., Legrain, A., Lahrichi, N. and Rousseau, L.-M. (2019) A set partitioning heuristic for the home health care routing and scheduling problem. *European Journal of Operational Research*, **275**(1), 295–303.

Guo, C., Bodur, M., Aleman, D.M. and Urbach, D.R. (2021) Logic-based Benders decomposition and binary decision diagram based approaches for stochastic distributed operating room scheduling. *INFORMS Journal on Computing*, **33**(4), 1551–1569.

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

Hashemi Doulabi, H., Pesant, G. and Rousseau, L.-M. (2020) Vehicle routing problems with synchronized visits and stochastic travel and service times: Applications in healthcare. *Transportation Science*, **54**(4), 1053–1072.

Heching, A., Hooker, J.N. and Kimura, R. (2019) A logic-based Benders approach to home healthcare delivery. *Transportation Science*, **53**(2), 510–522.

Jepsen, M., Petersen, B., Spoorendonk, S. and Pisinger, D. (2008) Subset-row inequalities applied to the vehicle-routing problem with time windows. *Operations Research*, **56**(2), 497–511.

Kovacs, A.A., Golden, B.L., Hartl, R.F. and Parragh, S.N. (2015a) Vehicle routing problems in which consistency considerations are important: A survey. *Networks*, **64**(3), 192–213.

Kovacs, A.A., Golden, B.L., Hartl, R.F. and Parragh, S.N. (2015b) The generalized consistent vehicle routing problem. *Transportation Science*, **49**(4), 796–816.

Lian, K., Milburn, A.B. and Rardin, R.L. (2016). An improved multi-directional local search algorithm for the multi-objective consistent vehicle routing problem. *IIE Transactions*, **48**(10), 975–992.

Liu, R., Yuan, B. and Jiang, Z. (2019) A branch-and-price algorithm for the home-caregiver scheduling and routing problem with stochastic travel and service times. *Flexible Services and Manufacturing Journal*, **31**(4), 989–1011.

Malagodi, L., Lanzarone, E. and Matta, A. (2021) Home care vehicle routing problem with chargeable overtime and strict and soft preference matching. *Health Care Management Science*, **24**(1), 140–159.

Mosquera, F., Smet, P. and Vanden Berghe, G. (2019) Flexible home care scheduling. *Omega*, **83**, 80–95.

Mostafayi Darmian, S., Fattahi, M. and Keyvanshokooh, E. (2021) An optimization-based approach for the healthcare districting under uncertainty. *Computers & Operations Research*, **135**, 105425.

Munari, P., Moreno, A., De La Vega, J., Alem, D., Gondzio, J. and Morabito, R. (2019) The robust vehicle routing problem with time windows: Compact formulation and branch-price-and-cut method. *Transportation Science*, **53**(4), 1043–1066.

Naderi, B., Begen, M.A., Zaric, G.S. and Roshanaei, V. (2023) A novel and efficient exact technique for integrated staffing, assignment, routing, and scheduling of home care services under uncertainty. *Omega*, **116**, 102805.

Naderi, B., Roshanaei, V., Begen, M.A., Aleman, D.M. and Urbach, D.R. (2021) Increased surgical capacity without additional resources: Generalized operating room planning and scheduling. *Production and Operations Management*, **30**(8), 2608–2635.

Park, J. and Kim, B.-I. (2010) The school bus routing problem: A review. *European Journal of Operational Research*, **202**(2), 311–319.

Pecin, D., Contardo, C., Desaulniers, G. and Uchoa, E. (2017) New enhancements for the exact solution of the vehicle routing problem with time windows. *INFORMS Journal on Computing*, **29**(3), 489–502.

Rabbouch, B., Mraihi, R. and Saâdaoui, F. (2018) A recent brief survey for the multi depot heterogenous vehicle routing problem with time windows, in *Hybrid Intelligent Systems*, Springer, Cham, pp. 147–157.

Ronen, D. (1995) Dispatching petroleum products. *Operations Research*, **43**(3), 379–387.

Rostami, B., Desaulniers, G., Errico, F. and Lodi, A. (2021) Branch-price-and-cut algorithms for the vehicle routing problem with stochastic and correlated travel times. *Operations Research*, **69**(2), 436–455.

Sarin, S.C., Sherali, H.D., Judd, J.D. and Tsai, P.-F.J. (2014) Multiple asymmetric traveling salesmen problem with and without precedence constraints: Performance comparison of alternative formulations. *Computers & Operations Research*, **51**, 64–89.

Sauré, A., Begen, M.A. and Patrick, J. (2020) Dynamic multi-priority, multi-class patient scheduling with stochastic service times. *European Journal of Operational Research*, **280**(1), 254–265.

Shao, Y., Bard, J.F. and Jarrah, A.I. (2012) The therapist routing and scheduling problem. *IIE Transactions*, **44**(10), 868–893.

Shapiro, A., Homem-de Mello, T. and Kim, J. (2002) Conditioning of convex piecewise linear stochastic programs. *Mathematical Programming*, **94**(1), 1–19.

Song, Y., Ulmer, M.W., Thomas, B.W. and Wallace, S.W. (2020) Building trust in home services–stochastic team-orienteering with consistency constraints. *Transportation Science*, **54**(3), 823–838.

Stavropoulou, F. (2022) The consistent vehicle routing problem with heterogeneous fleet. *Computers & Operations Research*, **140**, 105644.

Trautsamwieser, A. and Hirsch, P. (2014) A branch-price-and-cut approach for solving the medium-term home health care planning problem. *Networks: An International Journal*, **64**(3), 143–159.
World Health Organization (2014) Ageing and life-course. https://www.who.int/ageing/about/facts/en/. Accessed December 27, 2018.
Yang, M., Ni, Y. and Yang, L. (2021) A multi-objective consistent home healthcare routing and scheduling problem in an uncertain environment. *Computers & Industrial Engineering*, **160**, 107560.
Zhan, Y., Wang, Z. and Wan, G. (2021) Home care routing and appointment scheduling with stochastic service durations. *European Journal of Operational Research*, **288**(1), 98–115.
Zhang, Y., Zhang, Z., Lim, A. and Sim, M. (2021) Robust data-driven vehicle routing with time windows. *Operations Research*, **69**(2), 469–485.
Zhang, Z., Luo, Z., Qin, H. and Lim, A. (2019) Exact algorithms for the vehicle routing problem with time windows and combinatorial auction. *Transportation Science*, **53**(2), 427–441.
Zhao, Y. and Alfandari, L. (2020) Design of diversified package tours for the digital travel industry: A branch-cut-and-price approach. *European Journal of Operational Research*, **285**(3), 825–843.
Zhen, L., Gao, J., Tan, Z., Wang, S. and Baldacci, R. (2023) Branch-price-and-cut for trucks and drones cooperative delivery. *IISE Transactions*, **55**(3), 271–287.