Production and Operations Management

Last-Mile Attended Home Healthcare Delivery: A Robust Strategy to Mitigate Cascading Delays and Ensure Punctual Services

Mingda Liu
Department of Industrial Engineering, Tsinghua University, Beijing 100084, China.

Yanlu Zhao
Durham University Business School, Durham University, Durham DH1 1SL, United Kingdom.

Xiaolei Xie
Department of Industrial Engineering, Tsinghua University, Beijing 100084, China.

Abstract: The attended home healthcare (AHH) industry is experiencing rapid growth due to the rising demand from an aging population and the potential benefits of alleviating pressures on traditional healthcare resources. However, ensuring timely one-on-one AHH services for homebound patients remains a challenge because of cascading delays arising from uncertainties in travel and service times. To address this issue in last-mile homecare delivery, we develop a systematic cascading delay mitigation strategy to ensure patients receive dependable homecare services. Specifically, we introduce a compound set reliability index (CSRI) that captures risk exposure by separately characterizing distinct travel and service time uncertainties, instead of approaching them as a single type of uncertainty in previous studies. The CSRI-based service-level constraints are then integrated into a set-partitioning formulation to mitigate the cascading delays. We devise an exact branch-price-and-cut framework and employ a variable neighborhood search metaheuristic to achieve fast-effective solutions. Numerical experiments with benchmark and real-world datasets validate the effectiveness of our methods, underscore the benefits of adopting a systematic cascading delay mitigation strategy, and provide insights to AHH service providers regarding the impact of crucial managerial parameters on delay manifestation. The CSRI constraints and dedicated solution methods can effectively support practical decision-making and enhance the punctuality of AHH services, leading to better service dependability and heightened stakeholder satisfaction.

Key words: Attended home healthcare, Cascading delays, Compound set reliability index, Distributionally robust optimization, Branch-price-and-cut

History: Received: October 2023; Accepted: November 2024 by Michael Pinedo after two revisions.

1 Introduction

As we step into the future, an extraordinary shift is taking place in global population dynamics. The world is witnessing incredible growth in both the size and proportion of the elderly population. By 2050, 22% of the global population is projected to be over 60 years old, which is double the figure in 2015 reported by World Health Organization (2022). While this demographic enjoys longer life spans, the healthcare sector grapples with increasing pressure as medical demand outpaces both population growth and the availability of resources. The burgeoning healthcare needs of elderly individuals, primarily driven by their heightened susceptibility to health issues, are pushing service demand to unprecedented levels.

For example, a US-based study highlighted that the healthcare expenditure for those aged 65 and above was $22,356 per person in 2020, more than five times the spending per child ($4,217) and nearly three times that of working-age individuals ($9,154) (Centers for Medicare & Medicaid Services 2023). Global healthcare systems are facing a daunting challenge in preparation for this demographic shift, necessitating innovative solutions to meet these escalating medical demands (Green 2012).

In response, *attended home healthcare* (AHH) services are emerging as promising alternatives to traditional hospital care, offering multifaceted benefits (Rowe et al. 2016). Specifically, AHH caregivers, such as registered nurses, physical therapists, or personal support workers, are responsible for administering medical treatment, managing palliative care plans, and delivering superior one-on-one healthcare services at the patient's residence. This new health system presents significant advantages in terms of cost, convenience, prevention or postponement of hospital readmission, and less strain on mainstream medical resources (Cire and Diamant 2022). For example, the NHS (2022) in the UK advocated homecare treatments for mobility-challenged patients owing to their adaptability, comfort, and cost-effectiveness, pricing treatments as low as £20. Such services have proven invaluable, particularly during pandemics, by safeguarding clinically vulnerable populations through essential interventions such as vaccinations and PCR testing (NHS 2020). In the US, it is estimated that up to $265 billion worth of care services for Medicare fee-for-service and Medicare Advantage beneficiaries could shift to the home by 2025 (Mckinsey 2022a). Furthermore, a comprehensive survey involving 393,858 homecare visits revealed that prolonging treatment during a homecare visit by just one minute could reduce the likelihood of hospital readmission by 1.39%. A 10% extension in treatment time could reduce this likelihood by an additional 6% (Song et al. 2022), showcasing the immense value offered by AHH services to patients and healthcare infrastructures alike.

Parallel to this trend, there is a growing interest in reenvisioning the future of *care at home* ecosystems (Mckinsey 2022b). Generally, the primary goal of AHH is to provide professional services in residential and community settings, supporting patients in various aspects of health and social care, including home care, long-term care, assisted living, and substance use disorder treatments. Some of these services, such as insulin injections (Fikar and Hirsch 2017), heart failure care (NHS 2023), and surgical wound care (Sessler 2006), are critically time-sensitive. Any delays, even minor delays, can result in severe complications, deterioration in patient health, or rehospitalization. As Song et al. (2022) stressed, each minute of AHH services is invaluable in reducing readmission rates. Additionally, delays that may seem trivial in other contexts, such as grocery delivery, are unacceptable in the AHH context, especially for patients enduring suffering. Even delays in less time-critical services can result in patient dissatisfaction or loss of demand. A survey of over 9,000 homecare clients indicated that unmet expectations, particularly regarding punctual and accountable caregivers, are among the most common complaints (Kumar Saha 2020, Julie 2022). On the other hand, service delays can also lead to extended working hours, causing caregiver complaints or voluntary turnover (Kong et al. 2022, Bergman et al. 2023). Thus, effective delay management is crucial in the AHH context.

However, delays are omnipresent as evidenced by our field investigation of the Chinese AHH service provider, *Pinetree*. An analysis of approximately 150 real-world schedules and visit routes from the company revealed an average delay of 8.59 minutes per patient. Figure 1a illustrates the delay distribution for routes catering to at least two patients. The pronounced tail in this histogram highlights significant delays relative to the expected schedules, with some stretching to 110 minutes. Interestingly, our data analysis flagged a *cascading effect*, as shown in Figure 1b. This effect suggests that the average delay (i.e., the postponed service starting time) tends to amplify progressively, with subsequent visits inheriting and potentially exacerbating previous delays. This phenomenon resembles the backward propagation bullwhip effect in classical supply chain models (Lee et al. 1997) and is also observed in other areas, such as the increasing interappointment times in the

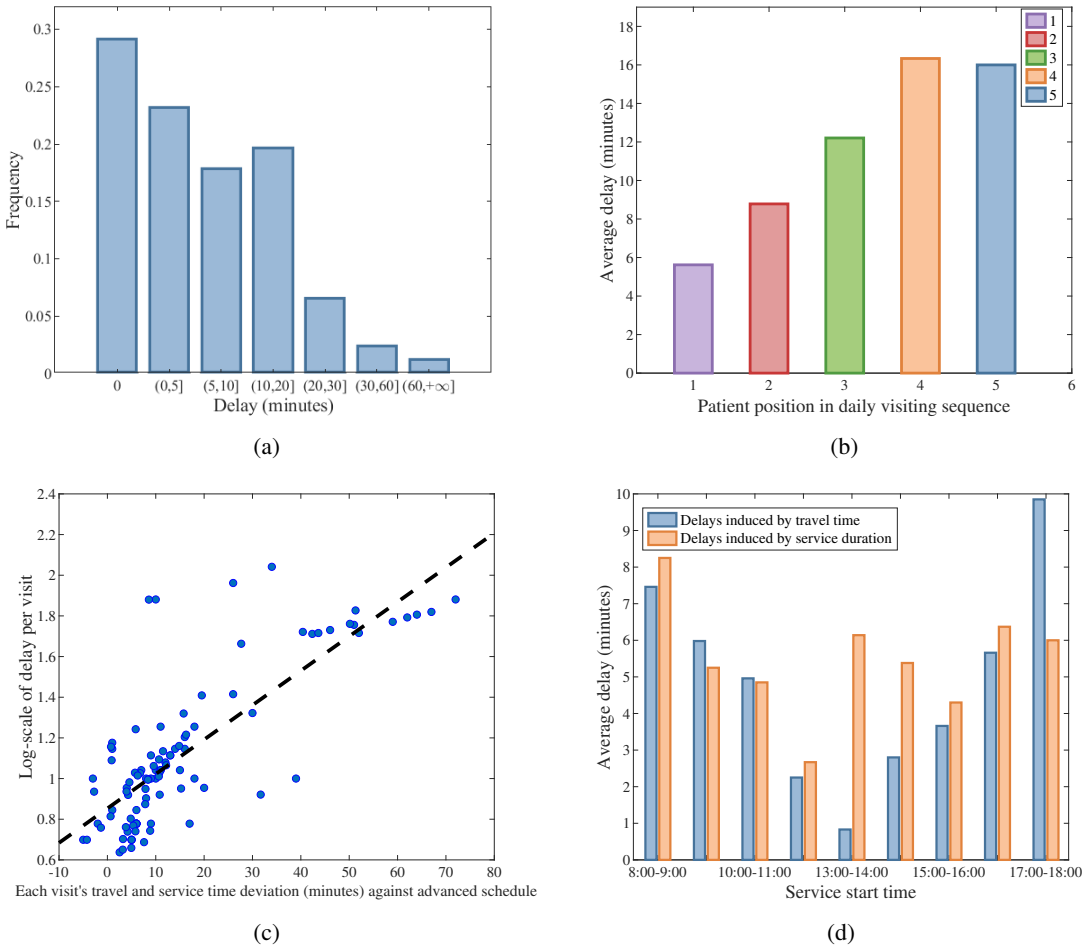


Figure 1 Description of the Practical Scheduling Situations for an AHH Service Provider

appointment scheduling sector (Benjaafar et al. 2023). However, to our knowledge, this cascading effect has not been explored and addressed well in the AHH domain, especially from scheduling and routing perspectives.

After identifying the cascading effect, a pertinent follow-up question emerges: *What are the primary factors and how do they result in and then propagate these delays along visiting paths?* An intuitive reason can be traced back to the *convolutional* uncertainties of travel and service times that arise when caregivers deliver AHH services in line with the initial plans. To quantify the pattern in which convolutional uncertainties contribute to delays, we empirically studied the correlation between the deviation (from the planned schedule) associated with the last (i.e., most recent) single travel and service time and the actual delay for the realized AHH delivery, as illustrated in Figure 1c. The findings indicate a positive correlation between the latest deviation for each visit and its delay value on a log scale, with an R^2 value of 0.625. This signals that, in addition to the latest deviation, the previous accumulation is also responsible for causing delays and the resulting cascading effects. We further investigated the impact of endogeneity (time dependency) on the delay behavior. Interestingly, while uncertain travel and service time delays contribute almost equally to the overtime, they display distinct temporal patterns[1], as presented in Figure 1d. Delays arising from travel time reach 'peak' during morning and evening rush hours, possibly because of heightened congestion from commuting traffic. In contrast, delays induced by service time exhibit a more consistent trend across different times of the day. In light of these observations, we conclude that the

[1] Note that the *delay* associated with travel and service times measures the difference between pre-planned and actual realizations. Ideally, a zero delay for travel and service times indicates adherence to the planned schedule.

uncertain nature of both travel and service times is the primary source of cascading delays. They exhibit different features and introduce a fresh challenge in maintaining a resilient schedule, delivering services as promised, and adhering to the working time regulation.

To mitigate delays, both academic and industrial efforts have strived to develop sophisticated techniques for ensuring timely service and completion. This surge of interest has led to a flurry of activities in the community adopting modern operations research tools to achieve on-time and reliable schedules (Lu and Shen 2021). Notably, the Markov decision process formulation (Sauré et al. 2020), the stochastic mixed-integer programming model (Zhan et al. 2021), and the mixed-integer robust optimization model (Naderi et al. 2023) have been established to counter potential delays in AHH service delivery. Unfortunately, technical challenges such as intractability have limited studies ensuring on-time service and completion when accounting for unknown distributions of travel and service times. Their distributions are seldom accurately derived from observations, yet they are the primary reasons for delays. To tackle this, distributionally robust optimization (DRO), which assesses the worst-case over the empirical distribution vicinity, has been introduced to reduce service delays in robust vehicle routing studies. However, the majority of studies in this stream either focus on individual uncertainty components (travel/service time) or treat them as homogeneous (Zhang et al. 2019, 2021, 2024, Tsang and Shehadeh 2023), thereby overlooking their *delineated characteristics*. This oversight fails to accurately capture real-world scenarios, particularly when the uncertainties exhibit distinct features that are not easily defined by a single ambiguity set. Unluckily, this is exactly the case confirmed by the descriptive analytics using real-world AHH data from our case company (details are presented in Figure 1d and EC.1). We find that uncertain travel and service times are equally to blame for causing delays but present inherently different characteristics. Indeed, travel times in urban areas are correlated, presenting an obvious time-dependent (endogenous) property (Parent and LeSage 2010, Xing and Zhou 2011). On the other hand, service times seem generally independent and exhibit lower variability (Liu et al. 2021, 2024a). As a result, ignoring these unique attributes and characterizing them through a single ambiguity set will lead to inferior precision in risk assessment and subsequently, diminished performance. For example, Chen et al. (2023) highlighted that while the moment-based ambiguity set effectively captures the interdependency of a multivariate random variable's components, its application for representing independently distributed random variables leads to a significant overestimation of actual risks. Moreover, Jiang et al. (2017) adopted a first-moment ambiguity set for heterogeneous random no-shows and service durations simultaneously, and their distinct features prevent the capture of more information (e.g., variance) through a single ambiguity set. Hence, it is crucial to govern the inherent features of uncertain travel and service times separately, and address them convolutively by employing a *compound ambiguity set* to alleviate the cascading effect.

Despite various delay-reduction initiatives, comprehensive studies in the AHH domain that explore the adverse effects of propagating delays in terms of both probability and magnitude are still scarce. The challenges in ensuring on-time service and completion in scenarios plagued by uncertain travel and service times often render delay assessments futile (Tsang and Shehadeh 2023). To our knowledge, no existing study has evaluated or measured AHH cascading delays stemming from cumulative uncertainties. Instead, many AHH studies resort to basic assumptions such as the linear penalty function for overtime expectations when scheduling caregiver visits, underestimating and oversimplifying the repercussions of delay propagation for downstream patients caused by the cascading effect (Zhang et al. 2021). To rectify this, we introduce a patient-centric delay metric capped by a threshold, aiming to counteract the cascading effect. More importantly, these service-level constraints, which establish a ceiling on the specific risk measures for each node, have proven effective in other domains, such as appointment scheduling (Benjaafar et al. 2023), inventory routing (Cui et al. 2023), and surgical scheduling (Zhou et al. 2021).

In this article, we investigate a robust AHH scheduling and routing problem (RSRP) for managing last-mile homecare deliveries. This problem is categorized as a *robust heterogeneous site-dependent vehicle routing problem with time windows (RHSDVRPTW)*. To this end, we propose a comprehensive framework that can address such a complicated problem effectively, that is, a set partitioning formulation encompassing various operational features in the AHH sector, and an exact solution approach that is able to solve the prescribed model efficiently. Leveraging this framework, we aim to answer the following questions: (i) How should uncertain travel and service times be accurately calibrated considering their distinct characteristics? (ii) What strategies can be employed to assign and schedule visits in advance for cost-optimal delivery, while ensuring on-time service and completion? To answer these questions, we contribute as follows:

- *Systematic cascading delay mitigation strategy*: (i) As AHH operational practices present different uncertainties, we introduce a compound ambiguity set that reconciles multivariate random variables with independent distributions and distinct features, enhancing risk assessment accuracy over conventional ambiguity sets. (ii) Moreover, to assess punctuality violations and quantify the risk exposure of scheduling outcomes, we adopt a tailored decision criterion — *compound set reliability index* (CSRI), which gauges delays from both probability and magnitude perspectives based on the compound ambiguity set. We also theoretically demonstrate its effectiveness against the prevalent lateness probability index (LPI; Adulyasak and Jaillet 2016). (iii) Finally, to mitigate cascading delays, our approach is rooted in the CSRI-based service-level constraints, which are further embedded in the set-partitioning model and evaluated leveraging three customized methods in the developed solution approach.
- *Dedicated cascading delay mitigation approach*: A holistic exact solution approach, built upon the branch-price-and-cut framework (`CSRI-BPC`), is developed to solve the RSRP, which subtly embodies flexibility against the CSRI constraints in the pricing subproblems amidst uncertain travel and service times. Among others, techniques such as the CSRI-based variable neighborhood search metaheuristic (`CSRI-VNS`) and CSRI-based dominance rule, are further designed to enhance the solution efficiency. Numerical evaluations validate the ability of our solution approach to address both benchmark and practical instances with up to 100 patients. This efficiency can be generalized to other stochastic programming or (distributionally) robust optimization problems by bridging our `CSRI-BPC` framework with the tailored risk measures, satisficing measures, and disutility functions therein.
- *Pragmatic cascading delay mitigation implications*: Empirical results and comparisons with other risk measures underscore the efficacy of our CSRI-based strategy in mitigating delay propagation. Remarkably, the average delay tested on real-world instances decreases from 8.59 minutes to a minuscule 0.16 minutes, with the cascading effect being nearly imperceptible. Moreover, we find that the delay manifestation is contingent on the topological structure of the graph. Specifically, clustered nodes seem to be susceptible to amplifying the cascading effect compared with randomly distributed nodes, thereby leading to larger expected lateness times. Such observations could guide decision-makers in wisely configuring on-time service levels to align with patient distributions in different districts. Additionally, the delay mitigation strategy also offers profound insights into optimizing multiple model choices against operating costs, customizing patient service times, and regulating caregiver working hours in AHH service management.

2 Literature Review

This research intersects with literature streams on AHH services and vehicle routing problems (VRPs) that consider uncertainty. In the following, we review seminal works in each domain and elucidate how our study can be distinguished from these frontiers.

2.1 Research on Attended Home Healthcare Services

The AHH scheduling and routing problem, which has been extensively investigated in recent decades, is essentially a variant of the VRP with time windows (VRPTW) that features unique constraints stemming from the AHH landscape. For a comprehensive overview, we suggest referring to Fikar and Hirsch (2017).

A significant difference in AHH services relative to conventional VRPTWs is the heterogeneity of caregivers. This disparity is underscored by attributes such as individual home addresses, skill portfolios, and demographic features (Naderi et al. 2023). Traditional AHH research typically categorizes caregivers by their skill levels and assumes that patient requirements are fulfilled by caregivers with higher skill levels (Bard et al. 2014). To describe divergent and incompatible skills and patient preferences, Cire and Diamant (2022) defined skill sets and adopted a hierarchy to verify caregiver competence, leading to a site-dependent VRP to ensure skill-matching requirements (Baldacci et al. 2010).

However, deterministic schedules often fail to deliver AHH services as promised facing uncertainty. Considering uncertain patient demand, Cire and Diamant (2022) developed a discrete-time, rolling-horizon, infinite-stage Markov decision process and proposed an approximate dynamic programming approach to solve the problem. Kong et al. (2020) deployed a distributionally robust model to handle time-dependent patients with no-show behavior and reformulated the problem as a copositive program. Considering the uncertain times embedded in AHH services, Sauré et al. (2020) formulated a Markov decision model that captures stochastic service times, prioritizing cost-effectiveness in overtime and idle times. Naderi et al. (2023) incorporated uncertain travel time and employed a hybrid of interval and polyhedral uncertainty sets to formulate the protection function for deadline violations. The authors subsequently designed a logic-based Benders branching-decomposition algorithm to solve the problem.

The robust AHH scheduling and routing problem is further complicated by time window constraints. To achieve on-time performance, most studies explicitly formulate time-related decision variables to incorporate the overtime and idle time penalties for each patient in the objective minimization, which is associated with a substantial computational burden. For example, Zhan et al. (2021) developed a stochastic mixed integer programming (MIP) model and proposed an integer L-shaped method to solve the sample average approximation (SAA) version of the problem. More recently, Tsang and Shehadeh (2023) applied mean-support and 1-Wasserstein ambiguity sets to characterize uncertain times and derived equivalent MIP reformulations of both DRO models. However, their maximum-scale tested instances include only 10 patients, and the proposed approaches are insufficient to support practical operations. Despite sharing certain similarities with current studies (Tsang and Shehadeh 2023, Naderi et al. 2023, Liu et al. 2024a), our work departs from the traditional robust AHH stream in terms of comprehensive AHH formulation, compound uncertainty characterization, and dedicated solution approach capable of handling real-world instances.

2.2 Research on the Vehicle Routing Problem with Uncertain Factors

Previous studies have extensively examined the deterministic VRPTW, yet comparatively little effort has been devoted to exploring its uncertain counterparts. However, the practical significance of uncertainty has been gradually acknowledged, catalyzing growing interest in integrating data into the model-building process for robust vehicle routing optimization. For example, Ghosal and Wiesemann (2020) and Ghosal et al. (2024) developed a unified branch-and-cut framework to solve the capacitated vehicle routing problem (CVRP) with uncertain demands. Nevertheless, this framework is not applicable to VRP with sequence-related constraints (e.g., time windows, distance, or working length restrictions) because it violates the permutation invariant assumption. We close this gap and investigate the RSRP with cascading delays in the AHH context, where the visiting sequence needs to be explicitly considered and two sources of uncertainty exist with different

characteristics. To this end, we reviewed relevant studies on stochastic VRP optimization and particularly, (distributionally) robust optimization.

In the context of stochastic VRP, most studies assume independent travel and service times with fully known distributions and endeavor to minimize deadline violations (e.g., caregiver's working time regulation). For example, Laporte et al. (1992) defined illegal routes by bounding the probability of late return in the chance-constrained model and demonstrated that the model is similar to a deterministic VRP model if travel and service times have additive probability distributions. In terms of the penalty cost of exceeding the route duration, the authors proposed recourse models and developed a branch-and-cut approach for solving the problem. Incorporating customer time windows renders the problem more challenging. Hashemi Doulabi et al. (2020) formulated the VRPTW with stochastic travel and service times as a two-stage stochastic integer programming model without big-M constraints, which was solved by an L-shaped algorithm with the maximum tested instances including 20 patients. In practice, however, the distribution functions typically cannot be accurately obtained for stochastic programming, and the involved calculation is time-consuming since the multivariate integral is #P-hard (Esfahani and Kuhn 2018).

Some studies subsequently adopted robust optimization paradigms and characterized unknown durations with different classes of uncertain sets. For example, Munari et al. (2019) addressed the VRPTW with uncertain travel time through a budgeted polyhedral uncertainty set. The authors proposed a branch-price-and-cut (BPC) method that relies on a robust resource-constrained elementary shortest path problem to generate robust routes in terms of both vehicle capacity and customer time windows. Bartolini et al. (2021) studied a robust traveling salesman problem with time windows in which the travel times are within a knapsack-constrained uncertainty set. The authors devised an exact method based on column generation and route enumeration. However, the robust optimization approach used by this stream of studies is excessively conservative as the extreme situations included in the uncertain set definition generally incur more costs and poor out-of-sample performance.

To ensure timely service and completion, certain initiatives have integrated the DRO approach with crafted decision criteria regarding delays to strengthen scheduling resilience. Specifically, Zhang et al. (2019) proposed a DRO model with a cross moment ambiguity set for uncertain travel times and the essential riskiness index to characterize delays in the traveling salesman problem. Zhang et al. (2021) solved the VRPTW counterpart with a Wasserstein distance-based ambiguity set and proposed a new decision criterion service fulfillment risk index (SRI). The authors incorporated the optimality and feasibility constraints of SRI into a branch-and-cut scheme to eliminate violated solutions. More recently, Zhang et al. (2024) designed a general decision criterion, termed the generalized riskiness index, and developed a BPC algorithm that can consistently solve Solomon's instances with up to 50 nodes. However, they assumed the known distribution of uncertainty to evaluate the index value, which fails to work when the uncertainty is represented by ambiguity sets. Although adopting similar decision criteria as the above studies, our work captures the delineated uncertainties in the AHH context and demonstrates that precisely governing the different features of travel and service times with compound ambiguity set can effectively enhance the efficacy of mitigating cascading delays. More specifically, our work combines the SRI risk measure with a convolutional ambiguity set that is composed of different types of uncertainty sets, versus the same type of uncertainty sets with different parameters, to achieve better risk management. Additionally, the current stream of studies usually optimizes the total risk measures within a budget constraint, which may incur risk imbalance among individual nodes and lead to severe delays due to propagation. As a result, we adhere to a different formulation by restricting the CSRI constraints to each node and guaranteeing individual service quality.

3 Problem Description and Formulation

In this section, we describe the business problem (i.e., RSRP) observed from the AHH practice and present a set-partitioning formulation to characterize the decision process.

3.1 RSRP Description

The RSRP depicts a general AHH scheduling and routing problem that assigns available caregivers to geographically dispersed patients and plans their visiting routes. Before each service day, the service provider manually finalizes critical operational decisions to minimize the operating costs, which include: (i) assigning patients to specific caregivers; (ii) designing the visiting sequence for each caregiver; and (iii) determining the arrival time at each patient's residence. In fact, the service provider pays caregivers proportional to their total engagement time, which includes both service duration and travel time[2]. Since all patients must be served and their random no-shows are omitted, the total service duration is predetermined. By normalizing the cost related to total service duration to zero, consequently, the objective of minimizing total cost is equivalent to minimizing total travel cost, thus they are used interchangeably hereafter. To obtain an optimal solution, these decisions are constrained by various AHH features in practice, including but not limited to the following:

- *Stakeholder heterogeneity and skill matching*: Caregivers and patients differ in terms of health skills, service requirements, period availability, service duration, and region of residence. This results in several remarks. First, advanced AHH studies assume that caregivers depart from and return to their homes and have diverse availability patterns for providing services (Mosquera et al. 2019, Liu et al. 2024a). Second, the depots (caregivers' homes) are predefined. Finally, each caregiver is characterized by qualifications and skillsets, as well as some demographic characteristics, such as gender and age (Cire and Diamant 2022).
- *Time window and working time regulation*: When providing AHH services, respecting time window and working time regulation is critical for ensuring stakeholder satisfaction as previously emphasized. The former refers to the specific intervals during which a patient can receive services, accommodating her availability and preferences. The latter pertains to the maximum shift lengths and mandatory rest periods for caregivers, complying with labor laws and caregiver well-being considerations. To hedge against uncertain durations, an on-time service-level constraint is necessary to ensure that the schedule respects these two requirements to a prescribed extent (Benjaafar et al. 2023, Cui et al. 2023, Zhou et al. 2021).
- *Workload, fatigue and medical resources*: Even when adhering to working time regulations, consecutive services, especially for patients who are clustered with trivial travel times, can lead to excessive workload and fatigue in caregivers, compromising the quality of care provided by them. Additionally, caregivers require various medical consumables for patient care, such as insulin syringes, venipuncture needles, blood collection tubes, nebulizers, and wound care supplies. Consequently, there is a prescribed capacity for each caregiver, representing the maximum number of patients or medical resources they can effectively manage within a given period, typically a shift. This constraint helps prevent caregivers from becoming overburdened, reducing the risk of fatigue-related errors and ensuring the well-being of both patients and caregivers (Yang et al. 2021).
- *Continuity of care*: In the AHH sector, continuity of care is an important factor for maintaining high service quality and patient satisfaction (Cappanera et al. 2018, Liu et al. 2024a). This constraint requires that the number of caregivers assigned to a single patient over the time horizon is limited within a cohort, enabling consistent and coordinated

[2] Note that the service provider does not account for uncertainty when calculating payments for caregivers, as incorporating such factors would significantly complicate the operational process.

care delivery. By restricting caregiver changes, continuity of care enhances trust and rapport between patients and caregivers, minimizes disruptions, and ensures that caregivers are well-acquainted with their patients' specific needs and preferences.

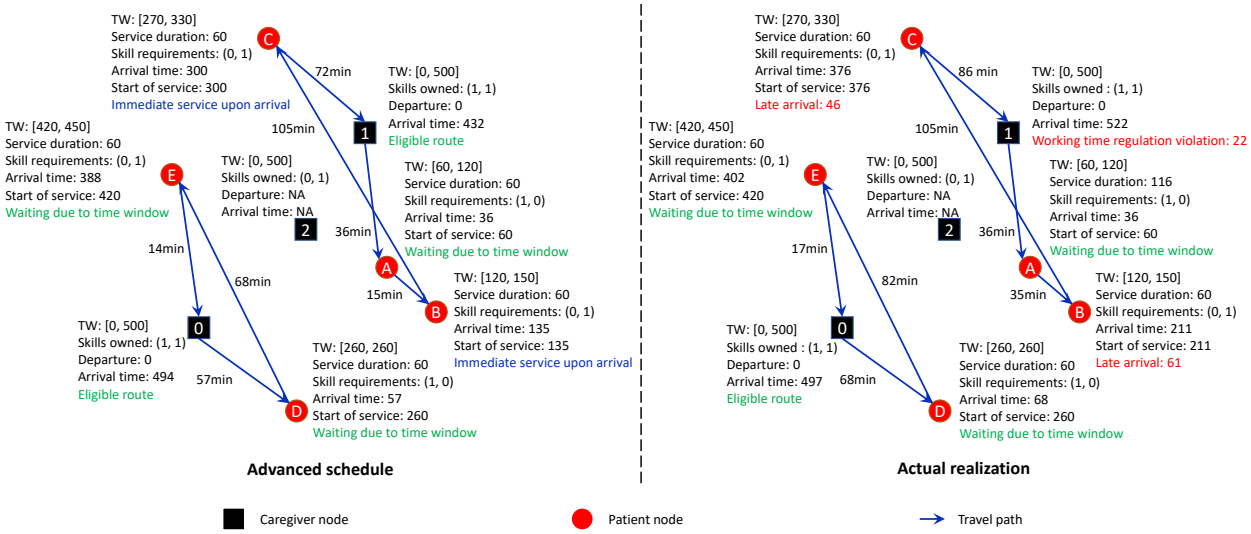


Figure 2 Example of Trade-offs in the RSRP

On each working day, caregivers depart from their homes to visit the assigned patients sequentially as indicated by the agendas, and finally return home after completing all the assigned tasks. In practical scenarios, however, the actual realization may deviate from the expectation significantly due to advanced (and static) assignments and routing decisions made without accounting for the uncertain nature of travel and service times. To better illustrate the downside caused by uncertain durations, we present a simplified realistic instance in Figure 2 with five patients and three caregivers, which also indicates the underlying trade-offs in the choice of robustness and operating costs. The data triplet next to each patient node displays individual information, including the prescribed time window, skill-matching requirements, service duration, arrival time, and service start time. The optimal routes under deterministic circumstances are described by blue solid lines, in which each patient can receive AHH service at the prescribed time slot, constituting the advanced schedule. However, upon actual realization, route $1 \to A \to B \to C \to 1$ results in late arrivals due to disturbances in several travel and service times. These situations are common in practice because of unexpected traffic congestion and caregivers spending additional time in patients' homes to handle emergencies. Note that even though the subsequent visits proceed as expected after arriving at patient B, patient C still suffers from a delay of 46 minutes due to the cascading effect, which further propagates and finally leads to a violation of working time regulation with up to 22 minutes. In contrast, route $0 \to D \to E \to 0$ is robust and remains eligible in both scenarios. As a remedy for this scenario, if rescheduling the patient C to caregiver 2 satisfies all constraints, then this new arrangement could mitigate delays and preclude potential patient dissatisfaction, at the expense of extra travel costs.

Therefore, considering these features in AHH practice, we formulate the RSRP as a mathematical program and aim to achieve a satisfactory trade-off between scheduling resiliency and operating costs, aided by the risk measure CSRI to ensure the promised on-time service level (for each patient) and completion target (for each caregiver).

3.2 Set Partitioning Formulation

The RSRP is defined on a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{A}$ represents the set of arcs. The node set $\mathcal{V}$ is partitioned as $\mathcal{V} = I \cup \mathcal{J}$, with $I = \{0, ..., |I|-1\}$ representing the caregiver locations, and $\mathcal{J} = \{|I|, ..., |I|+|\mathcal{J}|-1\}$ denoting patient addresses. The arc set $\mathcal{A}$ is defined as $\{(v,v')|v,v' \in \mathcal{V}, v \neq v'\} \backslash \{(v,v')|v,v' \in I\}$. Let $\mathcal{M}$ represent the set of patients' service requirements as well as caregivers' skills. Each node $v \in \mathcal{V}$ is associated with a series of binary parameters $\upsilon_{vm} \in \{0,1\}$, a demand $d_v \in \mathbb{R}_+$, an uncertain service time $\tilde{s}_v$ and a time window $[e_v, l_v]$, where υ_{vm} denotes whether caregiver or patient $v \in \mathcal{V}$ owns (or requires) skill $m \in \mathcal{M}$, if yes, then $\upsilon_{vm} = 1$; otherwise, $\upsilon_{vm} = 0$. The parameters $e_v \in \mathbb{R}_+$ and $l_v \in \mathbb{R}_+$ prescribe the earliest and latest service start times, respectively. Following convention, in case the earliest or latest time is not explicitly specified for some particular node $v \in \mathcal{V}$, we assign $e_v = 0$ or $l_v = \infty$ to represent the absence of such constraints, respectively. As a result, the node sets with explicit earliest and latest time window constraints are denoted as $\underline{\mathcal{V}} = \{v \in \mathcal{V} \mid e_v > 0\}$ and $\overline{\mathcal{V}} = \{v \in \mathcal{V} \mid l_v < \infty\}$. For example, we have $I \subset \overline{\mathcal{V}}$ where $l_{i \in I}$ represents the working time regulation for returning to the depot. For each patient $j \in \mathcal{J}$, the service start time cannot be earlier than e_j or later than l_j. Otherwise, early arrival will result in idle waiting time until e_j, whereas late arrival after l_j will lead to patient dissatisfaction. We assume a prescribed capacity $Q_i \in \mathbb{N}_+$ of each caregiver $i \in I$ due to the limitation for carrying medical instruments to serve patients. Without loss of generality, we define $\tilde{s}_{i \in I} = 0$, $d_{i \in I} = 0$ and $d_{j \in \mathcal{J}} = 1$ such that Q_i represents the maximum number of patients to be visited by a caregiver i. For each arc $(v,v') \in \mathcal{A}$, the travel cost and uncertain travel time are denoted as $c_{vv'} \in \mathbb{R}_+$ and $\tilde{t}_{vv'}$, respectively. Let $\tilde{\xi}_v$ denote the *uncertain* delay of node $v \in \mathcal{V}$ along the given route, then the corresponding service level can be evaluated by the CSRI $\rho_\gamma(\tilde{\xi}_v)$, which is a risk measure of delay with the risk aversion parameter $\gamma \in [0,1]$ and will be explained in more detail in Section 4.2. Finally, let $\beta \in \mathbb{R}_+$ be a prescribed threshold to restrict the CSRI for each node. Note that, a smaller β value reflects more stringent punctuality requirements, indicating the decision-maker's conservative attitude toward prioritizing service quality over minimizing the operating costs. In contrast, a larger β value signifies that the decision-maker values the cost objective over the schedule dependability. For ease of readability, we have described all the parameters and decision variables in Table EC.1 of the e-companion.

We assume that $\mathcal{G}$ is an acyclic graph, and a feasible route corresponds to an elementary path $r = (i, j_1, ..., j_k, i)$ for a caregiver $i \in I$ and visiting patients $j_1, ..., j_k \in \mathcal{J}$ such that

(i) the sum of patient demands cannot exceed caregiver's capacity: $\sum_{j \in \{j_1,...,j_k\}} d_j \leq Q_i$;

(ii) the caregiver possesses the specific skill that the visited patients require: $\upsilon_{im} \geq \upsilon_{jm}, \forall j \in \{j_1, ..., j_k\}, m \in \mathcal{M}$;

(iii) the service level upon each node $v \in \{j_1, ..., j_k, i\}$ satisfies the predefined risk level: $\beta \in \mathbb{R}_+$: $\rho_\gamma(\tilde{\xi}_v) \leq \beta, \forall v \in \{j_1, ..., j_k, i\}$. Note that, this condition primarily encompasses the consideration of time windows for patients and working time regulation for caregivers.

To characterize this problem, let $\tilde{\mathcal{R}}$ be the set of all feasible routes (i.e., elementary routes that satisfy the capacity, skill-matching and CSRI constraints of each node) and let $c_r \in \mathbb{R}_+$ denote the corresponding travel cost for route $r \in \tilde{\mathcal{R}}$, in other words, the total travel cost on arcs connecting all traversed nodes within the route. Furthermore, let binary parameter b_v^r indicate whether node $v \in \mathcal{V}$ is traversed ($b_v^r = 1$) or not ($b_v^r = 0$) by the route r.

For each route $r \in \tilde{\mathcal{R}}$, we define the binary decision variables $z_{r \in \tilde{\mathcal{R}}}$, where $z_r = 1$ indicates that route r is selected in the solution; otherwise, $z_r = 0$. Thus, the RSRP can be formulated as the following set partitioning model $\mathcal{F}_{\text{SPF}}$:

$$[\mathcal{F}_{\text{SPF}}] \qquad \min \sum_{r \in \tilde{\mathcal{R}}} c_r z_r \tag{1a}$$

$$\text{s.t.} \sum_{r\in\tilde{\mathcal{R}}} b_i^r z_r \le 1, \qquad \forall i \in I, \tag{1b}$$

$$\sum_{r\in\tilde{\mathcal{R}}} b_j^r z_r = 1, \qquad \forall j \in \mathcal{J}, \tag{1c}$$

$$z_r \in \{0,1\}. \qquad \forall r \in \tilde{\mathcal{R}}. \tag{1d}$$

The objective function (1a) aims to minimize the total travel cost. Constraints (1b)-(1c) ensure the availability of caregivers and patients with medical needs being visited. Finally, constraints (1d) define the domains of the decision variables. For clarity, we also present a compact formulation to characterize the RSRP in EC.2.

Note that a significant benefit of formulating the RSRP as $\mathcal{F}_{\text{SPF}}$ is that some of the constraints, such as the capacity, skill-matching and service-level constraints, do not need to be explicitly expressed but are instead implicitly considered in the definition of route $r \in \tilde{\mathcal{R}}$. This allows feasibility checks for these constraints to be integrated into the column generation procedure presented in Section 5.1.2. However, two major challenges are still waiting to be solved: First, evaluating the service-level constraint for each node is not trivial, and depends on the decision variable $\boldsymbol{z}$, uncertain service times $\tilde{\boldsymbol{s}}$ and uncertain travel times $\tilde{\boldsymbol{t}}$. Indeed, the service-level constraints are formally defined as follows.

$$\rho_\gamma(\tilde{\xi}_v(\boldsymbol{z},\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}})) \le \beta, \qquad \forall v \in \overline{\mathcal{V}}. \tag{2}$$

In the next section, we will explain how incorporating these CSRI-based service-level constraints helps mitigate delay risk. Second, $\mathcal{F}_{\text{SPF}}$ focuses solely on single-period scheduling and routing decisions to develop an effective cascading delay mitigation strategy, and thus neglects the continuity of care requirements in the multiple-period setting for ease of computational burden. As an extension, in EC.7, we introduce the model and solution approach that incorporates the continuity constraints. Additionally, we conduct further numerical experiments to evaluate the performance of the delay mitigation strategy across various continuity scenarios.

4 CSRI-Based Delay Risk Mitigation

In this section, we first introduce the compound ambiguity set for governing distinct travel and service time uncertainties. Next, we present CSRI as a metric to assess schedule performance, discuss the implications of CSRI constraints (2), and analyze its structural properties for tractable reformulations from three different perspectives. Finally, we theoretically compare the CSRI with the prevalent decision criterion LPI.

4.1 Compound Ambiguity Set

In real-world situations, the exact distributions $\mathbb{P}_{\tilde{\boldsymbol{t}}}$ and $\mathbb{P}_{\tilde{\boldsymbol{s}}}$ of travel and service times usually cannot be obtained exactly from observations. However, with some information on how they have evolved from historical records, we can approximate them by assuming that the true distributions $\mathbb{P}_{\tilde{\boldsymbol{t}}}$ and $\mathbb{P}_{\tilde{\boldsymbol{s}}}$ belong to certain families of distributions $F_{\tilde{\boldsymbol{t}}}$ and $F_{\tilde{\boldsymbol{s}}}$, respectively. In this regard, existing DRO studies typically assume the same type of ambiguity set for uncertain travel and service times (Tsang and Shehadeh 2023, Zhang et al. 2021). Nevertheless, sometimes travel and service times manifest different characteristics, especially in the AHH context as previously depicted in Figure 1d. Consequently, the classical single ambiguity set, even with different parameters, turns out to be "coarse" for precise risk evaluation, as we will see later in the next section. To assess risk more accurately, we separately capture the characteristics of travel and service times by constructing the most suitable ambiguity set individually, and cast them into a *compound ambiguity set* as follows.

DEFINITION 1. The compound ambiguity set is defined as the Cartesian product of an ambiguity set $F_{\tilde{\boldsymbol{t}}}$ for travel time and an ambiguity set $F_{\tilde{\boldsymbol{s}}}$ for service time, given by

$$F = F_{\tilde{\boldsymbol{t}}} \times F_{\tilde{\boldsymbol{s}}}. \tag{3}$$

In fact, some existing studies have implicitly adopted a compound ambiguity set in cases where $F_{\tilde{\boldsymbol{t}}}$ and $F_{\tilde{\boldsymbol{s}}}$ belong to the same type of ambiguity set and their compound ambiguity set coincidentally reduces to a classical single ambiguity set. For instance, when both $F_{\tilde{\boldsymbol{t}}}$ and $F_{\tilde{\boldsymbol{s}}}$ are modeled as mean-support ambiguity sets, their Cartesian product remains a single mean-support ambiguity set (Tsang and Shehadeh 2023). However, if $F_{\tilde{\boldsymbol{t}}}$ and $F_{\tilde{\boldsymbol{s}}}$ are based on different types of ambiguity sets, then their Cartesian product cannot be represented as a single unified type of ambiguity set.

This is exactly the case for the compound ambiguity set in the AHH context, which necessitates the examination of two different ambiguity sets. Our study is also the first to consider this scenario in general. As previously revealed in Figure 1d, travel times are highly time-dependent and correlated, and may even exhibit considerable fluctuations due to unforeseen emergencies. In contrast, service times remain independent and relatively stable, as the duration is primarily managed by caregivers. These distinct features necessitate a compound ambiguity set for describing travel and service times separately. For travel time uncertainty, some studies often circumvent interdependencies by presuming that travel times are independently distributed (Tsang and Shehadeh 2023, Zhang et al. 2021). As this assumption simplifies risk assessment, the actual risk could be underestimated in the absence of correlations (Chen et al. 2023). On the contrary, ambiguity sets that depend on decision variables to fully counteract endogenous uncertainty may lead to considerable computational intricacy (Kong et al. 2020). Therefore, we employ the following cross moment ambiguity set to hedge against uncertain travel times by resorting to a variance-covariance matrix, which is sufficient to characterize their correlations (Prakash and Srinivasan 2018, Rostami et al. 2021) and avoids excessive conservatism associated with only specifying marginal distributions (Chen et al. 2022).

$$F_{\tilde{\boldsymbol{t}}}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^{|\mathcal{A}|}) \,\middle|\, \begin{array}{l} \mathbb{E}_{\mathbb{P}}(\tilde{\boldsymbol{t}}) = \boldsymbol{\mu} \\ \mathbb{E}_{\mathbb{P}}\left((\tilde{\boldsymbol{t}}-\boldsymbol{\mu})(\tilde{\boldsymbol{t}}-\boldsymbol{\mu})^{\top}\right) = \boldsymbol{\Sigma} \end{array} \right\}, \tag{4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{A}|}_{++}$ is a positive vector, $\boldsymbol{\Sigma} \succ 0$ is a positive definite matrix, and $\mathcal{P}$ represents the set of all probability distributions on $\mathbb{R}^{|\mathcal{A}|}$.

With respect to the service time uncertainty, although a diagonal covariance matrix can describe uncorrelated distributed random variables, the cross moment ambiguity set may not be sufficient for capturing their independence, which is a more stringent condition than simply being uncorrelated. In other words, relying solely on the cross moment ambiguity set for robust risk evaluation may result in a significant overestimation of the actual risk (Chen et al. 2023). Therefore, the Wasserstein ambiguity set over the empirical distribution vicinity is more advantageous for effectively utilizing reliable historical data (Esfahani and Kuhn 2018, Zhang et al. 2021, Chen et al. 2024) and circumvent the omission of unobserved scenarios encountered by the ϕ divergence-based ambiguity set (Ben-Tal et al. 2013, Gao and Kleywegt 2023). Let $\hat{\boldsymbol{s}}_{\omega}$ denote the empirical service times for scenario $\omega \in \Omega$, where $\Omega = \{1, 2, \ldots, N\}$ represents all possible scenarios. Then the empirical distribution of the service time is

$$\mathbb{P}^{\dagger}\left[\tilde{\boldsymbol{s}}^{\dagger} = \hat{\boldsymbol{s}}_{\omega}\right] = \frac{1}{N}, \quad \forall \omega \in \Omega. \tag{5}$$

The Wasserstein ambiguity set, defined with a radius $\theta \in \mathbb{R}_{+}$, is given by:

$$F_{\tilde{\boldsymbol{s}}}(\theta) = \left\{ \mathbb{P} \in \mathcal{P}(\mathcal{W}) \,\middle|\, \begin{array}{l} \tilde{\boldsymbol{s}} \sim \mathbb{P}, \tilde{\boldsymbol{s}}^{\dagger} \sim \mathbb{P}^{\dagger} \\ d_{\mathcal{W}}\left(\mathbb{P}, \mathbb{P}^{\dagger}\right) \leq \theta \end{array} \right\}, \tag{6}$$

where $\mathcal{W} = \{\boldsymbol{s} \mid \boldsymbol{s} \geq \underline{\boldsymbol{s}}\}$ is the corresponding support set with the lower bounds $\underline{\boldsymbol{s}}$ and

$$d_{\mathcal{W}}\left(\mathbb{P}, \mathbb{P}^{\dagger}\right) = \inf\left\{ \int_{\Omega} ||\boldsymbol{s}_1 - \boldsymbol{s}_2||_p \Pi(d\boldsymbol{s}_1, d\boldsymbol{s}_2) : \begin{array}{l} \Pi \text{ is a joint distribution of } \boldsymbol{s}_1 \text{ and } \boldsymbol{s}_2 \\ \text{with marginals } \mathbb{P} \text{ and } \mathbb{P}^{\dagger} \text{ respectively} \end{array} \right\}, \tag{7}$$

where $||\cdot||_p$ represents a polynomial norm for which $p \geq 1$. Note that the case of $\theta = 0$ ignores distributional ambiguity and corresponds to the SAA approach.

Conclusively, we obtain the compound ambiguity set $F_{\tilde{\boldsymbol{t}}}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \times F_{\tilde{\boldsymbol{s}}}(\theta)$ through the Cartesian product. Our numerical analysis (see Section 6.3) indicates that using this compound ambiguity set for disjoint travel and service times can effectively enhance robustness compared with only employing a single ambiguity set for embedded uncertainties.

4.2 Compound Set Reliability Index

The service fulfillment risk index (SRI) is an effective decision criterion for evaluating both the probability and magnitude of lateness, where the uncertainty is governed by a single ambiguity set as the service times are assumed to be implicitly included in the travel times (Zhang et al. 2021). Building on this concept, we extend it and obtain a new risk measure, the CSRI, by explicitly incorporating the service time as an independent multivariate random variable and constructing a compound ambiguity set for disjoint travel and service times. We show that the CSRI could enhance precision in risk assessment theoretically and numerically, but also lead to new solution challenges, as Section 4.3 elaborates.

Given a vehicle routing solution $\boldsymbol{z}$, we can extract the corresponding route(s) in the form of node sequence(s). Consider a route $(i, j_1, j_2, \ldots, j_{k-1}, j_k, \ldots, i')$ that visits node $v = j_k \in \mathcal{V}$ and ends at dummy depot i' (i.e., a duplicate of i). Regarding the partial path to node v, we define the sets of visited nodes and traversed arcs as

$$\mathcal{N}_v(\boldsymbol{z}) = \{i, j_1, j_2, \ldots, j_{k-1}, v\} \quad \text{and} \tag{8}$$

$$\mathcal{A}_v(\boldsymbol{z}) = \{(i, j_1), (j_1, j_2), \ldots, (j_{k-1}, v)\}, \tag{9}$$

respectively. For each node $v \in \mathcal{V}$, corresponding to $\mathcal{N}_v(\boldsymbol{z})$ and $\mathcal{A}_v(\boldsymbol{z})$, let us define $\boldsymbol{\pi}_v(\boldsymbol{z})$ and $\boldsymbol{\varsigma}_v(\boldsymbol{z})$ as 0-1 vectors such that $\pi_v^j(\boldsymbol{z}) = 1$ if and only if node $j \in \mathcal{N}_v(\boldsymbol{z})$, and $\varsigma_v^a(\boldsymbol{z}) = 1$ if and only if arc $a \in \mathcal{A}_v(\boldsymbol{z})$. Similarly, we define the set of traversed nodes and arcs (starting) from an upstream node $v' = j_{k'} \in \mathcal{N}_v(\boldsymbol{z})$ to v as

$$\mathcal{N}_{v'v}(\boldsymbol{z}) = \{v', j_{k'+1}, \ldots, j_{k-1}\} \quad \text{and} \tag{10}$$

$$\mathcal{A}_{v'v}(\boldsymbol{z}) = \mathcal{A}_v(\boldsymbol{z}) \backslash \mathcal{A}_{v'}(\boldsymbol{z}) = \{(v', j_{k'+1}), (j_{k'+1}, j_{k'+2}), \ldots, (j_{k-1}, v)\}, \tag{11}$$

respectively, and $\boldsymbol{\pi}_{v'v}(\boldsymbol{z})$ and $\boldsymbol{\varsigma}_{v'v}(\boldsymbol{z})$ are the corresponding 0-1 vectors as we defined above but acting on $\mathcal{N}_{v'v}(\boldsymbol{z})$ and $\mathcal{A}_{v'v}(\boldsymbol{z})$. It is important to note that $\mathcal{N}_v(\boldsymbol{z})$ includes node v, whereas $\mathcal{N}_{v'v}(\boldsymbol{z})$ does not. We now introduce the calculation of *service start time* and *delay function*, which are widely adopted in the VRPTW literature (see e.g., Zhang et al. 2019, 2021). The separation of travel and service times can be incorporated straightforwardly as follows.

LEMMA 1. *Given a routing solution $\boldsymbol{z}$, a realization $\boldsymbol{s}$ of service times and a realization $\boldsymbol{t}$ of travel times, the service start time for each node $v \in \mathcal{V}$ is determined by the function*

$$S_v(\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{t}) = \max_{v' \in \mathcal{N}_v(\boldsymbol{z})} \left\{ e_{v'} + \sum_{j \in \mathcal{N}_{v'v}(\boldsymbol{z})} s_j + \sum_{a \in \mathcal{A}_{v'v}(\boldsymbol{z})} t_a \right\}. \tag{12}$$

We present the proof in EC.3.1. Following this, we introduce the delay function, which plays a pivotal role in characterizing the CSRI metric.

DEFINITION 2. Given the service start time, the delay function upon a node $v \in \mathcal{V}$ is defined as

$$\xi_v(\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{t}) = S_v(\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{t}) - l_v. \tag{13}$$

Clearly, the delay function captures the temporal difference between the realized service start time and the latest time window as promised. By analyzing such discrepancies across all travel and service time realizations, a comprehensive view of the delay emerges. However, as the distributions of travel times $\tilde{\boldsymbol{t}}$ and service times $\tilde{\boldsymbol{s}}$ are unobservable by nature, the delay $\tilde{\xi}(\boldsymbol{z}, \tilde{\boldsymbol{s}}, \tilde{\boldsymbol{t}})$ is also a random variable[3].

[3] For clarity in our exposition, we denote random variables with a tilde ($\tilde{}$) symbol. In contrast, variables without a symbol or variables with a hat ($\hat{}$) symbol, represent their realized values. We denote the empirical distribution $\mathbb{P}^\dagger$ with a dagger superscript (†). For example, $\tilde{\boldsymbol{s}}^\dagger$ denotes the random variable governed by the empirical distribution $\mathbb{P}^\dagger$ of service times, whereas $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$ represent a specific realization of these service times.

DEFINITION 3. (Compound set reliability index, CSRI) Given a random delay denoted by random variable $\tilde{\xi}$ with probability distribution $\mathbb{P}$ and a service level $\gamma \in [0,1]$, we define the CSRI as

$$\rho_\gamma(\tilde{\xi}) = \min\left\{\alpha \geq 0 \mid \text{F-CVaR}_\gamma(\max\{\tilde{\xi}, -\alpha\}) \leq 0\right\}, \tag{14}$$

where $\text{F-CVaR}_\gamma(\tilde{\xi})$ is the worst-case *conditional value-at-risk* (CVaR) for random variable $\tilde{\xi}$ over the compound ambiguity set F of distribution $\mathbb{P}$:

$$\text{F-CVaR}_\gamma(\tilde{\xi}) = \min_\alpha \left\{\alpha + \frac{1}{1-\gamma}\sup_{\mathbb{P}\in F}\mathbb{E}_\mathbb{P}\left[(\tilde{\xi}-\alpha)^+\right]\right\}. \tag{15}$$

Intuitively, the CSRI of each node is the minimum nonnegative constant α such that the F-CVaR_γ of the *regularized delay* does not exceed 0, where the regularized delay is acquired by forcing the excess early arrival time beyond α equal to α, that is, let $\tilde{\xi} = -\alpha$ when $\tilde{\xi} \leq -\alpha$. Given the CSRI definition, we aim to answer the following questions.

First, *what do we lose if we implicitly embed service time uncertainty in travel time counterparts?* From the robust optimization standpoint, the worst-case delay is less encapsulated for the joint ambiguity set of the sum of travel and service times than the compound ambiguity set. This distinction arises from the subadditivity property of the supremum operator. That is,

$$\sup_{\hat{\boldsymbol{s}}\in\mathcal{W}_{\boldsymbol{s}},\,\hat{\boldsymbol{t}}\in\mathcal{W}_{\boldsymbol{t}}} \xi(z,\hat{\boldsymbol{s}},\hat{\boldsymbol{t}}) \geq \sup_{\hat{\boldsymbol{t}}\in\mathcal{W}_{\boldsymbol{s}+\boldsymbol{t}}} \xi(z,\mathbf{0},\hat{\boldsymbol{t}}), \tag{16}$$

where $\mathcal{W}_{\boldsymbol{t}}$ and $\mathcal{W}_{\boldsymbol{s}}$ denotes the data-driven support set for travel and service times, respectively, and $\mathcal{W}_{\boldsymbol{s}+\boldsymbol{t}}$ signifies the counterpart with the assumption that the service time at each node is included in the travel time of its outgoing arcs. This underscores the imperative to handle uncertain travel and service times separately in the context of robust optimization. We illustrate this result by using an example constructed on the Wasserstein ambiguity set.

EXAMPLE 1. We reveal the distinction between the compound Wasserstein ambiguity sets based on separate and independent travel and service times versus the single ambiguity set for their sum. A comparison of these two methods can be found in Table 1, where different on-time metrics are analyzed under $\theta = 0$ and $\gamma = 0.1$ without loss of generality. Notably, despite the duration samples (columns 'Uncertain deviation') being identical, the single and compound ambiguity sets present remarkably divergent results. Specifically, the former over the empirical distribution $\mathbb{P}(\xi) = \frac{2}{3}\delta(\xi+1) + \frac{1}{3}\delta(\xi+4)$, where $\delta(\cdot)$ denotes the Dirac distribution, manifests a pronounced reliability of the current schedule, with all the metrics indicating no delay risks (columns 'Single'). It seems that the potential delays are entirely avoided. However, this evaluation is obviously impractical, as scenarios with delays can easily be identified. For example, if the uncertain deviations for both travel and service times are 5 minutes, as historically recorded, then a worst-case delay of 10 minutes would be observed. In contrast to the "satisfactory" outcome obtained through the single ambiguity set, metrics derived from the compound ambiguity set raise serious concerns about punctuality (columns 'Compound'). As shown in Table 1, the lateness probability is as high as 33%, with an expected delay of 1.78 minutes. The result indicates poor on-time performance, which is not perceived as an acceptable outcome. This is because the compound ambiguity set successfully captures the delay records for separate travel and service times (e.g., the 5-minute delay for service time in scenario 1), offering more risk assessment precision than the one implicitly blending them. ■

Table 1 Comparison between the Single and Compound Wasserstein Ambiguity Sets

	Uncertain deviation			**Single**				**Compound**			
Scenario	Service	Travel	Total	Lateness probability(%)	Expected delay	Worst-case delay	SRI	Lateness probability(%)	Expected delay	Worst-case delay	CSRI
1	5	−6	−1								
2	−6	5	−1	0	0	0	0	33	1.78	10	4.76
3	−2	−2	−4								

Second, *why are CSRI constraints effective in addressing the cascading effect?* To elucidate this, we analyze how the CSRI constraints function via another illustrative example. As mentioned previously, the literature has not yet fully investigated or addressed the cascading effect. For example, by degenerating the compound ambiguity set into a single Wasserstein ambiguity set, Zhang et al. (2021) focused on minimizing the total index metric of customer nodes within a specified travel cost budget. Similar research ventures have considered other index metrics (e.g., Jaillet et al. 2016, Zhang et al. 2019, 2024). The primary objective of these studies has been to control the overall delay risk, which is a crucial concern for practitioners, but the reason for the cascading effect has not been discussed. In other words, little attention is directed toward balancing the CSRI for managing individual node-level delays. While the overall CSRI values may exhibit trivial differences, different formulations for integrating CSRI (i.e., minimizing total CSRI versus restricting CSRI for individual nodes) can lead to varying robustness levels for individual nodes.

EXAMPLE 2. We illustrate the service-level fairness of restricting the CSRI for each patient in Figure 3. The figure elucidates three potential routes, each characterized by distinct uncertain delay distributions (i.e., uncertain delay realization and probability) for each patient. Here, we use τ to represent a large constant value and $\gamma = 0.1$. In terms of the objective of minimizing the overall CSRI, route 2 emerges as the top choice, boasting the lowest total CSRI value of 1.18. However, in the worst-case scenario, patient *A* on route 2 might endure a waiting time of 20 minutes (with a probability of 0.05), which is unacceptable and bound to cause dissatisfaction and complaints. Conversely, the maximum waiting time on routes 1 and 3 is only 10 minutes, which is considered trivial. Route 1, in particular, stands out as it optimally distributes the CSRI, ensuring timely service for each patient. ■

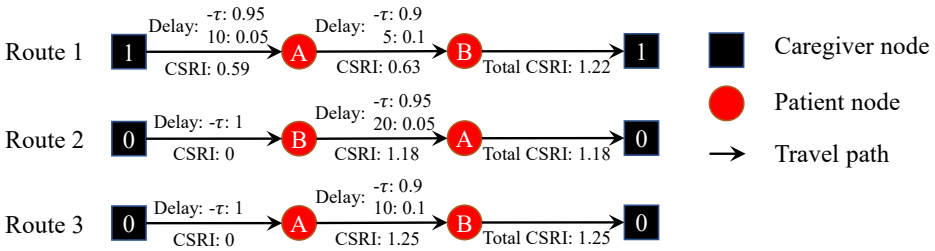


Figure 3 Comparison of Delay Mitigation Strategies: Individual Restriction versus Total Minimization

This example clearly demonstrates that setting a cap on the CSRI for each node is crucial and necessary to prevent situations from deteriorating beyond the prescribed criterion and boost the satisfaction of both patients and caregivers. The implementation of CSRI constraints is deemed an effective tool to mitigate cascading delays caused by the convolutional travel and service uncertainties, which distinguishes this work from Zhang et al. (2021). Therefore, we incorporate this measure into our formulation and explore its closed forms in the following subsection.

4.3 Evaluating the CSRI in Closed Forms

To embody the left side of constraints (2), namely the CSRI value, we reformulate $\rho_\gamma(\tilde{\xi}_v(z,\tilde{s},\tilde{t}))$ for node $v \in \overline{\mathcal{V}}$ under a given routing solution z as the following semi-infinite programming problem (Zhang et al. 2021, Theorem 1).

$$[\mathcal{F}_{\text{CSRI-SIP}}] \quad \rho_\gamma(\tilde{\xi}_v(z,\tilde{s},\tilde{t})) = \min \alpha \tag{17a}$$

$$\text{s.t.} \sup_{\mathbb{P}\in F_{\tilde{t}}(\mu,\Sigma)\times F_{\tilde{s}}(\theta)} \mathbb{E}_{\mathbb{P}}\left[\left(\tilde{\xi}_v(z,\tilde{s},\tilde{t})+\alpha\right)^+\right] \le (1-\gamma)\,\alpha, \tag{17b}$$

$$\alpha \ge 0. \tag{17c}$$

Note that $\mathcal{F}_{\text{CSRI-SIP}}$ is intractable because of the complexity related to high-dimensional integration. To address such a technical challenge, in this section, we develop three different CSRI solution approaches by exploiting the structure and deriving tractable reformulations. Specifically, we introduce the SAA approach $\Phi_{\text{CSRI-SAA}}$ in Section 4.3.1 and the exact CSRI approach $\Phi_{\text{CSRI-EXA}}$ as well as the sufficient CSRI acceleration approach $\Phi_{\text{CSRI-SUF}}$ in Section 4.3.2 to obtain the CSRI value for each node given the preceding partial path under the assumption of implicit time windows. These approaches can easily be adapted for CSRI-feasibility checks and are applicable to any exact or heuristic algorithm. In Section 5, we develop an exact and metaheuristic algorithm integrated with these CSRI-tailored closed forms to ensure on-time service and completion requirements.

4.3.1 Sample Average Approximation

SAA is a simulation-based approach proposed by Kleywegt et al. (2002) to handle stochastic discrete optimization problems through the basic idea that the expected objective value of the stochastic problem can be approximated by the corresponding average value of sampling problems. Owing to the intractability of calculating the sum of random variables, we reformulate $\mathcal{F}_{\text{CSRI-SIP}}$ with the SAA method into the following linear programming (LP) model where the ambiguity set is a singleton that contains only the empirical distribution.

$$
\begin{aligned}
[\mathcal{F}_{\text{CSRI-SAA}}] \quad & \min \alpha && && (18a)\\
& \text{s.t.} \ \frac{1}{N} \sum_{\omega \in \Omega} y_\omega \leq (1-\gamma)\, \alpha, && && (18b)\\
& y_\omega \geq \xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_\omega, \hat{\boldsymbol{t}}_\omega) + \alpha, && \forall \omega \in \Omega, && (18c)\\
& y_\omega \geq 0, && \forall \omega \in \Omega, && (18d)\\
& \alpha \geq 0, && && (18e)
\end{aligned}
$$

where $\hat{\boldsymbol{s}}_\omega$ and $\hat{\boldsymbol{t}}_\omega$ denote the ω-th realization of $\tilde{\boldsymbol{s}}^\dagger$ and $\tilde{\boldsymbol{t}}^\dagger$, respectively, and each scenario $\omega \in \Omega$ occurs with even probability. The auxiliary decision variables $y_\omega = (\xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_\omega, \hat{\boldsymbol{t}}_\omega) + \alpha)^+, \omega \in \Omega$ are introduced to linearize the reformulation. $\mathcal{F}_{\text{CSRI-SAA}}$ can be solved directly via state-of-the-art commercial solvers such as CPLEX and Gurobi and adapted in column generation to check route feasibility. We thus represent this evaluation procedure as the $\Phi_{\text{CSRI-SAA}}$ approach.

4.3.2 Closed-Form Evaluation

The SAA method estimates the objective value via the empirical distribution, assigning an equal mass of $1/N$ to each historical sample. However, this approach may lead to estimation errors and yield inferior decisions, which perform poorly in practice. To leverage the advantages of CSRI constraints over the compound ambiguity set and solve the $\mathcal{F}_{\text{CSRI-SIP}}$ for all nodes in $\overline{\mathcal{V}}$, we utilize duality theory and obtain its tractable reformulation as follows.

THEOREM 1. *$\mathcal{F}_{\text{CSRI-SIP}}$ can be equivalently represented as the following optimization problem with second-order conic constraints.*

$$
\begin{aligned}
[\mathcal{F}_{\text{CSRI-SOC}}] \quad & \min \alpha && && (19a)\\
& s.t.\ \theta\, |\mathcal{A}_v(\boldsymbol{z})|^{\frac{p-1}{p}} + \frac{1}{N} \sum_{\omega \in \Omega} \alpha_\omega \leq (1-\gamma)\alpha, && && (19b)\\
& \quad \boldsymbol{\varsigma}_{v'v}(\boldsymbol{z})^\top \boldsymbol{\Sigma} \boldsymbol{\varsigma}_{v'v}(\boldsymbol{z}) \leq 4\alpha_\omega \left(\alpha_\omega - \alpha - e_{v'} + l_v - \boldsymbol{\varsigma}_{v'v}(\boldsymbol{z})^\top \boldsymbol{\mu} - \boldsymbol{\pi}_{v'v}(\boldsymbol{z})^\top \hat{\boldsymbol{s}}_\omega\right), && \forall \omega \in \Omega, v' \in \mathcal{N}_v(\boldsymbol{z}), && (19c)\\
& \quad \alpha_\omega \in \mathbb{R}_+, && \forall \omega \in \Omega, && (19d)\\
& \quad \alpha \in \mathbb{R}_+, && && (19e)
\end{aligned}
$$

where α_ω is the auxiliary decision variable for each scenario $\omega \in \Omega$.

The proof of Theorem 1 follows from the concurrent duality over the two ambiguity sets and subsequent reformulations, which subtly integrate the duality results of the Wasserstein distance-based ambiguity set in Zhang et al. (2021) and that of the cross moment ambiguity set in Zhang et al. (2019). Note that Theorem 1 indicates that the CSRI of a node under a given routing solution can be addressed via a conic program comprising at least $O(N)$ decision variables and $O(N)$ constraints even if $e_{v'} = 0$ for all $v' \in \mathcal{V}$, which is of the same scale as $\mathcal{F}_{\text{CSRI-SAA}}$ and thus tends to be computationally demanding. In other words, it is computationally prohibitive to solve the compact formulation by substituting the CSRI constraints with $\mathcal{F}_{\text{CSRI-SOC}}$ and solving the resulting mixed-integer second-order conic program using state-of-art techniques (Drewes and Ulbrich 2009), as the basic VRP is still NP-hard. To accelerate the procedure, we introduce the following proposition to highlight the properties of this conic program given graph $\overline{G} = (\overline{\mathcal{V}}, \mathcal{A})$.

PROPOSITION 1. *For each node $v \in \overline{\mathcal{V}}$, the solution of $\mathcal{F}_{\text{CSRI-SOC}}$ must fulfill the following constraints:*

$$\frac{\theta \left|\mathcal{A}_v(\boldsymbol{z})\right|^{\frac{p-1}{p}} + \frac{1}{N}\sum_{\omega\in\Omega}\alpha_\omega}{1-\gamma} = \alpha = \alpha_\omega - \frac{\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top \boldsymbol{\Sigma} \boldsymbol{\varsigma}_v(\boldsymbol{z})}{4\alpha_\omega} - \xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_\omega, \boldsymbol{\mu}), \quad \forall \omega \in \Omega. \tag{20}$$

Proposition 1 provides the optimal conditions of the CSRI solution, which supports us in developing the following CSRI evaluation approaches. Before formalizing the theorem, we first denote the CVaR of the delay given the empirical distribution of service times and the expected travel times as $\text{CVaR}_\gamma(\tilde{\xi}_v(\boldsymbol{z}, \tilde{\boldsymbol{s}}^\dagger, \boldsymbol{\mu}))$. With the result presented in Sarykalin et al. (2008), we can compute $\text{CVaR}_\gamma(\tilde{\xi}_v(\boldsymbol{z}, \tilde{\boldsymbol{s}}^\dagger, \boldsymbol{\mu}))$ as

$$\text{CVaR}_\gamma(\tilde{\xi}_v(\boldsymbol{z}, \tilde{\boldsymbol{s}}^\dagger, \boldsymbol{\mu})) = \sum_{\omega=1}^{\lfloor(1-\gamma)N\rfloor} \frac{\xi_v\left(\boldsymbol{z}, \hat{\boldsymbol{s}}_{(\omega)}, \boldsymbol{\mu}\right)}{(1-\gamma)N} + \left(1 - \frac{\lfloor(1-\gamma)N\rfloor}{(1-\gamma)N}\right) \times \xi_v\left(\boldsymbol{z}, \hat{\boldsymbol{s}}_{(\lfloor(1-\gamma)N\rfloor+1)}, \boldsymbol{\mu}\right), \tag{21}$$

where $\xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_{(\omega)}, \boldsymbol{\mu})$ represents the decreasing *order statistics* of the empirical delays such that $\xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_{(1)}, \boldsymbol{\mu}) \geq \xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_{(2)}, \boldsymbol{\mu}) \geq \ldots \geq \xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_{(N)}, \boldsymbol{\mu})$.

THEOREM 2. *(Sufficient CSRI) The CSRI in constraints* (2) *of node $v \in \overline{\mathcal{V}}$ under a given routing solution $\boldsymbol{z}$ can be equivalently reformulated as*

$$[\mathcal{F}_{\text{CSRI-SUF}}] \quad \min \alpha \tag{22a}$$

$$s.t.\ \theta \left|\mathcal{A}_v(\boldsymbol{z})\right|^{\frac{p-1}{p}} + \frac{1}{N}\sum_{\omega\in\Omega} \frac{(\alpha + \xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_\omega, \boldsymbol{\mu})) + \sqrt{(\alpha + \xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_\omega, \boldsymbol{\mu}))^2 + \boldsymbol{\varsigma}_v(\boldsymbol{z})^\top \boldsymbol{\Sigma} \boldsymbol{\varsigma}_v(\boldsymbol{z})}}{2} \leq (1-\gamma)\alpha, \tag{22b}$$

$$\alpha \in \mathbb{R}_+, \tag{22c}$$

Furthermore, if $\text{CVaR}_\gamma(\tilde{\xi}_v(\boldsymbol{z}, \tilde{\boldsymbol{s}}^\dagger, \boldsymbol{\mu})) \leq -\Gamma_v(\boldsymbol{z})$*, then constraints* (2) *are satisfied when*

$$\max\left\{\max_{n\in\{1,2,\ldots,\lfloor(1-\gamma)N\rfloor\}}\left\{\frac{\sum_{\omega=1}^{n}\xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_{(\omega)}, \boldsymbol{\mu}) + (1-\gamma)N\Gamma_v(\boldsymbol{z})}{(1-\gamma)N - n}\right\}, \Gamma_v(\boldsymbol{z})\right\} \leq \beta, \tag{23}$$

where $\Gamma_v(\boldsymbol{z}) = \frac{2\theta\left|\mathcal{A}_v(\boldsymbol{z})\right|^{\frac{p-1}{p}} + \sqrt{\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top \boldsymbol{\Sigma} \boldsymbol{\varsigma}_v(\boldsymbol{z})}}{2(1-\gamma)}$.

Note that $\mathcal{F}_{\text{CSRI-SUF}}$ favourably harbors both modeling fidelity and computational traceability. Compared with the same measure index under the Wasserstein ambiguity set for joint travel and service times in Zhang et al. (2021), $\mathcal{F}_{\text{CSRI-SUF}}$ fully captures the structure of the variance-covariance matrix of travel times with the term $\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top \boldsymbol{\Sigma} \boldsymbol{\varsigma}_v(\boldsymbol{z})$, avoiding the computational intractability or the simplified assumption that correlation exists only between adjacent links for computational convenience (Rostami et al. 2021). If travel times are deterministic as in the mean values, then constraint (22b) degenerates into the form $\theta \left|\mathcal{A}_v(\boldsymbol{z})\right|^{\frac{p-1}{p}} + \frac{1}{N}\sum_{\omega\in\Omega}(\alpha + \xi_v(\boldsymbol{z}, \hat{\boldsymbol{s}}_\omega, \boldsymbol{\mu}))^+ \leq (1-\gamma)\alpha$, which is exactly the variant in Zhang et al. (2021). In this regard, $\mathcal{F}_{\text{CSRI-SUF}}$ imposes a more rigorous criterion for measuring travel time correlation.

The left-hand side of Eq. (23) serves as a conservative (larger) approximation of the CSRI value. The advantage of this sufficient condition is that it can be assessed easily in a constant time. Nonetheless, this condition is not a necessary condition for the fulfillment of $\rho_\gamma(\tilde{\xi}_v(\boldsymbol{z},\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}}))$. In other words, it is possible to construct problem instances that satisfy constraints (2) but violate Eq. (23). Our preliminary experimental results show that the approach is less computationally intensive but overly conservative such that a portion of feasible routes are ruled out. In light of these observations, we apply Eq. (23) first in column generation as a preliminary check to accelerate CSRI-feasibility evaluation by avoiding exact examination with a larger computational burden, which is denoted as the $\Phi_{\text{CSRI-SUF}}$ approach hereafter. However, this approach may mistakenly exclude some feasible routes as a sufficient condition for constraints (2). To check the CSRI feasibility of each route exactly, we develop the sufficient and necessary conditions as follows.

THEOREM 3. *(Exact CSRI) Suppose that $e_{v'}=0$ for $\forall v'\in\mathcal{V}$, we have $\rho_\gamma(\tilde{\xi}_v(\boldsymbol{z},\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}}))<+\infty$ if and only if there exists an $\alpha^*\in\mathbb{R}_+$ such that*

$$\gamma-\frac{1}{2}+\frac{1}{N}\sum_{\omega\in\Omega}\frac{\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})}{2\sqrt{\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})^2+\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top\boldsymbol{\Sigma}\boldsymbol{\varsigma}_v(\boldsymbol{z})}}\le 0, \tag{24}$$

$$\gamma-\frac{1}{2}+\frac{1}{N}\sum_{\omega\in\Omega}\frac{\alpha^*+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})}{2\sqrt{\left(\alpha^*+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})\right)^2+\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top\boldsymbol{\Sigma}\boldsymbol{\varsigma}_v(\boldsymbol{z})}}=0, \tag{25}$$

$$\theta\left|\mathcal{A}_v(\boldsymbol{z})\right|^{\frac{p-1}{p}}+\frac{1}{N}\sum_{\omega\in\Omega}\frac{\alpha^*+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})+\sqrt{\left(\alpha^*+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})\right)^2+\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top\boldsymbol{\Sigma}\boldsymbol{\varsigma}_v(\boldsymbol{z})}}{2}+(\gamma-1)\alpha^*\le 0, \tag{26}$$

which serve as the sufficient and necessary conditions of the existence of $\rho_\gamma(\tilde{\xi}_v(\boldsymbol{z},\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}}))$. If $\rho_\gamma(\tilde{\xi}_v(\boldsymbol{z},\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}}))<+\infty$, then constraints (2) hold if one of the following two inequalities is satisfied:

$$\gamma-\frac{1}{2}+\frac{1}{N}\sum_{\omega\in\Omega}\frac{\beta+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})}{2\sqrt{\left(\beta+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})\right)^2+\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top\boldsymbol{\Sigma}\boldsymbol{\varsigma}_v(\boldsymbol{z})}}\ge 0, \tag{27}$$

$$\theta\left|\mathcal{A}_v(\boldsymbol{z})\right|^{\frac{p-1}{p}}+\frac{1}{N}\sum_{\omega\in\Omega}\frac{\beta+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})+\sqrt{\left(\beta+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})\right)^2+\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top\boldsymbol{\Sigma}\boldsymbol{\varsigma}_v(\boldsymbol{z})}}{2}+(\gamma-1)\beta\le 0. \tag{28}$$

Note that the check of condition (24) is trivial. However, the expression of α^* by known parameters is impracticable. Fortunately, we can utilize gradient descent to evaluate the CSRI existence and a subsequent bisection algorithm to obtain the exact CSRI value efficiently. Specifically, consider a function $f:\mathbb{R}_+\mapsto\mathbb{R}$, given by

$$f(\alpha)=\theta\left|\mathcal{A}_v(\boldsymbol{z})\right|^{\frac{p-1}{p}}+\frac{1}{N}\sum_{\omega\in\Omega}\frac{\alpha+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})+\sqrt{\left(\alpha+\xi_v(\boldsymbol{z},\hat{\boldsymbol{s}}_\omega,\boldsymbol{\mu})\right)^2+\boldsymbol{\varsigma}_v(\boldsymbol{z})^\top\boldsymbol{\Sigma}\boldsymbol{\varsigma}_v(\boldsymbol{z})}}{2}+(\gamma-1)\alpha. \tag{29}$$

Since the function $f(\alpha)$ is convex (see EC.3.6 for proof details), the α^* can thus be easily obtained through iterative descent searches. That is, we start with an initial point $\alpha_0=0$ and successively generate a list of points $\alpha_1,\alpha_2,\ldots$ via Newton's method such that f is decreased in each iteration. Once a nonpositive f value is obtained, the remaining conditions (25)-(26) are guaranteed. Then, the bisection method is utilized on the last iterative interval to find the value of CSRI with $f(\rho_\gamma(\tilde{\xi}_v(\boldsymbol{z},\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}})))=0$. The above process is described in Algorithm 1 in EC.4 and is denoted as $\Phi_{\text{CSRI-EXA}}$ for the CSRI feasibility check.

4.4 Comparison with the Lateness Probability Measure

When measuring on-time service and completion in routing optimization, the lateness probability index (LPI) is perhaps the most natural and ubiquitous decision criterion to improve scheduling resiliency (Adulyasak and Jaillet 2016). For example, in our case, this approach involves considering the same setting of $\mathcal{F}_{\text{SPF}}$ but replacing the CSRI constraints (2) with the LPI constraints as the measure of punctuality. In this section, we theoretically illustrate that the CSRI is more sensitive to the magnitude of delays than the general LPI measure. We first formally present the LPI definition concerning the compound ambiguity set as follows.

DEFINITION 4. (Lateness probability index under the compound ambiguity set, LPI_c) Given a random delay denoted by the random variable $\tilde{\xi}$ with probability distribution $\mathbb{P}$, we define the LPI_c as

$$\chi(\tilde{\xi}) = \min\left\{\alpha \geq 0 \mid \sup_{\mathbb{P}\in F_{\tilde{s}}(\theta)\times F_{\tilde{t}}(\boldsymbol{\mu},\boldsymbol{\Sigma})} \mathbb{P}\left(\tilde{\xi} > 0\right) \leq \alpha\right\}. \tag{30}$$

LEMMA 2. *Given routing solution $\boldsymbol{z}$, the LPI_c for node $v \in \overline{\mathcal{V}}$ can be obtained via the semi-infinite program as follows.*

$$[\mathcal{F}_{\textit{LPI-SIP}}] \quad \min \alpha \tag{31a}$$

$$s.t. \sup_{\mathbb{P}\in F_{\tilde{\boldsymbol{t}}}(\boldsymbol{\mu},\boldsymbol{\Sigma})\times F_{\tilde{\boldsymbol{s}}}(\theta)} \text{F-CVaR}_{1-\alpha}(\tilde{\xi}_v(\boldsymbol{z},\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}})) \leq 0, \tag{31b}$$

$$\alpha \geq 0. \tag{31c}$$

Intuitively, the LPI_c calculates the least quantile to curb the conditional expectation of lateness and provides decision-makers with the flexibility to customize service levels in terms of probabilistic guarantees of on-time service and completion. Although the same objective, risk threshold and CVaR are adopted to quantify the tail lateness, the CSRI procures the least normalized value for regularizing the excess early arrival time instead of the least quantile as the LPI_c. As a result, the CSRI is more sensitive than the LPI_c to the *magnitude* of lateness rather than early arrival. For example, the LPI_c might prioritize a delay with a lower probability but notably longer duration, as long as the lateness magnitude is negligible compared with the early arrival. The CSRI comprehensively quantifies both the lateness magnitude and probability, thereby avoiding such a dilemma. Furthermore, contrary to the computational intractability for $\mathcal{F}_{\text{LPI-SIP}}$, $\mathcal{F}_{\text{CSRI-SIP}}$ allows concise reformulations in Section 4.3, which enables us to develop efficient solution algorithms in Section 5.

5 Solution Approaches

The challenges for addressing $\mathcal{F}_{\text{SPF}}$ are twofold. On the one hand, the CSRI constraints (2), which can be computed through the evaluation approaches in Section 4.3, i.e., $\Phi_{\text{CSRI-SAA}}$, $\Phi_{\text{CSRI-EXA}}$, $\Phi_{\text{CSRI-SUF}}$, are not linear constraints; thus, $\mathcal{F}_{\text{SPF}}$ cannot be directly solved by off-the-shelf solvers such as CPLEX and Gurobi. On the other hand, even excluding the CSRI constraints, solvers can rarely provide optimal solutions for medium- or large-size instances in a reasonable computing time, because the reduced problem, as a variant of VRP, is still a well-known NP-hard problem.

To address these concerns and solve RSRP instances more efficiently, this section designs an exact algorithm (`CSRI-BPC`) based on the $\mathcal{F}_{\text{SPF}}$ formulation and a metaheuristic method (`CSRI-VNS`). Specifically, `CSRI-BPC` can solve the problem to proven optimality, whereas `CSRI-VNS` achieves high-quality solutions efficiently. Moreover, `CSRI-BPC` can be enhanced with `CSRI-VNS` to achieve fast-effective performance. For conciseness, **we provide only the main components for each approach here and present more details in EC.5-EC.6.**

5.1 Main Components of CSRI-BPC

Branch-price-and-cut (BPC) is a method of combinatorial optimization to solve MIP models where both column generation (CG) and the separation of cutting planes are exploited simultaneously to solve the restricted master problem at each node of the branch-and-bound tree (Costa et al. 2019).

5.1.1 Master problem.

The master problem (MP) has been presented in the $\mathcal{F}_{\text{SPF}}$ formulation. To solve this integer programming problem, we first relax $\mathcal{F}_{\text{SPF}}$ to the LP model by setting the decision variables $\boldsymbol{z}$ to be continuous. Additionally, the cuts (see Section 5.1.3) and branching rules (see Section 5.1.5) are applied to this relaxed MP to obtain integer solutions efficiently. The restricted master problem (RMP) is derived by replacing $\tilde{\mathcal{R}}$ with a subset $\hat{\mathcal{R}} \subseteq \tilde{\mathcal{R}}$ in the MP, which expands through iterative CG and finally converges to optimal solutions.

5.1.2 Pricing subproblem.

Let ψ_i $(i \in I)$ and ψ_j $(j \in \mathcal{J})$ represent the dual variables of constraints (1b) and constraints (1c), respectively. The reduced cost for a route $r \in \tilde{\mathcal{R}}$ is given by

$$\overline{C}_r = c_r - \sum_{i \in I} b_i^r \psi_i - \sum_{j \in \mathcal{J}} b_j^r \psi_j = \sum_{(v,v') \in \mathcal{A}} b_{vv'}^r \overline{c}_{vv'} \text{ (where } \overline{c}_{vv'} = c_{vv'} - \psi_v). \tag{32}$$

The binary parameter $b_{vv'}^r$ indicates whether an arc (v,v') is selected in the route $r \in \tilde{\mathcal{R}}$, which is generated through the label-setting algorithm. We can conveniently obtain a support graph $\mathcal{G}_i$ $(i \in I)$ with the modified cost $\overline{c}_{vv'}$ only on arcs by changing the values of its outgoing arcs according to Eq. (32). Therefore, the original problem of minimizing the reduced cost $\overline{C}_r$ can be converted to the shortest path problem under the capacity, skill-matching and CSRI constraints on graph $\mathcal{G}_i$. That is, the pricing subproblem is an NP-hard *elementary shortest path problem with resource constraints*.

Now we introduce the label-setting algorithm by only focusing on one support graph $\mathcal{G}_i$ to simplify the exposition, where a list of labels at each node is defined to represent the information of partial paths. More specifically, a label L_j is a tuple $L_j = (v_j, \overline{C}_j, q_j, \rho_j, (\beth_j^v)_{v \in \mathcal{V}})$ representing the partial path $r(L_j)$ originating from depot i and extending (directly/indirectly) through the network to node j $(j \in \mathcal{V})$. The elements in L_j are described as follows:

- v_j: the last vertex along path $r(L_j)$, $v_j = -1$ if j is the initial depot;
- $\overline{C}_j$: the reduced cost of path $r(L_j)$;
- q_j: the accumulated demand over path $r(L_j)$;
- ρ_j: the CSRI of node j along path $r(L_j)$;
- $\beth_j^v$: a binary variable. $\beth_j^v = 1$ if node v has been visited or unreachable from L_j; otherwise $\beth_j^v = 0$.

After label initialization at depot i by setting all elements to 0 except $v_i = -1$, new labels are generated iteratively by performing forward extensions along feasible arcs through *resource extension functions* (REFs). The feasibility of extending each selected node is determined according to the skill-matching, CSRI and capacity constraints, thereby excluding infeasible nodes from the paths. Specifically, let $\oplus$ be the concatenation symbol, and the extension of path $r(L_j)$ along arc (j,k) to obtain new path $r(L_k) = r(L_j) \oplus (j,k)$ is performed as the following REFs:

$$v_k = j;\ \overline{C}_k = \overline{C}_j + \overline{c}_{jk};\ q_k = q_j + d_k; \tag{33a}$$

$$\rho_k \leftarrow \rho_\gamma\Big(\tilde{\xi}_k\big(r(L_k), \tilde{\boldsymbol{s}}, \tilde{\boldsymbol{t}}\big)\Big); \tag{33b}$$

$$\beth_k^v = \begin{cases} \beth_j^v + 1, & \text{if } v = k; \\ \max\{\beth_j^v, UR_v(L_k)\}, & \text{otherwise.} \end{cases} \tag{33c}$$

Here, $\tilde{\xi}_k(r(L_k),\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}})$ is the uncertain delay of node k along path $r(L_k)$ and $\rho_\gamma\Big(\tilde{\xi}_k\big(r(L_k),\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}}\big)\Big)$ is the CSRI value, which can be computed through the approaches (i.e., $\Phi_{\text{CSRI-SAA}}$, $\Phi_{\text{CSRI-EXA}}$, $\Phi_{\text{CSRI-SUF}}$) in Section 4.3. Furthermore, $UR_v(L_k)$ is a binary variable indicating whether the node v is unreachable from label L_k. Specifically, $UR_v(L_k)=1$ when at least one of the following conditions is violated.

- *Workload, fatigue and medical resources*: the sum of patient demands cannot exceed the caregiver's capacity, $q_k \le Q_i$;
- *Skill matching*: the caregiver possesses the specific skill that the next node v requires, $\upsilon_{im} \ge \upsilon_{vm}, \forall m \in \mathcal{M}$;
- *Time window*: the extension to the next node v satisfies the predefined risk level, $\rho_\gamma\Big(\tilde{\xi}_v\big(r(L_k)\oplus(k,v),\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}}\big)\Big) \le \beta$;
- *Working time regulation*: after completing the service at node v, the caregiver can directly return to the depot, $\rho_\gamma\Big(\tilde{\xi}_i\big(r(L_k)\oplus(k,v)\oplus(v,i),\tilde{\boldsymbol{s}},\tilde{\boldsymbol{t}}\big)\Big) \le \beta$.

As the continuity of care constraints impose restrictions on the number of caregivers serving each patient, it is handled in the master problem with additional *robust cuts*. We introduce the formulation and solution approach in EC.7.

5.1.3 Subset row inequalities and the CSRI-based dominance rule.

To strengthen the quality of the MP lower bound, we adopt subset-row cuts (SRCs) in the `CSRI-BPC` procedure. Accordingly, the dominance rule considering the impact of non-robust SRCs must be updated in the subproblem.

5.1.4 Warm-up and acceleration techniques.

The warm-up procedure is incorporated to generate the initial column pool $\hat{R}$ and obtain an initial upper bound. We first propose a *backtracking algorithm* and further develop a `CSRI-VNS` algorithm (Section 5.2) to close the initial solution gap. Furthermore, we adopt strategies such as *ng-route*, *heuristic pricing* and *digital tree (trie)* to accelerate the solving process of the subproblem.

5.1.5 Branching scheme.

To obtain integer solutions, we enforce the following branching rules in the branch-and-bound search tree: (i) on the fractional total flow of an arc visited by all caregivers; (ii) on the node visited a fractional number of times; and (iii) on the fractional total flow of an arc visited by one caregiver.

5.2 Main Components of `CSRI-VNS`

To solve large-scale real-world instances and obtain high-quality solutions, we develop a metaheuristic framework based on the variable neighborhood search (VNS), which can be integrated with our CSRI approach to obtain suboptimal or even optimal solutions for the RSRP (referred to as `CSRI-VNS`). The main components are described as follows:

- **Initial solution generation.** The initial solutions (i.e., routes) are constructed by the backtracking algorithm outlined in Section 5.1.4.
- **Local search procedure.** Four operators (i.e., relocate, exchange, reverse and interchange) are adopted to produce more promising solutions.
- **Diversification procedure.** To escape from the incumbent local optimum, we randomly remove several patient nodes and then insert them back via the backtracking method.

Specifically, for each new route generated either by local search or diversification, the CSRI evaluation approaches are applied to calculate the CSRI value and ensure that it does not exceed the threshold β; otherwise, the route is excluded. Overall, the `CSRI-VNS` complexity is $O(|\mathcal{V}|)$.

As mentioned above, the CSRI-VNS algorithm includes the backtracking algorithm as its initial baseline, enabling it to produce better solutions without consuming excessive computational effort. Therefore, it outperforms the backtracking algorithm in terms of solution quality and efficiency. We thus adopt CSRI-VNS to quickly obtain a suboptimal, or even optimal, solution to warm up CSRI-BPC for a tighter upper bound, which terminates when either it does not improve the best solution for 10 iterations or the one-hour time limit is reached.

6 Numerical Experiments

In this section, we provide a comprehensive experimental analysis that explores the efficiency and effectiveness of our models, methodologies, and algorithms. We begin by introducing the benchmark and real-world instances used for our experiments in Section 6.1. In Section 6.2.1, we evaluate the performance of the exact methods outlined in Section 5.1 by comparing different tailored CSRI-solution approaches and explore the searchability and performance of the metaheuristic approach described in Section 5.2. Next, we investigate the impacts of the topological structure on the solution performance in Section 6.2.2. In Section 6.3, we highlight the advantage of incorporating the compound ambiguity set relative to the single joint ambiguity set. Finally, we present the performance of our methods in solving real-world instances and derive pragmatic managerial insights in Section 6.4.

We implement the designed algorithm in C++ and use IBM ILOG CPLEX V20.1 to solve the LP and MIP models. Our experimental environment is an AMD Ryzen Threadripper 3990x 64-core, 128-thread CPU. All the experiments are run on a single thread. The overall time limit to terminate the solving process is 3,600 seconds (denoted as T.L.).

6.1 Datasets and Methods

Through the tests, we selected two datasets: a benchmark dataset and an industrial dataset. The benchmark dataset is widely adopted in the literature, notably in the SRI study (Zhang et al. 2021), whereas the industrial dataset aims to provide more managerial insights from a practitioner's perspective. The construction of the two datasets is as follows.

The benchmark instances tested in our experiments are adapted from the widely used deterministic VRPTW instances with tight time windows proposed by Solomon (1987), which consist of 100 customer nodes and 29 instances in total. Instances are divided into three classes on the basis of the nodes' geographic distribution: *c*-type (clustered), *r*-type (random), and *rc*-type (mixed clustered-random). Furthermore, these instances are commonly categorized into three different scales based on selecting the first 25, 50, and 100 nodes, with the corresponding numbers of available vehicles being 8, 15, and 25, respectively. For brevity, we use the notation $(8,25)$ to represent an instance with 8 vehicles and 25 customers. This results in a total of 87 (29×3) instances used in the benchmark testbed. We compute the travel time and cost for each arc based on the Euclidean distance and rounded them down to the first decimal place. To account for uncertainty, we modify the deterministic travel and service times using asymmetric two-point distributions supported on $\mu - \sigma/\sqrt{3}$ and $\mu + \sqrt{3}\sigma$ with probabilities of 0.75 and 0.25 respectively (Zhang et al. 2021). Here, μ is the original value, and $\sigma = \lambda\mu$ is its standard deviation, where λ is generated randomly from the interval $[0.1, 0.5]$.

Furthermore, we obtain an operational dataset from our AHH industry partner. This dataset contains 142 AHH visiting routes with 332 detailed appointments from November 1 to November 6, 2019, including information such as the actual service time, appointment time (i.e., the latest service time required by the patient and the working time regulation), the longitude and latitude of the caregiver and patient locations, and patient conditions and caregiver skills. The distance matrix is acquired with the longitudes and latitudes of locations from *Amap* (*https://developer.amap.com*), i.e., the shortest integer travel time by taking a combination of metro and bus between any two nodes (the travel times satisfy the triangular inequality conditions). We generate 50 medium-scale instances with 15 caregivers and 50 patients and 50 large-scale instances with

25 caregivers and 100 patients. The patients are randomly selected based on the five most common medical services, and we preprocess the data to ensure compatibility between caregivers and patients by matching their skills and requirements. To reflect real-world situations, we retrieve 100 copies of travel times via Amap and service times for selected patients at different times and dates. Overall, we utilize this dataset and its generated instances as a practice testbed to study the AHH industry and analyze the performance of different algorithms.

To construct the cross moment ambiguity set, we empirically estimate the means $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ based on the samples from the joint travel time distribution $\mathbb{P}_t^{\dagger}$, as in many studies (e.g., Zhang et al. 2019, Ghosal and Wiesemann 2020). For our analysis, we set the risk aversion parameter γ to 0.1 following the convention. Unless specified otherwise, we always set the CSRI threshold $\beta = 0.2$, the Wasserstein radius $\theta = 0.05$ and use $N = 20$ travel and service time samples across all tests, which are calibrated via a preliminary fourfold cross-validation technique (Esfahani and Kuhn 2018). For comprehensive study purposes, we conducted a sensitivity analysis to investigate the impact of key robustness parameters (θ, N and γ) and provided managerial explanations in EC.9.

In addition to defining the benchmark and real-world testbeds, we evaluate the effectiveness of the following methods in our experiments:

i. `CSRI-VNS`: The variable neighborhood search algorithm developed in Section 5.2 and equipped with the $\Phi_{\text{CSRI-EXA}}$ approach to solve the RSRP. Furthermore, let `SRI-VNS` denote the variable neighborhood search algorithm developed for handling the CSRI governed by a single Wasserstein ambiguity set that embeds service time uncertainty in travel times, as described in Zhang et al. (2021).

ii. `CSRI-BPC`: The branch-price-and-cut algorithm proposed in Section 5.1 to solve the RSRP. We use $\texttt{BPC}_{\texttt{SAA}}$, $\texttt{BPC}_{\texttt{E}}$, and $\texttt{BPC}_{\texttt{SE}}$ to represent the `CSRI-BPC` equipped with the $\Phi_{\text{CSRI-SAA}}$, $\Phi_{\text{CSRI-EXA}}$ and $\Phi_{\text{CSRI-SUF}}$+$\Phi_{\text{CSRI-EXA}}$ approaches (mentioned in Section 4.3), respectively. In addition, we define the $\texttt{BPC}_{\texttt{VNS}}$to indicate that `CSRI-VNS` is adopted in the warm-up procedure (see Subsection 5.1.4). Therefore, enhanced by `CSRI-VNS`, the `CSRI-BPC` algorithm further evolves into the $\texttt{BPC}_{\texttt{VNS+SAA}}$, $\texttt{BPC}_{\texttt{VNS+E}}$ and $\texttt{BPC}_{\texttt{VNS+SE}}$.

After obtaining the routing and scheduling solution, we perform the out-of-sample evaluation by testing the solution on another 10,000 newly generated travel and service time samples. Finally, we report the following characteristics of the solutions: *Instance solved/tested* (number of tested and solved instances for each type), $T(s)$ (the runtime in seconds), Obj (the optimal or best-known value, where '–' indicates that an upper bound with at least one integer solution cannot be derived, and thus no corresponding gap exists), $Gap(\%)$ (the relative gap computed with respect to the lower bound and best-known values obtained in the branch-and-bound tree, (best-known value - lower bound)/lower bound $\times$ 100), *MaxProb* (the maximum lateness probability across all nodes), *AveProb* (the average lateness probability across all nodes), *MaxExp* (the maximum expected lateness time across all nodes), *AveExp* (the average expected lateness time across all nodes), and *SumCSRI* (the sum of out-of-sample CSRI values for all nodes).

6.2 Comparison of Solution Methods and Topological Structures

6.2.1 Experiments on Algorithm Performance Evaluation

To evaluate the computational performance of our proposed exact algorithm framework, which incorporates various CSRI solution approaches, we investigate the $\texttt{BPC}_{\texttt{VNS+SAA}}$, $\texttt{BPC}_{\texttt{E}}$, $\texttt{BPC}_{\texttt{VNS+E}}$, and $\texttt{BPC}_{\texttt{VNS+SE}}$ methods for solving benchmark instances. We present the overall performance summary in Table 2 and show the detailed results in Tables EC.2-EC.7.

The results indicate that $\texttt{BPC}_{\texttt{VNS+SE}}$ generally outperforms the other methods, achieving the highest number of solved instances and the smallest exit gap. This superior performance can be attributed to the reduced computational time spent on

Table 2 Computational Performance Summary of the CSRI−BPC with Different CSRI Evaluation Approaches

	$\text{BPC}_{\text{VNS+SAA}}$			BPC_{E}			$\text{BPC}_{\text{VNS+E}}$			$\text{BPC}_{\text{VNS+SE}}$		
Instance type	Instance solved/tested	$T(s)$	$Gap(\%)$	Instance solved/tested	$T(s)$	$Gap(\%)$	Instance solved/tested	$T(s)$	$Gap(\%)$	Instance solved/tested	$T(s)$	$Gap(\%)$
c	0/27	T.L.	-	5/27	3460	6.52	9/27	2568	3.95	10/27	2507	3.75
r	0/36	T.L.	-	12/36	2555	1.63	12/36	2440	1.48	16/36	2359	0.66
rc	0/24	T.L.	-	8/24	2623	5.25	9/24	2538	2.79	10/24	2323	1.58
All	0/87	T.L.	-	25/87	2842	4.14	30/87	2379	2.61	36/87	1742	1.89

the CSRI-feasibility checks. Specifically, the $\Phi_{\text{CSRI-SUF}}$ approach embedded in the $\text{BPC}_{\text{VNS+SE}}$ avoids unnecessary CSRI-feasibility checks with trivial assessments that take constant computational time. Additionally, both the $\Phi_{\text{CSRI-SUF}}$ approach and the $\Phi_{\text{CSRI-EXA}}$ approach are more efficient than the $\Phi_{\text{CSRI-SAA}}$ approach, which involves a linear program with $O(N)$ decision variables and $O(N)$ constraints, rendering it quite computationally demanding. This explains why $\text{BPC}_{\text{VNS+SAA}}$ cannot obtain a feasible upper bound for even one instance. It is worth mentioning that the $\text{BPC}_{\text{VNS+SE}}$ is capable of solving most instances with the shortest computational time across all methods. Specifically, 28 of the 29 (8,25)-instances are solved to optimality, and the average unsolved gap for all 87 instances is merely 1.89%. This performance outperforms existing exact algorithms dedicated to solving DRO models for VRPTW (e.g., Zhang et al. 2019, 2021). Therefore, it is reasonable to use $\text{BPC}_{\text{VNS+SE}}$ as the primary representation of exact methods in future experiments unless otherwise specified. To further justify this decision, we compare the performance profiles across all methods on the benchmark testbed. As shown in Figure 4, $\text{BPC}_{\text{VNS+SE}}$ is capable of solving most instances to proven optimality with the shortest computational time on average, which reinforces the notion that it is the most efficient and effective exact solution approach.

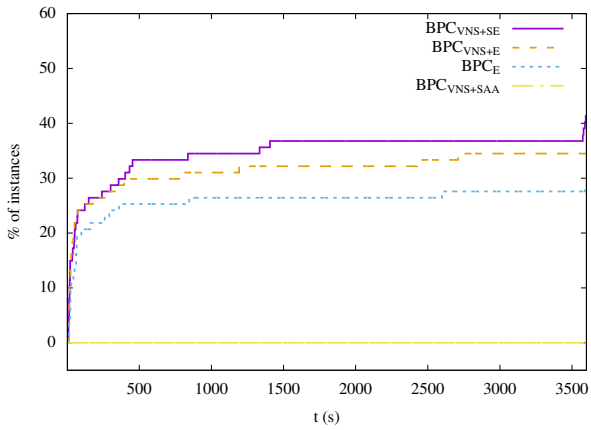


Figure 4 Performance Profile of the $\text{BPC}_{\text{VNS+SE}}$, $\text{BPC}_{\text{VNS+E}}$, BPC_{E} and $\text{BPC}_{\text{VNS+SAA}}$ Algorithms: Percentage of Instances Solved to Optimality within the Given Computing Times

To investigate the computational enhancement achieved by the metaheuristic, we examine the CSRI−VNS and $\text{BPC}_{\text{VNS+SE}}$ methods with more benchmark instances. We subsequently evaluate the performance of these two methods through out-of-sample tests. The average performance is presented in Table 3 grouped by instance type and the detailed results can be found in Tables EC.6-EC.10.

Table 3 indicates that CSRI−VNS exhibits excellent searchability, as exact methods fail to provide better solutions for 21 of the 29 (8,25)-instances and can only obtain the same solutions as those produced by the CSRI−VNS. Moreover, when solving larger-scale instances, CSRI−VNS can quickly find near-optimal or potentially even optimal solutions. The experiments indicate that a graph structure with up to 25 caregivers and 100 patients might be a suitable scale for efficient optimization. Finally, the out-of-sample results suggest that the solutions generated by CSRI−VNS are as robust as those obtained by the $\text{BPC}_{\text{VNS+SE}}$, as evidenced by the corresponding out-of-sample indicators, e.g., smaller $AvgExp$ values. More importantly, these results also demonstrate that our formulations can significantly balance patient satisfaction and cost

Table 3 Computational and Out-of-Sample Performance of CSRI−VNS, SRI−VNS and $\text{BPC}_{\text{VNS+SE}}$ for Different Scale Instances

			Performance							
Method	Scale	*Instance*	$T(s)$	Obj	$Gap(\%)$	*AveProb*	*MaxProb*	*AveExp*	*MaxExp*	*SumCSRI*
CSRI−VNS	(8,25)	c	7	242.4	-	0.01	0.10	0.23	2.38	1.88
		r	8	424.0	-	0.01	0.12	0.05	0.92	2.68
		rc	7	332.9	-	0.02	0.13	0.11	0.97	2.65
		Average	7	342.5	-	0.01	0.12	0.12	1.39	2.42
	(15,50)	c	39	494.2	-	0.01	0.10	0.17	2.22	3.85
		r	49	731.4	-	0.02	0.17	0.11	1.47	5.27
		rc	40	799.5	-	0.03	0.21	0.27	3.05	5.55
		Average	43	676.6	-	0.02	0.16	0.17	2.14	4.90
	(25,100)	c	233	1104.7	-	0.01	0.20	0.29	6.50	8.19
		r	356	1102.6	-	0.02	0.21	0.13	2.65	9.81
		rc	365	1317.2	-	0.02	0.21	0.19	3.24	10.67
		Average	320	1162.4	-	0.02	0.20	0.20	4.01	9.55
SRI−VNS	(8,25)	c	2	241.0	-	0.01	0.10	0.20	2.33	3.30
		r	2	416.7	-	0.01	0.15	0.12	1.36	3.12
		rc	1	334.5	-	0.02	0.21	0.25	2.92	3.34
		Average	2	339.5	-	0.01	0.15	0.18	2.09	3.24
	(15,50)	c	14	494.7	-	0.01	0.14	0.21	3.43	6.70
		r	12	712.9	-	0.03	0.22	0.22	2.42	6.52
		rc	7	727.5	-	0.04	0.25	0.45	4.35	6.77
		Average	11	649.2	-	0.03	0.20	0.28	3.27	6.64
	(25,100)	c	95	1102.3	-	0.01	0.17	0.29	6.53	13.47
		r	121	1086.9	-	0.02	0.22	0.12	2.34	13.64
		rc	114	1248.2	-	0.03	0.26	0.29	4.12	13.68
		Average	111	1136.2	-	0.02	0.21	0.22	4.13	13.60
$\text{BPC}_{\text{VNS+SE}}$	(8,25)	c	644	241.9	0.09	0.01	0.08	0.18	1.55	1.87
		r	33	421.7	0.00	0.01	0.12	0.05	0.94	2.68
		rc	153	327.3	0.00	0.02	0.13	0.11	0.98	2.56
		Average	256	339.9	0.03	0.01	0.11	0.11	1.14	2.39
	(15,50)	c	3219	492.1	4.50	0.01	0.12	0.17	2.22	3.85
		r	3374	722.9	1.03	0.02	0.18	0.11	1.45	5.25
		rc	2934	726.9	2.10	0.03	0.21	0.29	2.91	5.47
		Average	3204	625.3	2.40	0.02	0.16	0.18	2.09	4.88
	(25,100)	c	3658	1099.2	6.67	0.01	0.20	0.29	6.50	8.17
		r	3668	1092.4	0.95	0.02	0.21	0.15	2.68	9.76
		rc	3712	1294.3	3.29	0.02	0.21	0.16	2.74	10.53
		Average	3677	1150.2	3.38	0.02	0.20	0.19	3.88	9.48

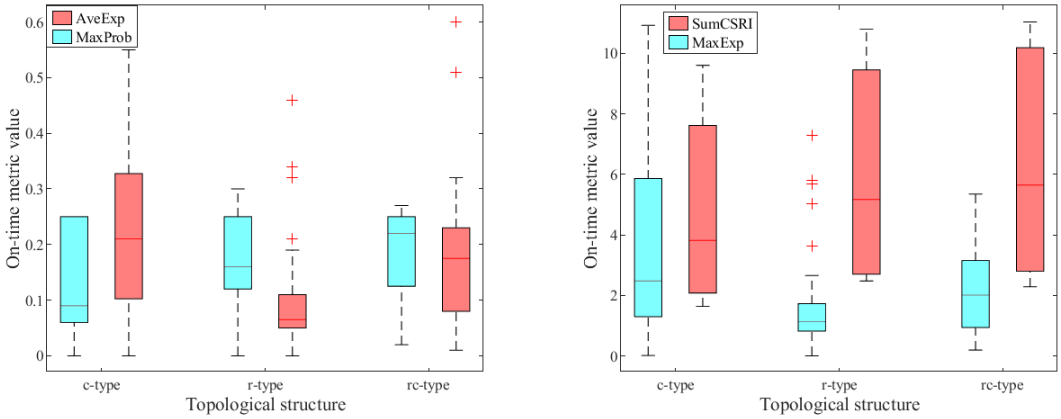
reduction, aligning with our initial motivation to enhance caregiver efficiency and reduce operating costs by allowing for a slight amount of patient-tolerant overtime. For example, the average lateness probability (*AveProb*) for the (15,50)-instances of the CSRI−VNS solution is 0.02, and the average expected lateness time (*AveExp*) is 0.17 minutes. These results reveal the potential to pursue operational excellence by allowing trivial lateness. Therefore, by restricting the CSRI constraints for each patient and caregiver, the CSRI formulations can assist practitioners in pursuing economic benefits while minimizing negative impacts on patient satisfaction.

6.2.2 Impact of Topological Structures

Based on the results presented in Table 2, the performance of $\text{BPC}_{\text{VNS+E}}$ and $\text{BPC}_{\text{VNS+SE}}$ is significantly influenced by the graph's topological structure, i.e., the geographical distribution of nodes. The *r*-type instances appear to be the most straightforward to optimize and have smaller exit gaps. This might be attributed to the fact that clustered nodes result in relatively shorter travel times, thereby introducing more combination patterns and expanding the solution-searching space. Consequently, decision-makers should evaluate the distribution of patients and arrange them into clusters to streamline scheduling and improve service efficiency. Following these insights, the existing literature has developed a plethora of clustering policies for informed decisions in the AHH sector (e.g., Cire and Diamant 2022, Pahlevani et al. 2022).

However, few past studies have investigated the implications of the graph's topological structure on the manifestation of delays. Interestingly, we find that schedules for various geographical distributions exhibit different overtime risk patterns

because of the cascading effect. As depicted in Figure 5a, despite having similar out-of-sample CSRI levels, patients distributed randomly are more likely to either experience no overtime or a trivial delay with a higher frequency. In contrast, patients in clustered distributions generally endure delays with more consistent low-level overtime probabilities but larger average expected lateness times, as shown in Figure 5b. The reasoning behind this is clear: Clustering patients geographically can lead to more efficient routing and scheduling, as caregivers can visit multiple patients in close proximity within a single trip. This reduces the proportion of time spent traveling compared to providing service, which in turn increases the number of visiting patients, intensifying the cascading effect caused by the aggregated uncertain durations. Finally, the same CSRI threshold could lead to different delay manifestations for different topological structures. Hence, decision-makers should factor in patient distribution when setting the CSRI threshold to strike a balance between operating costs and service level. Clustered patients combined with more restrictive service-level requirements are recommended to enhance visiting effectiveness as well as prevent potential nontrivial delays.



(a) (b)

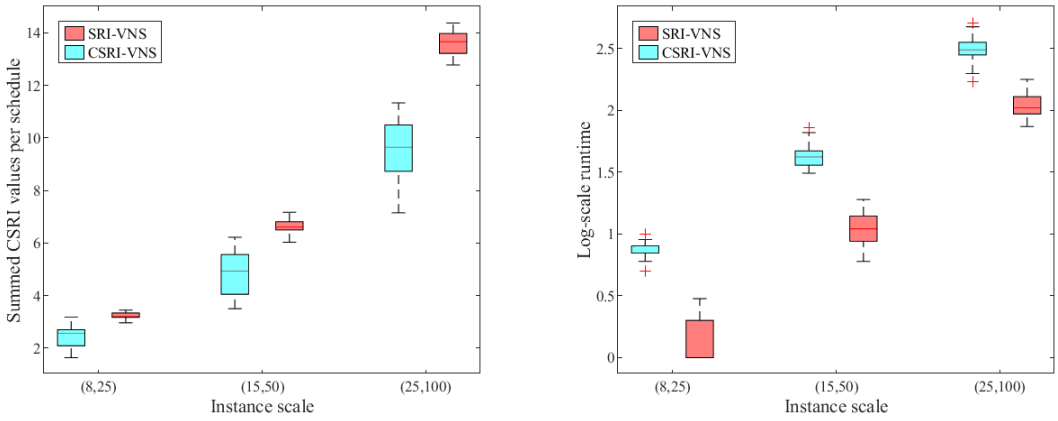
Figure 5 Comparison of Topological Structures: (a) *MaxProb* and *AveExp*; (b) *MaxExp* and *SumCSRI*

6.3 Comparison of Compound vs. Single Ambiguity Sets

To understand the benefits of using a compound ambiguity set in contrast to a single ambiguity set for integrated travel and service times, we compared the `CSRI-VNS` and `SRI-VNS` methods over Solomon's instances, which were adopted for SRI tests in Zhang et al. (2021). We assessed the effectiveness of these two approaches via out-of-sample tests, highlighting the significance of individually addressing travel and service times due to their distinct intrinsic attributes. The mean performance for each approach is displayed in Table 3, categorized by instance type. The detailed outcomes are elaborated in Tables EC.8-EC.13.

As expected, Table 3 indicates that `CSRI-VNS` exhibits better on-time service performance than `SRI-VNS` does when evaluated under identical CSRI constraints. This advantage is particularly evident in instances involving 100 customers. Figure 6a visually reinforces the notion that the compound ambiguity set, as defined in Section 4.1, outperforms the single Wasserstein ambiguity set, especially in terms of punctuality across various instance sizes. For example, in the context of the (25,100)-instances for the `CSRI-VNS` solution, the sum of out-of-sample CSRI values (*SumCSRI*) reaches 9.55. In contrast, the `SRI-VNS` strategy yields a value that is increased by 42.41%. Moreover, the `CSRI-VNS` method consistently outperforms the `SRI-VNS` technique when metrics such as *AveProb*, *MaxProb*, *AveExp*, and *MaxExp* are taken into account, accomplishing better punctuality with a mere 2.31% increase in the total cost. Our analysis suggests that

the enhanced performance stemming from the compound ambiguity set is due to its refined ability to encompass and synchronize a diverse array of uncertainties. Specifically, the variance-covariance matrix adeptly captures the correlations in uncertain travel times, whereas the Wasserstein ambiguity set leverages the historical data of fluctuating service periods. Together, these elements aid in the creation of more resilient scheduling.



(a) (b)

Figure 6 Comparison of `CSRI-VNS` and `SRI-VNS`: (a) Summed CSRI Values and (b) Runtime

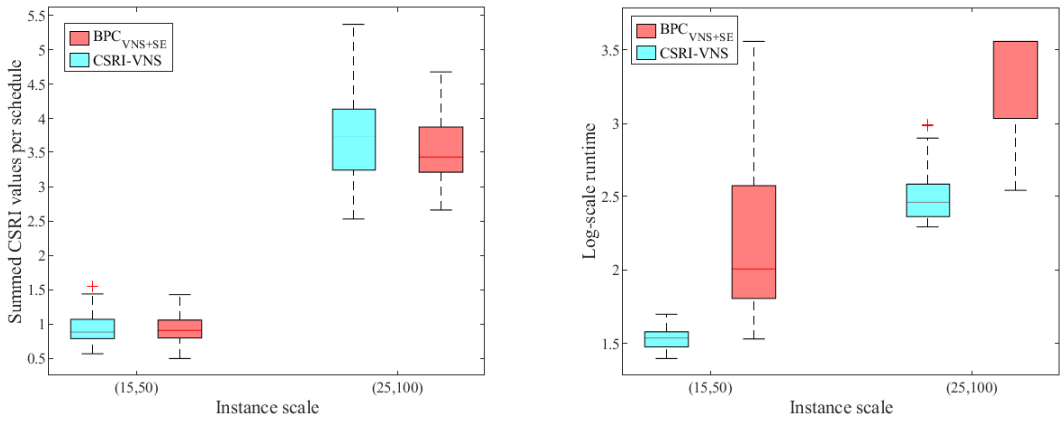
While the compound ambiguity set yields improved out-of-sample performance, it does entail a marginally increased computational time (see Figure 6b). As a result, for those targeting near-optimal solutions, the `SRI-VNS` method might be more efficient than the `CSRI-VNS`. However, the observed difference in computational time between the two methods is marginal and remains within acceptable limits for practical applications. Despite its slightly elevated computational requirements, the `CSRI-VNS` approach is both efficient and insightful for exploring real-world situations, as shown in the upcoming Section 6.4. In addition, CSRI outperforms the traditional LPI in terms of both computational efficiency and on-time performance, as detailed in EC.8.

6.4 Real-World Dataset Tests

To evaluate the performance of our algorithms in solving practical problems, we apply the `CSRI-VNS` and $\texttt{BPC}_{\texttt{VNS+SE}}$ methods to solve real-world AHH instances, and summarize the detailed results in Tables EC.14-EC.17.

(1) Summary of average performance. The numerical results demonstrate that the `CSRI-VNS` and $\texttt{BPC}_{\texttt{VNS+SE}}$ methods are effective when handling real-world instances. Specifically, both methods achieve satisfactory out-of-sample performance, with average *MaxProb* and *MaxExp* values of less than 25% and 5 minutes, respectively, for all instances. Notably, as depicted in Figure 7a, there is no significant difference regarding the average summed CSRI values per schedule. Computationally, for the (15,50)- and (25,100)-instances, $\texttt{BPC}_{\texttt{VNS+SE}}$ takes 1,478 seconds, on average, solving 71 (out of 100) instances to optimality. As a comparison, Figure 7b shows that `CSRI-VNS` is faster by at least one order of magnitude and acquires equivalent-quality solutions. Therefore, `CSRI-VNS` is more advantageous when handling practical instances, with a focus on both computational efficiency and solution resiliency.

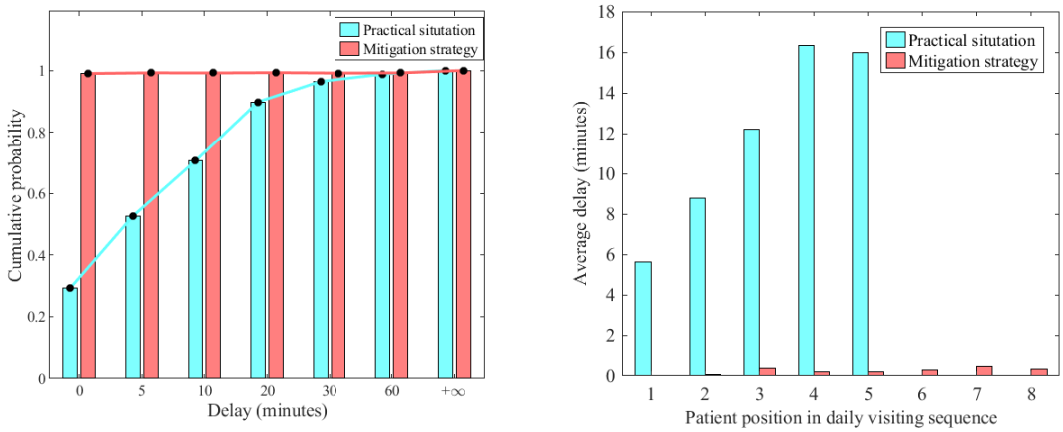
(2) Effectiveness of cascading delay mitigation. Next, we highlight the effectiveness of our methods in mitigating the cascading delays that arise due to uncertainties in AHH scheduling and routing practices. Figure 8 illustrates the performance of our model and dedicated solution approaches when addressing the cascading effect. To facilitate clear comparisons, we duplicate the actual cascading delays presented in Figure 1. Specifically, as Figure 8a shows, the `CSRI-VNS` solution achieves an average expected delay of 0.16 minutes (compared with 8.59 minutes for manual schedules), with a 0.96%



(a) (b)

Figure 7 Average Performances for Real-World Instances: (a) Summed CSRI Values and (b) Runtime

lateness probability (compared with 70.83% for manual schedules) for all patients within the (15,50)-instances. This significant improvement in punctuality effectively eliminates the vast majority of delays in the manual schedules. Furthermore, as shown in Figure 8b, the cascading effect has been effectively curtailed, with the (originally) increasing tendency being flattened (almost unobservable compared with the original case) through our CSRI-tailored mitigation strategy.

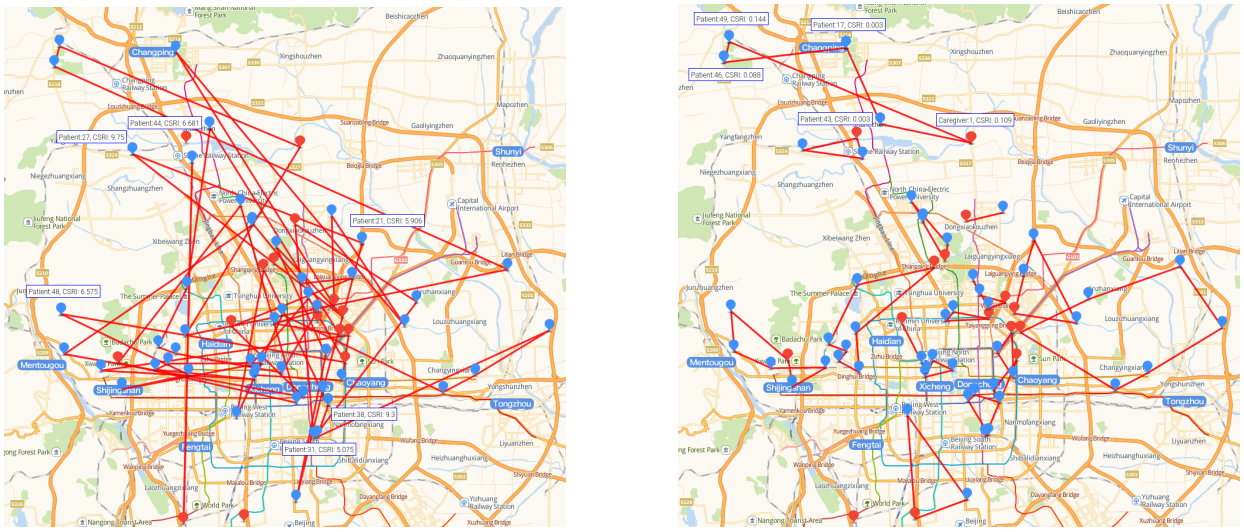


(a) (b)

Figure 8 Delay Description of All (15,50)-Instances Solved by `CSRI-VNS`

The $\texttt{BPC}_{\texttt{VNS+SE}}$ method also yields similar results when solving both (15,50)- and (25,100)-instances. Note that the average travel times for each node are 10.8 and 16.4 minutes for (15,50)- and (25,100)-instances, respectively. After analyzing the instance features, we attribute this superiority to the tendency of patients to cluster in different urban regions, such as neighborhoods or residential zones, whereas these urban regions themselves are randomly scattered and often far apart. However, while this *rc*-type pattern (mixed clustered-random) facilitates cost-effective solutions, it also necessitates decision-makers to enforce stricter on-time requirements (i.e., smaller β) to mitigate the cascading effect as previously analyzed.

In summary, the above findings demonstrate that our proposed algorithms significantly improve caregiver visiting effectiveness, reduce operating costs, and minimize patient dissatisfaction (caused by delays) compared with the manual schedule. We further visualize this comparison for a (15, 50) instance in Figure 9, which displays the visiting sequence for each patient and labels the route with the maximum individual CSRI value. Figure 9a depicts the current real-world schedule, which is executed independently by different schedulers and lacks comprehensive optimization. For comparison, Figure



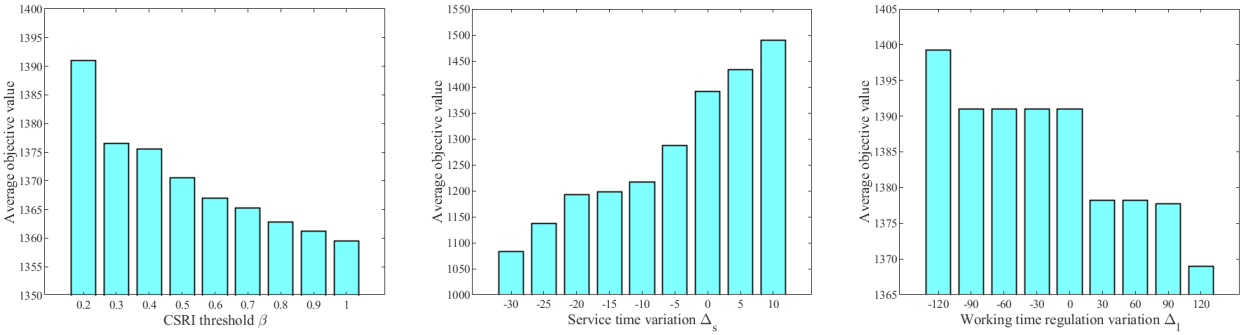
(a) Industry partner's manual schedule (b) The schedule obtained by $\texttt{BPC}_{\texttt{VNS+SE}}$

Figure 9 Comparison between Manually Obtained Schedules and Those Generated by $\texttt{BPC}_{\texttt{VNS+SE}}$

Notes. The red and blue icons represent caregivers' and patients' locations, respectively, and the edges connecting nodes represent the caregivers' scheduling and movement paths.

9b presents an optimized schedule via our $\texttt{BPC}_{\texttt{VNS+SE}}$ method. All patients in densely populated residential areas are properly allocated and sequenced, and caregivers living nearby can efficiently provide services without traversing the whole city, thereby avoiding costly commuting routes. Moreover, the cascading delays are effectively managed for the clustering structure. As indicated, the maximum patient CSRI value has been reduced from 9.75 to 0.144, showcasing a substantial improvement in the on-time performance of the schedule. The total travel cost has been significantly reduced from 2025 to 741. This comparison illustrates that the current manual schedule is dominated by our recommended solutions, which highlights the effectiveness of the proposed method and its potential for practical applications.

(3) Impact of the key managerial parameters. Finally, we aim to explore pragmatic improvements that practitioners can undertake for operational excellence. To ensure computational convenience while maintaining generality, we randomly select several instances with scales of (15,50) and (25,100). Then, we vary the parameters within a range, including the CSRI threshold β, service time variation Δ_s, and working time regulation variation Δ_l. Finally, the results are summarized in Tables EC.26-EC.28, and graphical comparisons are provided in Figure 10. Specifically,



(a) (b) (c)

Figure 10 Description of the Key Managerial Parameter Tests

(i) *Impact of the on-time service requirement.* The results demonstrate that, as the on-time service and completion requirements gradually shift from leniency to stringency (i.e., β decreases), the objectives described in Figure 10a increase for boosting out-of-sample performance, whereas computational efficiency improves (see Table EC.26). For example,

in the case of instance '25_100_01', the objective increases from 2,063 to 2,197 (from $\beta = 1$ to $\beta = 0.2$), with $T(s)$ decreasing from 2,475 seconds to 764 seconds, and *AveExp* from 0.21 minutes to 0.07 minutes. However, as expected, some instances (e.g., '25_100_03') become infeasible when the punctual criterion is excessively strict, which should serve as a warning to overly conservative decision-makers who prioritize service quality over all else. Therefore, we conclude that stringent on-time service and completion requirements inevitably lead to additional operating costs with improved computational efficiency, posing challenges for decision-makers who must balance schedule resilience and efficiency against operating costs.

(ii) *Impact of the service time duration.* When service times are reduced by a certain number of minutes, the objective improves significantly, as shown in Figure 10b, but the computational burden also increases. For example, in the case of '25_100_01' with $\Delta_s = -10$, the objective value decreases more than 10% from 2,103 to 1,868, whereas $T(s)$ increases from 455 seconds to 1,460 seconds. This outcome aligns with our expectation that shorter service times allow for more flexible travel schedules, leading to diverse visiting routes and additional computational demands. Therefore, when decision-makers consider hiring additional caregivers with specialized skills to provide more scheduling flexibility, employing proficient caregivers is crucial, as shorter service times can substantially reduce the operating costs associated with medical services.

(iii) *Impact of the working time regulation.* Finally, the results in Figure 10c indicate that the current working time regulation, which was designed to cover all possible patient time windows, is overly conservative. Reducing the working time regulation by 120 minutes would lead to a substantial improvement in caregiver satisfaction without incurring significant additional costs. That is, if the service provider could adopt our solutions, then they might be able to cut down the working time length appropriately while still serving all patients and eliminating delays.

In summary, this section presents various numerical experiments that validate the efficiency and effectiveness of our cascading delay mitigation strategy and solution methodology. We truly hope that these insights can support managers in achieving operational excellence, particularly in the context of AHH scheduling and routing.

7 Conclusion

In this article, we investigate a last-mile attended home healthcare (AHH) delivery problem, which has gained considerable attention in recent years due to the aging population trend. We analyze operational data from a service provider and discover a cascading delay effect in the AHH business that has not been fully discussed. This phenomenon is significant as delay propagation impairs patient satisfaction and companies' operational performance. To address this concern, we attribute the key reasons for cascading delays to uncertain travel and service times and develop a systematic strategy to mitigate these issues. Specifically, to capture the inherent features of travel and service time uncertainties, we construct a compound ambiguity set by adopting cross moment and Wasserstein ambiguity sets. Then, we propose a new decision criterion, termed the CSRI, to gauge delays in terms of their probability and magnitude. We present three approaches to evaluate the CSRI values. With the proposed risk assessment measure, we develop a set-partitioning model, which utilizes CSRI-based service-level constraints for each node to curb cascading delays. While similar problems have been investigated in the literature, to the best of our knowledge, there is no distributionally robust model that reconciles distinct uncertain travel and service times into a compound ambiguity set for risk assessment of individual nodes. To efficiently solve this problem, we develop an exact branch-price-and-cut solution framework with acceleration strategies (`CSRI-BPC`) and a fast-effective metaheuristic (`CSRI-VNS`). Finally, numerical experiments on benchmark and real-world datasets demonstrate the computational efficiency and effectiveness of our proposed methods in mitigating cascading delays. For example, our methods reduce the average delay from 8.59 minutes to 0.16 minutes and almost eliminate the cascading effect through practical tests.

This paper presents valuable insights for AHH operations in several aspects. For accurate risk assessment, we demonstrate the imperative and advantages of adopting the compound ambiguity set theoretically and numerically when the characteristics of uncertainties are distinct. Our numerical results also elucidate the impact of the graph's topological structure on the computational efficiency and the cascading effect. This information can help decision-makers allocate patients and employ caregivers with suitable geographical distributions to achieve better scheduling flexibility. Finally, we conduct sensitivity analyses for robustness and time-related parameters to support informed decisions. Overall, we believe that these insights could effectively aid decision-making scenarios in scheduling and routing management for AHH service delivery, and help balance scheduling resilience and efficiency against operating costs. In conclusion, this article not only complements the robust vehicle routing literature with a systematic strategy, but also offers practical guidance for AHH practitioners to mitigate the cascading effect and achieve operational excellence.

Our study indicates several interesting future research directions that could be explored and advanced, not only in the AHH sector, but also in other service industries. First, one may find the research of the unified compound ambiguity set framework, which covers all combinations of different types of ambiguity sets in general for all metrics (risk measures, satisficing measures, or disutility functions), presents a promising avenue for future research, such as the design of patient visit itineraries in tandem systems (Liu et al. 2024b), and airline fleet assignment problems with uncertain boarding and flying times. Moreover, adopting machine learning tools in the preprocessing stage to exploit the uncertainty structure can help obtain a better sample set with a proper size. This, in turn, can lead to more efficient and robust solutions via the proposed algorithms (Wang et al. 2023).

Acknowledgments

The authors gratefully thank the Department Editor Michael Pinedo, the Senior Editor and the three anonymous reviewers for their helpful and insightful comments that have significantly improved the paper. This work is supported by the National Natural Science Foundation of China (NSFC) under Grants (No.71872093, No.723B2014), National Social Science Fund of China (No.21&7D128) and the Durham University Business School Primary Research Funding.

References

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

Baldacci, Roberto, Enrico Bartolini, Aristide Mingozzi, Roberto Roberti. 2010. An exact solution framework for a broad class of vehicle routing problems. *Computational Management Science*, 7 (3), 229-268.

Bard, Jonathan F., Yufen Shao, Xiangtong Qi, Ahmad I. Jarrah. 2014. The traveling therapist scheduling problem. *IIE Transactions*, 46 (7), 683-706.

Bartolini, Enrico, Dominik Goeke, Michael Schneider, Mengdie Ye. 2021. The robust traveling salesman problem with time windows under knapsack-constrained travel time uncertainty. *Transportation Science*, 55 (2), 371-394.

Ben-Tal, Aharon, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59 (2), 341-357.

Benjaafar, Saif, David Chen, Rowan Wang, Zhenzhen Yan. 2023. Appointment scheduling under a service-level constraint. *Manufacturing & Service Operations Management*, 25 (1), 70-87.

Bergman, Alon, Guy David, Hummy Song. 2023. "I Quit": Schedule volatility as a driver of voluntary employee turnover. *Manufacturing & Service Operations Management*, 25 (4), 1416-1435.

Cappanera, Paola, Maria Grazia Scutellà, Federico Nervi, Laura Galli. 2018. Demand uncertainty in robust home care optimization. *Omega*, 80 95-110.

Centers for Medicare & Medicaid Services. 2023. NHE Fact Sheet. `https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet`. Accessed March 27, 2023.

Chen, Li, Chenyi Fu, Fan Si, Melvyn Sim, Peng Xiong. 2023. Robust optimization with moment-dispersion ambiguity. *SSRN Electronic Journal*, URL `https://api.semanticscholar.org/CorpusID:260594845`.

Chen, Louis, Will Ma, Karthik Natarajan, David Simchi-Levi, Zhenzhen Yan. 2022. Distributionally robust linear and discrete optimization with marginals. *Operations Research*, 70 (3), 1822-1834.

Chen, Zhi, Daniel Kuhn, Wolfram Wiesemann. 2024. Technical note—data-driven chance constrained programs over wasserstein balls. *Operations Research*, 72 (1), 410-424.

Cire, Andre A., Adam Diamant. 2022. Dynamic scheduling of home care patients to medical providers. *Production and Operations Management*, 31 (11), 4038-4056.

Costa, Luciano, Claudio Contardo, Guy Desaulniers. 2019. Exact branch-price-and-cut algorithms for vehicle routing. *Transportation Science*, 53 (4), 946-985.

Cui, Zheng, Daniel Zhuoyu Long, Jin Qi, Lianmin Zhang. 2023. The inventory routing problem under uncertainty. *Operations Research*, 71 (1), 378-395.

Drewes, Sarah, Stefan Ulbrich. 2009. *Mixed integer second order cone programming*. Verlag Dr. Hut Germany.

Esfahani, P Mohajerin, D. Kuhn. 2018. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171 (1), 115–166.

Fikar, Christian, Patrick Hirsch. 2017. Home health care routing and scheduling: A review. *Computers & Operations Research*, 77 86-95.

Gao, Rui, Anton Kleywegt. 2023. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48 (2), 603-655.

Ghosal, Shubhechyya, Chin Pang Ho, Wolfram Wiesemann. 2024. A unifying framework for the capacitated vehicle routing problem under risk and ambiguity. *Operations Research*, 72 (2), 425-443.

Ghosal, Shubhechyya, Wolfram Wiesemann. 2020. The distributionally robust chance-constrained vehicle routing problem. *Operations Research*, 68 (3), 716-732.

Green, Linda V. 2012. Om forum—the vital role of operations analysis in improving healthcare delivery. *Manufacturing & Service Operations Management*, 14 (4), 488-494.

Hashemi Doulabi, Hossein, Gilles Pesant, Louis-Martin Rousseau. 2020. Vehicle routing problems with synchronized visits and stochastic travel and service times: Applications in healthcare. *Transportation Science*, 54 (4), 1053-1072.

Jaillet, Patrick, Jin Qi, Melvyn Sim. 2016. Routing optimization under uncertainty. *Operations Research*, 64 (1), 186-200.

Jiang, Ruiwei, Siqian Shen, Yiling Zhang. 2017. Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research*, 65 (6), 1638-1656.

Julie, Redd. 2022. Top 10 complaints from home care clients. https://activatedinsights.com/articles/top-10-complaints-from-home-care-clients/. Accessed July 15, 2024.

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

Kong, Lu, Kejia Hu, Rohit Verma. 2022. Service chains' operational strategies: Standardization or customization? evidence from the nursing home industry. *Manufacturing & Service Operations Management*, 24 (6), 3099-3116.

Kong, Qingxia, Shan Li, Nan Liu, Chung-Piaw Teo, Zhenzhen Yan. 2020. Appointment scheduling under time-dependent patient no-show behavior. *Management Science*, 66 (8), 3480-3500.

Kumar Saha. 2020. The 10 most common complaints about home care and how we deal with them. https://www.myplacehomecare.ca/2020/09/10/the-10-most-common-complaints-about-home-care-and-how-we-deal-with-them/. Accessed July 16, 2024.

Laporte, Gilbert, François Louveaux, Hélène Mercure. 1992. The vehicle routing problem with stochastic travel times. *Transportation Science*, 26 (3), 161-170.

Lee, Hau L, Venkata Padmanabhan, Seungjin Whang. 1997. Information distortion in a supply chain: The bullwhip effect. *Management Science*, 43 (4), 546-558.

Liu, Mingda, Yanlu Zhao, Xiaolei Xie. 2024a. Continuity-skill-restricted scheduling and routing problem: Formulation, optimization and implications. *IISE Transactions*, 56 (2), 201-220.

Liu, Nan, Guohua Wan, Shan Wang. 2024b. Design of patient visit itineraries in tandem systems. *Manufacturing & Service Operations Management*, 26 (3), 972-991.

Liu, Sheng, Long He, Zuo-Jun Max Shen. 2021. On-time last-mile delivery: Order assignment with travel-time predictors. *Management Science*, 67 (7), 4095-4119.

Lu, Mengshi, Zuo-Jun Max Shen. 2021. A review of robust operations management under model uncertainty. *Production and Operations Management*, 30 (6), 1927-1943.

Mckinsey. 2022a. From facility to home: How healthcare could shift by 2025. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/from-facility-to-home-how-healthcare-could-shift-by-2025. Accessed December 27, 2022.

Mckinsey. 2022b. How 'care at home' ecosystems can reshape the way health systems envision patient care. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/how-care-at-home-ecosystems-can-reshape-the-way-health-systems-envision-patient-care. Accessed December 27, 2022.

Mosquera, Federico, Pieter Smet, Greet Vanden Berghe. 2019. Flexible home care scheduling. *Omega*, 83 80-95.

Munari, Pedro, Alfredo Moreno, Jonathan De La Vega, Douglas Alem, Jacek Gondzio, Reinaldo Morabito. 2019. The robust vehicle routing problem with time windows: Compact formulation and branch-price-and-cut method. *Transportation Science*, 53 (4), 1043-1066.

Naderi, Bahman, Mehmet A. Begen, Gregory S. Zaric, Vahid Roshanaei. 2023. A novel and efficient exact technique for integrated staffing, assignment, routing, and scheduling of home care services under uncertainty. *Omega*, 116 102805.

NHS. 2020. Coronavirus (covid-19): provision of home care. https://www.gov.uk/government/publications/coronavirus-covid-19-providing-home-care/coronavirus-covid-19-provision-of-home-care--2. Accessed December 27, 2022.

NHS. 2022. Help at home from a paid carer. https://www.nhs.uk/conditions/social-care-and-support-guide/care-services-equipment-and-care-homes/homecare/. Accessed December 27, 2022.

NHS. 2023. Managing heart failure home. https://www.england.nhs.uk/nhs-at-home/managing-heart-failure-at-home/. Accessed July 15, 2024.

Pahlevani, Delaram, Babak Abbasi, John W. Hearne, Andrew Eberhard. 2022. A cluster-based algorithm for home health care planning: A case study in australia. *Transportation Research Part E: Logistics and Transportation Review*, 166 102878.

Parent, Olivier, James P. LeSage. 2010. A spatial dynamic panel model with random effects applied to commuting times. *Transportation Research Part B: Methodological*, 44 (5), 633-645.

Prakash, A. Arun, Karthik K. Srinivasan. 2018. Pruning algorithms to determine reliable paths on networks with random and correlated link travel times. *Transportation Science*, 52 (1), 80-101.

Rostami, Borzou, Guy Desaulniers, Fausto Errico, Andrea Lodi. 2021. Branch-price-and-cut algorithms for the vehicle routing problem with stochastic and correlated travel times. *Operations Research*, 69 (2), 436-455.

Rowe, John W, Terry Fulmer, Linda Fried. 2016. Preparing for better health and health care for an aging population. *Jama*, 316 (16), 1643-1644.

Sarykalin, Sergey, Gaia Serraino, Stan Uryasev. 2008. Value-at-risk vs. conditional value-at-risk in risk management and optimization. *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*, chap. 13. INFORMS Tutorials in Operations Research (INFORMS, Catonsville, MD), 270-294.

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

Sessler, Daniel I. 2006. Non-pharmacologic prevention of surgical wound infection. *Anesthesiology Clinics of North America*, 24 (2), 279-297.

Solomon, Marius M. 1987. Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research*, 35 (2), 254-265.

Song, Hummy, Elena Andreyeva, Guy David. 2022. Time is the wisest counselor of all: The value of provider–patient engagement length in home healthcare. *Management Science*, 68 (1), 420-441.

Tsang, Man Yiu, Karmel S. Shehadeh. 2023. Stochastic optimization models for a home service routing and appointment scheduling problem with random travel and service times. *European Journal of Operational Research*, 307 (1), 48-63.

Wang, Yu, Yu Zhang, Minglong Zhou, Jiafu Tang. 2023. Feature-driven robust surgery scheduling. *Production and Operations Management*, 32 (6), 1921-1938.

World Health Organization. 2022. Ageing and life-course. `https://www.who.int/ageing/about/facts/en/`. Accessed December 27, 2022.

Xing, Tao, Xuesong Zhou. 2011. Finding the most reliable path with and without link travel time correlation: A lagrangian substitution based approach. *Transportation Research Part B: Methodological*, 45 (10), 1660-1679.

Yang, Meng, Yaodong Ni, Liu Yang. 2021. A multi-objective consistent home healthcare routing and scheduling problem in an uncertain environment. *Computers & Industrial Engineering*, 160 107560.

Zhan, Yang, Zizhuo Wang, Guohua Wan. 2021. Home care routing and appointment scheduling with stochastic service durations. *European Journal of Operational Research*, 288 (1), 98-115.

Zhang, Yu, Roberto Baldacci, Melvyn Sim, Jiafu Tang. 2019. Routing optimization with time windows under uncertainty. *Mathematical Programming*, 175 263–305.

Zhang, Yu, Zhenzhen Zhang, Andrew Lim, Melvyn Sim. 2021. Robust data-driven vehicle routing with time windows. *Operations Research*, 69 (2), 469-485.

Zhang, Zhenzhen, Yu Zhang, Roberto Baldacci. 2024. Generalized riskiness index in vehicle routing under uncertain travel times: Formulations, properties, and exact solution framework. *Transportation Science*, 58 (4), 761-780.

Zhou, Yun, Mahmut Parlar, Vedat Verter, Shannon Fraser. 2021. Surgical scheduling with constrained patient waiting times. *Production and Operations Management*, 30 (9), 3253-3271.