Les Cahiers du GERAD ISSN: 0711–2440

Consistent home health care routing and scheduling problem under time uncertainty

S.-S. Hosseini, Y. Adulyasak, L.-M. Rousseau

G–2023–45

October 2023
Revised: September 2025

CITATION ORIGINALE / ORIGINAL CITATION

S.-S. Hosseini, Y. Adulyasak, L.-M. Rousseau (January 2026). "Consistent home health care routing and scheduling problem under time uncertainty", *Transportation Research Part E: Logistics and Transportation Review*, vol. 205, p. 104509. https://doi.org/10.1016/j.tre.2025.104509.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2023
– Bibliothèque et Archives Canada, 2023

Legal deposit – Bibliothèque et Archives nationales du Québec, 2023
– Library and Archives Canada, 2023

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Consistent home health care routing and scheduling problem under time uncertainty

Seyede Saeede Hosseini [a, b, c]

Yossiri Adulyasak [a, d]

Louis-Martin Rousseau [b, c]

[a] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[b] *CIRRELT, Montréal (Qc), Canada, H3T 1J4*

[c] *Department of Industrial Engineering and Applied Mathematics, Polytechnique Montréal, Montréal (Qc), Canada, H3C 3A7*

[d] *Department of Logistics and Operations Management, HEC Montréal, Montréal (Qc), Canada, H3T 2A7*

seyede-saeede.hosseini@polymtl.ca
yossiri.adulyasak@hec.ca
louis-martin.rousseau@cirrelt.net

October 2023
Revised: September 2025
Les Cahiers du GERAD
G–2023–45

Abstract : This study addresses the challenge of routing and scheduling care workers for home health care logistics in a stochastic environment, where consistency in service delivery is crucial. The primary research question focuses on determining reliable schedules while ensuring timely care despite the uncertainty of travel and service times (TST). The objective is to maximize the number of new patients care workers can attend to while ensuring feasible and consistent schedules. To tackle this challenge, we propose a chance-constrained optimization modeling framework that ensures a likelihood of on-time arrivals, with arrival time distributions at patients estimated empirically and analytically via a discrete scenario set and an extreme value theory-based (EVT-based) approach, respectively. The EVT-based approximation incorporates nonlinear constraints that link patient visit times with the probability of on-time arrivals. The problem is decomposed into a master problem, which optimizes patient assignments, and subproblems, which generate feasible schedules and routes. To solve this problem, we propose a branch-and-check (B&Ch) algorithm, where the subproblems are solved efficiently via constraint programming. Computational results demonstrate that our solution approach, particularly with the EVT-based approximation, can efficiently handle practical benchmark instances while producing schedules with significantly higher service levels than the deterministic model in the literature.

Keywords : Service consistency; home health care routing; time uncertainty; branch-and-check; constraint programming

1 Introduction

Due to aging populations and governments' plans to decrease hospitalization costs (Lanzarone and Matta, 2014; Restrepo et al., 2020), home health care (HHC) has received significant attention over the past decade. It has become a highly active area of research (Wang et al., 2022; Xie et al., 2023). For example, in Canada, Deloitte (2021) estimates that HHC needs will increase by 615,479 patients, reaching a total of over 1.7 million by 2031. This is a 50% increase in demand compared to 2019. This surge in demand and an increase in the cost of care will double Canada's annual elder care costs to reach $58.5 billion in 2031.

The HHC industry can be classified into three structures: home hospitalization, home services, and home nursing services (Di Mascolo et al., 2017). These structures include various services, from basic housekeeping to professional care, such as physical therapy at people's homes. In HHC logistics, service quality plays an important role, and service consistency is necessary to ensure proper follow-ups and patient monitoring (Freeman and Hughes, 2010). AlayaCare, a Canadian company developing HHC software solutions for over 700 home care agencies worldwide, reported that, increasing care consistency is one of the most effective ways to improve care management (AlayaCare, 2023). Woodward et al. (2004) studied the importance of service consistency in HHC logistics by interviewing home care clients and their care workers. They state that consistency in care is supported by personnel consistency and time consistency. Supporting this, a survey of over 9,000 home care clients found that two of the ten most common complaints about home care agencies were the assignment of multiple care workers and irregularity in arrival times (Redd, 2022). Notwithstanding the prominent role of consistency in HHC, very few works in HHC scheduling consider it in their studies (Yang et al., 2021). Personnel consistency refers to assigning the same care worker or a limited number of different care workers to each client during their presence in the HHC system. This approach builds trust between clients and their assigned care workers over time and reduces information loss across care workers, thereby improving both care quality and service efficiency (Eveborn et al., 2006; Hewitt et al., 2016; Matta et al., 2014). In addition, time consistency means that service delivery to clients occurs at regular times at each visit. Receiving the service at regular times is essential for particular clients, especially those who need time-sensitive services, such as insulin injections, or generally those with time-specific routines (Eveborn et al., 2009). In addition to benefiting clients, time consistency also helps their families and relatives better plan their daily activities (Cappanera and Scutellà, 2022). In practice, employing care worker and time consistency yields greater patient satisfaction, which in turn fosters client loyalty and generates recurring revenue (Kovacs et al., 2014a; Nickel et al., 2012).

Many operations research problems are defined around HHC logistics planning issues in each strategic, tactical, and operational decision. One of the central and challenging problems at the operational level is the home health care routing and scheduling problem (HHCRSP) (Cissé et al., 2017). In HHCRSP, a health institution must decide the routes and times to provide health services at clients' homes for their care workers. This problem is a practical variant of the vehicle routing problem (VRP), which is very challenging in transportation planning and logistics (Di Mascolo et al., 2017). In the planning phase of HHCRSP, due to its complexity, decision makers often assume deterministic values for travel and service times (TST). However, during the execution phase, patient service times may be affected by random factors such as variations in patients' conditions (Shi et al., 2019; Lanzarone and Matta, 2014; Yuan et al., 2015), while travel times may fluctuate due to traffic congestion or weather conditions. Due to the stochastic nature of TST, care workers may arrive earlier or later than the scheduled appointment time. Ignoring the uncertainty of TST could result in inefficient (or even infeasible) schedules and service delays. While such delays may be considered trivial in contexts like grocery delivery, they are often unacceptable in HHC (Liu et al., 2025). In HHCRSP, with time consistency, a delay in service could result in patient dissatisfaction or pose risks to those whose treatments are time-sensitive. For example, patients who require medication provision at specific times could be severely affected by late service (Fikar and Hirsch, 2017).

Figure 1 illustrates the impact of travel time uncertainty on scheduling decisions, focusing on care worker A within a multi-worker, multi-day home health care scheduling problem. Case (1) represents the deterministic solution, while cases (2) and (3) show stochastic solutions with increasing travel time variability, where the coefficient of variation (CoV) is 25% and 50%, respectively. While the set of assigned patients remains unchanged between cases (1) and (2), the stochastic model adjusts routing and scheduling in response to uncertainty. In case (3), higher variability further alters assignments, routes, and schedules. Simulation results indicate that the average probability of delay for the deterministic solution is 23.4% at 25% CoV, rising to 33% at 50% CoV. In contrast, the stochastic model keeps the average probability of delay below 1% in both cases (2) and (3). These findings highlight the necessity of incorporating uncertainty into HHCRSP models to ensure robust scheduling, yet this remains an underexplored research area.

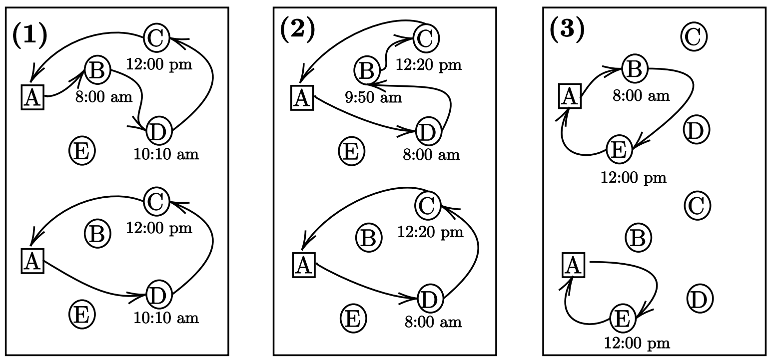

Figure 1: Robustness of Routes and Schedules Generated by Stochastic Model.
Note: The figure illustrates: (1) deterministic model, (2) stochastic model with lower travel time variation, and (3) stochastic model with higher travel time variation. Squares represent care worker A, while circles denote patients. Numbers below patient nodes indicate appointment times. Routes above correspond to Monday and Thursday, while routes below pertain to Tuesday, Wednesday, and Friday schedules for care worker A.

This research investigates the implications of uncertainty in TST and consistency in home health care logistics. While this work is inspired by the HHC industry, the planning problem is pertinent to the consistent VRP (ConVRP) due to the consideration of service consistency. The ConVRP has received significant attention in academic literature (Kovacs et al., 2014a). In addition to health care, the consistency feature is important in many fields such as the inventory routing problem (Diabat et al., 2021; Coelho et al., 2012), small package delivery (Groër et al., 2009), service network design (Liu et al., 2023; Yao et al., 2021), and rental-with-driver services (Mancini and Gansterer, 2021). To the best of our knowledge, no algorithms have been developed for the ConVRP that address hard consistency under stochastic TST. A recent study by Yang et al. (2021) introduced uncertainty in TST within a consistent HHCRSP, while assuming soft time windows for patients. In contrast, we impose hard time windows, meaning that the service cannot start before the beginning of the specified time frame. This consideration makes the problem more complex (Zhang et al., 2019); however, our approach can be adapted to handle soft time windows as well. Our problem shares similarities with the one outlined by Heching et al. (2019), but we expand its scope to incorporate service level constraints under stochastic TST. The service level requirements are guaranteed through a set of probabilistic constraints where hard time windows are imposed. In practice, this problem is solved to periodically update existing schedules of a home hospice care company to accommodate changes in patients' requests, including their admission to or discharge from the company. We aim to determine feasible schedules for care workers that maximize the number of newly accepted patients, or equivalently the total number of patients served, while respecting consistency. The problem of our interest comprises assignment and scheduling decisions. More specifically, the planner must simultaneously determine the optimal assignments of patients, care workers, and visit days, as well as the schedules for each care worker to visit the assigned patients that meet the service level requirements.

Our contribution is as follows. First, we present two formulations for the stochastic HHCRSP under uncertain travel and service times: one based on a discrete set of scenarios that characterize possible trip and service time realizations, and the other based on an extreme value theory (EVT) approximation. The first model relies on scenarios generated through a sampling approach but the model can be intractable when the number of scenarios is large (Gendreau et al., 2016). On the other hand, the EVT-based model eliminates the need for using a large number of scenarios as it utilizes a parametric approach to approximate the underlying probabilistic distributions of arrival times. At the same time, it introduces complexity due to the nonlinear constraints associated with the EVT-based function. In addition, we extend the EVT-based modeling framework to handle the case of dependent travel times, which can be influenced by common factors (such as weather or traffic conditions). Next, we propose a branch-and-check (B&Ch) algorithm to solve these models. In B&Ch, the problem is decomposed into a master problem (care worker and patient assignments) and subproblems (routing and scheduling of care workers). The subproblems are modeled using constraint programming (CP), which effectively handles both scenario-based and EVT-based formulations. Finally, we numerically validate the performance of our two modeling frameworks and the benefits of considering uncertainty at the planning level by comparing the service levels and optimal values of our stochastic models with those of the deterministic model proposed by Heching et al. (2019). Despite the fact that the stochastic and deterministic models generally accept a similar number of new patients, the deterministic model significantly underperforms in achieving the minimum service levels, while the stochastic models consistently fulfill this requirement. Out of the two stochastic models, the EVT-based model outperforms the scenario-based model in terms of computational efficiency while still producing high-quality solutions that meet service level requirements. We also present the numerical analyses for the case of dependent travel times.

The remainder of this paper is organized as follows. Section 2 discusses the recent literature on the consistent HHCRSP. Section 3 outlines mathematical formulations and solution approaches for the stochastic HHCRSP. Section 4 presents the experimental results, and Section 5 concludes with a summary of our work and future research directions.

2 Literature review

Although there is no standard version of the HHCRSP, it is closely related to the VRP (Grenouilleau et al., 2020; Di Mascolo et al., 2017). The service consistency feature in the HHCRSP makes a model similar to the ConVRP (Kovacs et al., 2014a). Nevertheless, the HHCRSP involves many additional operational constraints (Cissé et al., 2017). Our primary focus is on integrating consistency and uncertainty in travel and service times within the home health care context. To maintain a clear scope, we do not include studies on HHCRSP without consistency requirements in the main review; however, Table 1 compares our work with selected recent studies that address stochastic travel and service times but ignore consistency requirements. The table highlights key characteristics, including the planning horizon (single- vs. multi-period), the number of care workers (single vs. multiple), the treatment of time windows (patient-specified, care worker-specified, or endogenously determined), as well as the modeling approach for uncertainty and the solution methodology employed. Hereafter, the term HHCRSP refers specifically to the problem with consistency requirements.

Our work ensures consistency in personnel and visit times in the presence of uncertainty in travel and service times, which sets it apart from the existing literature. To our knowledge, no algorithms have been published for ConVRP with stochastic travel and service times, despite the significant attention ConVRP has received in academic literature (Kovacs et al., 2014a). In this section, we survey papers dealing with consistent VRP and HHCRSP. Table 2 summarizes relevant papers reviewed in this section. For the papers dealing with uncertain parameters, we indicate the stochastic parameters in the *Stochastic Information* column where "**-**" indicates that the paper considers a deterministic solution approach with no uncertain parameters. The *HHCRSP* column indicates whether papers deal with a home health care routing and scheduling problem.

Table 1: Recent literature on HHCRSP with stochastic travel and service times

Author(s)	Horizon	Care Worker	Time Window	Modeling Approach	Solution Method
Shi et al. (2018)	S	M	P	SPR	Simulated Annealing
Shi et al. (2019)	S	M	P	RO	Simulated Annealing/ Tabu Search
Liu et al. (2019)	S	M	P	CCP	Branch-and-Price
Zhan et al. (2021)	S	S	E	SPR	L-Shaped/Modified TSP
Tsang and Shehadeh (2023)	S	S	E/C	DRO	CPLEX
Naderi et al. (2023)	M	M	C	RO	Logic-Based Benders Decomposition
Liu et al. (2025)	S	M	P/C	DRO	Branch-Price-and-Cut
This paper	M	M	P/C/E	CCP	Branch-and-Check

M, Multiple; S, Single; P, Patient; C, Care Worker; E, Endogenous(Time Window Assignment); SPR, Stochastic Programming with Recourse; RO, Robust; DRO, Distributionaly Robust Optimization; CCP, Chance-Constrained Programming.

Table 2: Summary of related literature on ConVRP and HHCRSP

	Consistency		Solution Type		Stochastic	
Author(s)	Time	Personnel	Exact	Heuristic	**Information**	**HHCRSP**
Groër et al. (2009)	Hard	Hard	-	RTR[1]	-	No
Sungur et al. (2010)	-	Soft	-	TS[2]	Customer and Service Time	No
Bennett and Erera (2011)	Hard	-	-	Capacity-based	-	Yes
Tarantilis et al. (2012)	Hard	Hard	-	TS	-	No
Kovacs et al. (2014b)	Hard	Hard	-	ALNS[3]	-	No
Kovacs et al. (2015a)	Soft	Hard	-	LNS[4]	-	No
Spliet and Gabor (2015)	Hard	-	BPC[5]	-	Demand	No
Kovacs et al. (2015b)	Soft	Soft	ϵ[6]	MDLNS[7]	-	No
Jabali et al. (2015)	Soft	-	-	TS/LP[8]	Travel Times	No
Subramanyam and Gounaris (2017)	Hard	-	Decomp[9]	-	-	No
Subramanyam et al. (2018)	Hard	-	Decomp	-	Demand and Travel Times	No
Goeke et al. (2019)	Hard	Hard	CG[10]	LNS	-	No
Heching et al. (2019)	Hard	Hard	LBBD/B&Ch	-	-	Yes
Yang et al. (2021)	Soft	Soft	-	Artificial bee colony	Travel and Service Times	Yes
Demirbilek et al. (2021)	Hard	Hard	-	Scenario-based	Demand	Yes
Wang et al. (2021)	Hard	Hard	CG	-	-	No
Yang et al. (2022)	-	Soft	-	LNS-SA[11]	Demand, Travel and Service Times	No
Cappanera and Scutellà (2022)	Hard	Hard	-	Pattern-based	Demand	Yes
Güven-Koçak et al. (2024)	-	Soft	-	Petal-based	-	Yes
Liu et al. (2024)	-	Hard	BPC[5]	-	Travel and Service Times	Yes
Alvarez et al. (2024)	-	Soft	B&C[12]/BD[13]	-	Customer and Demand	No
Yu et al. (2024)	Hard	Hard	-	ALNS[3]	-	No
This paper	Hard	Hard	B&Ch	-	Travel and Service Times	Yes

[1] Record-To-Record, [2] Tabu Search, [3] Adaptive Large Neighborhood Search, [4] Large Neighborhood Search, [5] Branch-Price-and-Cut, [6] ϵ-constraint, [7] Multi-Directional LNS, [8] Linear Programming, [9] Decomposition, [10] Column Generation, [11] Simulated Annealing, [12] Branch-and-Cut, [13] Benders Decomposition

In routing problems in which customers require services periodically over time, consistency in visit time can play a significant role in improving service quality. Groër et al. (2009) introduced a new variant of the periodic VRP in which service consistency is incorporated into the problem and called it ConVRP. In this multiday VRP, in addition to the constraints on vehicle capacity and route length, there are additional consistency requirements, i.e., each customer must be served by the same driver (personnel consistency) at approximately the same time on each day (time consistency) when the service takes place. Several studies, e.g., Groër et al. (2009); Tarantilis et al. (2012), and Kovacs et al. (2014b), proposed heuristics to determine a template-based solution with consistency considerations in which a set of template routes (also called priority routes) including only customers requiring service on multiple days (frequent customers) is generated. Then, for each day, the daily routes are constructed by removing the customers who do not ask for service on that day and by inserting customers who require service on only that day (non-frequent customers). A generalized consistent vehicle routing problem was introduced by Kovacs et al. (2015a) in which the maximum difference in arrival times is penalized in the objective function instead of it being considered a hard constraint. To solve their problem, they proposed a large neighborhood search (LNS) heuristic, which outperforms the template-based heuristics of Groër et al. (2009); Tarantilis et al. (2012), and Kovacs et al. (2014b) in terms of both travel cost and time consistency. A multi-objective ConVRP that combines the consistency and cost objectives was considered by Kovacs et al. (2015b), and they proposed a multi-directional large neighborhood search heuristic to solve it. Yu et al. (2024) addressed a new ConVRP that, in addition to time and personnel consistency, incorporates route consistency and driver equity. Goeke et al. (2019) proposed an exact solution approach based on a column-and-cut generation (CCG) procedure to solve the ConVRP with driver and time consistencies. They also developed an LNS heuristic to tackle large instances. Recently, Wang et al. (2021) considered the ConVRP with route consistency and proposed an exact solution approach based on set partitioning-based models and CCG techniques to solve it.

Most of the studies in the ConVRP consider the deterministic case when all the parameters are assumed to be perfectly known. In other words, information related to customer demands, customer presence, service times, and travel times is often considered fully available prior to optimization. There are, however, a few papers that explore the approaches under stochastic information. For example, Sungur et al. (2010) studied a variant of VRP with soft time windows, called the courier delivery problem, in which service times and customer presence are uncertain. They modeled these uncertainties through robust optimization and scenario-based stochastic programming approaches. Nevertheless, they did not consider service consistency explicitly, but such consistency is encouraged by maximizing the similarity of routes across multiple scenarios. A tabu search heuristic was employed to solve this problem. In their study, Yang et al. (2022) explored a variant of ConVRP that focuses only on personnel consistency. They introduced a novel metric to quantify driver consistency and incorporated it into a weighted objective function alongside travel times. To address uncertainties in demand, travel times, and service times, they applied uncertainty theory, transforming these elements into a deterministic model. The problem was tackled using a hybrid algorithm combining LNS with simulated annealing techniques. Similar to Yang et al. (2022), Alvarez et al. (2024) studied soft personnel consistency under uncertainty. They penalized violations of a consistency target in the objective function as well as expected routing costs and penalties for unserved visits. The problem was modeled as a two-stage scenario-based stochastic programming: in the first stage, customers are assigned to drivers, in then in the second stage, for each scenario, a subset of customers is selected and routing decisions are made. They solved this problem using both branch-and-cut and Benders decomposition methods.

A similar VRP variant to ConVRP is the time window assignment vehicle routing problem (TWAVRP) (Spliet and Gabor, 2015). In this problem, the authors assumed that demands are not known when the decision-maker assigns time windows (with fixed-size width) to their customers. Once the demand is realized, they decide on a vehicle routing schedule to satisfy the demand of each customer at the assigned time window. In TWAVRP, however, personnel consistency is not considered, and only time consistency is imposed. Later on, Subramanyam et al. (2018) extended the uncertainty vector to incorporate demand and travel time uncertainties through a set of discrete scenarios. They employed a two-stage stochastic programming framework in which the first-stage decisions comprise time window assignments, and the second-stage decisions comprise routes to serve customers based on the assigned time windows. They adapted the decomposition algorithm of Subramanyam and Gounaris (2017) to solve their TWAVRP in an exact manner. Jabali et al. (2015) studied a self-imposed TWAVRP in which the provider decides the time window for customers. In this case, time windows are considered endogenous to the routing problem. Their model assumes that travel times are uncertain when time windows are assigned. They applied a buffer allocation model in order to protect scheduled time windows against travel time uncertainty. To solve this problem, they applied a tabu search heuristic to determine routing decisions. Afterward, they used a linear programming model to optimize the schedules for customers based on the routing decisions previously determined in the first step.

Consistency can be considered as one of the main features affecting HHC quality services (Woodward et al., 2004). However, despite its practical importance, it has not been extensively studied in the literature, and it only started gaining attention from 2018 (Di Mascolo et al., 2021). Personnel consistency, which is often referred to as continuity of care (Borsani et al., 2006), has been studied in the literature more than time consistency (Cappanera and Scutellà, 2022). Di Mascolo et al. (2021) emphasize the lack of approaches concerning consistency in time. In other words, there are only a few papers that study HHCRSP models with both personnel and time consistency simultaneously.

Bennett and Erera (2011) handled hard time consistency for a single-nurse HHCRSP through a rolling horizon myopic planning approach. They assumed that patients' requests are revealed dynamically over time. Their heuristic approach, which considers remaining schedule time when inserting patient requests, generally outperformed traditional distance-based methods in terms of accepted patients and visits per day. Similarly, Demirbilek et al. (2021) dealt with hard time consistency with a dynamic patient set, but for a multiple-nurse case of HHCRSP. Additionally, they considered personnel consistency for all accepted patients. They developed a scenario-based heuristic approach. This approach determines the acceptance of new patients and their schedules by generating several scenarios for future requests, in addition to considering the current schedules of the nurses.

Heching et al. (2019) studied a home health care delivery problem and they proposed an exact optimization method, called logic-based Benders decomposition (LBBD), to maximize the number of newly accepted patients while respecting service consistency for existing patients. The problem they addressed concerned updating a home hospice care company's weekly schedule in response to patient population changes to predict their staffing needs. Nevertheless, Heching et al. (2019) assumed deterministic TST. However, the routes and schedules determined by the deterministic model could lead to significant delays in services and treatments in the presence of travel and service times uncertainty. In our work, we aim to extend this context to consider stochastic TST and demonstrate the value of the stochastic optimization framework, which can improve patient service levels through reduced risks of service delays. Later on, Güven-Koçak et al. (2024) addressed a daily HHCRSP with only soft personnel consistency over a rolling horizon approach. They considered the existing schedule as a baseline and employed an updating process in response to frequent schedule changes. In their paper, the degree of consistency in patient-care worker assignments can be set by the cost of changing assignments.

To our knowledge, no literature on HHCRSP with hard service consistency tackles the challenge of stochasticity in TST. Cappanera and Scutellà (2022) addressed personnel and time consistency under demand uncertainty using a pattern-based heuristic that schedules visits based on predefined visit patterns. Yang et al. (2021) considered a multi-objective consistent home health care routing and scheduling problem with uncertain travel and service times in which consistencies were enforced through soft constraints. Three objectives were considered in this work: minimizing routing cost, increasing service consistency, and improving workload balance. The authors solved this problem heuristically via a multi-objective artificial bee colony algorithm. In addition to the fact that our focus is on the exact algorithmic framework, one distinct aspect of the problem considered in our work versus the problem in Yang et al. (2021) is that we incorporate complex uncertainty of arrival times that arise from hard time window restrictions in our stochastic framework. Most papers in HHC routing and scheduling problems assume hard time windows for service start times due to the time-sensitive nature of many tasks (Fikar and Hirsch, 2017; Di Mascolo et al., 2021). Furthermore, since the arrival time cannot be calculated by the sum of independent travel times of the arcs, hard time windows are more difficult to solve exactly (Hoogeboom et al., 2021). Additionally Yang et al. (2021) assumed independent uncertain travel times, but we also include dependency.

A recent study by Liu et al. (2024) addressed an HHCRSP that incorporated only personnel consistency, assuming exogenous patient time windows and accounting for stochastic TST. They applied a chance-constrained approach to model uncertainty using sample average approximation, and they proposed a branch-price-and-cut algorithm to solve the problem exactly. While our work shares some similarities with theirs, it differs in several important aspects. First, we determine endogenous consistent time windows for patients, which is inherently different from the classical VRP with time windows (Jabali et al., 2015). Jointly routing decisions and time window assignments under TST uncertainty adds complexity to the decision-making process. Second, unlike their model, we treat the visiting pattern for each patient as a decision variable rather than a fixed input, which expands the solution space for each accepted patient and improves scheduling flexibility. Most importantly, their chance-constrained model relied only on a discrete set of scenarios, which required more computational efforts as the number of scenarios increased. In contrast, we address this scalability issue by introducing an analytical approximation based on EVT which improves computational efficiency while maintaining high-quality solutions.

3 HHCRSP formulation and solution framework

This work extends the deterministic problem presented by Heching et al. (2019), which was derived from a real-world HHCRSP. The notation for sets and parameters is provided in Table 3. In this study, a set of patients P must be scheduled with a set of care workers C over a week, represented by set D. While a weekly schedule is common in many HHC problems (Hewitt et al., 2016), our framework can be generalized to accommodate longer planning horizons. The scheduling operates on a rolling basis: each *existing* patient has pre-scheduled visits and a pre-assigned care worker, which must remain unchanged to ensure service consistency. *New* patients may also be added to the schedule.

Table 3: Description of notation used in HHCRSP model

Set	Description
P	Set of all patients
P_{new}	Set of new patients
P_{existing}	Set of existing patients
C	Set of care workers
D	Set of days
Q_p	Set of required qualifications to serve patient p
Q_c	Set of qualifications of care worker c
D_p	Set of pre-assigned days to patient $p \in P_{\text{existing}}$
K_p	Set of vectors a feasible combination of visit days to the patient p
$\tilde{P}_{c,d}$	Set of patients assigned to care worker c on day d, determined by master problem
$\tilde{Z}_{c,d}$	Set of nodes in the route of care worker c on day d, including $\tilde{P}_{c,d}$ and care worker's nodes
I	Set of all nodes, including care workers and patients
Parameter	**Description**
w_p	Number of visits required for patient p
$[r_p, e_p]$	Time window of patient p
$\bar{s}_p$	Nominal service time of patient p
c_p	Pre-assigned care worker to patient $p \in P_{\text{existing}}$
l_p	Location of patient p
$[r_c, e_c]$	Time window of care worker c
l_c	Starting location of care worker c
l'_c	Ending location of care worker c
h_c	Maximum work time over the week of care worker c
$\bar{t}_{i,i'}$	Nominal travel time between node i and $i' \in I$
$\mathcal{L}$	Maximum allowable delay
α	Minimum acceptable service level
Pr	Probability function

Each patient requires a single care worker with matching qualifications to observe personnel consistency (same care worker for each visit). The assumption that a single care worker is assigned to each patient over the planning horizon is common in ConVRP literature. This should not hinder the practical application of our approach in addressing real-world problems. This is because the assignment of care workers to patients is a long-term decision not affected by the time aspect (Cissé et al., 2017; Song et al., 2020). However, we discuss later how to modify the proposed model when assuming multiple care workers. Additionally, each patient has a specified number of weekly visits, all of which are of the same type and have a nominal duration representing the average visit length. To maintain time consistency, all visits must occur at the same scheduled time within the patient's fixed time window across all assigned days. This means that once an appointment time is determined, it remains unchanged throughout the planning horizon. As a result, the model determines a narrower endogenous time window for each patient, allowing them to better organize their daily routines (Hoogeboom et al., 2021). The specific visiting pattern for each patient, referring to the selection of visit days, is also treated as a decision variable. Patients requiring less frequent visits (e.g., two or three times per week) may have constraints on visit days, specified by $\boldsymbol{K}$. For example, for patients with two weekly visits, close visits like Mondays and Tuesdays must be avoided through a separation constraint, which is incorporated into our mathematical formulation (discussed in Section 4).

Each care worker possesses a defined set of qualifications. Their daily working hours must align with their fixed time window that remains consistent across all days of the week. Additionally, their total weekly working hours must not exceed a specified maximum limit. Each care worker's route starts at their designated starting location, visits assigned patients, and returns to their ending location. Therefore, in this context, care workers are considered heterogeneous with respect to their skills and availabilities, and they may also depart from and return to different depots. The average travel time between any two locations is given as a nominal travel time.

The time constraint is hard, i.e., service cannot start earlier than the scheduled time, and care workers must wait if they arrive early (Bertels and Fahle, 2006). On time arrival is ensured through probabilistic constraints, defined by a service level α set by the decision-maker for all patients. We assume that an arrival delayed by at most $\mathcal{L}$ time units is still considered on time. This assumption can be easily relaxed by setting $\mathcal{L} = 0$. As noted in Fikar and Hirsch (2017); Cissé et al. (2017), in single-period HHC problems, traditional vehicle routing objectives, such as minimizing travel distance or travel costs, are a primary concern. However, in multi-period HHC problems, greater emphasis is placed on staffing and service-related factors, such as maximizing the number of assigned services or minimizing the number of care workers. In this regard, similar to Bennett and Erera (2011) and Nickel et al. (2012), our model aims to maximize the number of newly accepted patients, or equivalently, the total number of patients served, while ensuring consistency and meeting service level constraints. Additionally, maximizing the number of patients served is a widely used measure of care worker productivity, which is indirectly influenced by reduced travel time and idle time (Bennett and Erera, 2011; Heching et al., 2019). Taking all of these elements into account, this problem can be classified as a stochastic heterogeneous home healthcare routing and scheduling problem with multiple depots and endogenous time window assignment.

In the following section, a mathematical formulation for this problem is presented. The formulation is described by a master problem (MP) that addresses patient acceptance, care worker assignment, and visit days scheduling, along with subproblems (SPs) that determine appointment times and routing for each care worker. An overview of the decision variables used in MP and SPs is provided in Table 4.

3.1 Master Problem (MP)

The master problem defines three sets of decision variables: binary variable δ_p to determine whether patient p is accepted, binary variable $x_{c,p}$ to determine whether patient p is assigned to care worker c, and binary variable $y_{c,p,d}$ to determine whether patient p is visited by care worker c on day d. Additionally, $\boldsymbol{y}$ is a vector of variables $y_{c,p,d}$ that is described by a feasible set K_p. The master

Table 4: Decision variables in master problem and subproblem

Master Problem:	
δ_p	1 if patient p is accepted; and 0 otherwise
$x_{c,p}$	1 if patient p is assigned to care worker c; and 0 otherwise
$y_{c,p,d}$	1 if patient p is visited by care worker c on day d; and 0 otherwise
Subprblem:	
$\pi_{z,d}$	Interval variable for visit to node z on day d
ν_p	Integer variable for appointment time of patient p across all visits
λ_d	Integer variable for amount of work time of the care worker

problem is formulated as follows:

$$\max \sum_{p \in P_{\text{new}}} \delta_p \tag{1}$$

$$\text{subject to (s.t.)} \sum_{c \in C | Q_p \subseteq Q_c} x_{c,p} = \delta_p \qquad \forall p \in P, \tag{2}$$

$$y_{c,p,d} \leq x_{c,p} \qquad \forall c \in C, \forall p \in P, \forall d \in D, \tag{3}$$

$$\sum_{c \in C} \sum_{d \in D} y_{c,p,d} = w_p \delta_p \qquad \forall p \in P, \tag{4}$$

$$\delta_p = 1 \qquad \forall p \in P_{\text{existing}}, \tag{5}$$

$$x_{c,p} = 1 \qquad \forall p \in P_{\text{existing}}, c = c_p, \tag{6}$$

$$y_{c,p,d} = 1 \qquad \forall p \in P_{\text{existing}}, c = c_p, \forall d \in D_p, \tag{7}$$

$$\boldsymbol{y} \in K_p \qquad \forall p \in P, \tag{8}$$

$$\sum_{d \in D} \sum_{p \in \hat{P}_{c,d} | p \in P_{\text{new}}} (1 - y_{c,p,d}) \geq 1 \qquad \forall c \in C, \tag{9}$$

$$\delta_p, x_{c,p}, y_{c,p,d} \in \{0, 1\} \qquad \forall c \in C, \forall p \in P, \forall d \in D. \tag{10}$$

The objective (1) maximizes the number of newly accepted patients. Constraints (2) ensure that if a patient is accepted, one dedicated care worker will be assigned to that patient. Constraints (3) enforce that patients are only visited by care workers assigned to them. Constraints (4) ensure that the number of visits required by each accepted patient is satisfied. Constraints (5),(6), and (7) impose the inclusion, care worker assignment, and visit days assignment, respectively, for existing patients based on existing consistency requirements. Constraints (8) enforce that visits to patients must respect the feasible visit days described by set K_p. Note that this constraint is consistent with what is presented by Heching et al. (2019), and we explain in Section 4 how it can be described using a set of linear inequalities. Constraints (9) are no-good cuts, each of which removes at least one of the current assignments that produced the corresponding infeasible subproblem. These constraints will be generated by solving the SPs iteratively in the B&Ch procedure. Here, $\hat{P}_{c,d}$ includes the set of assigned patients to care worker c on day d that makes their schedule infeasible.

Indeed, some constraints can be modified to incorporate different requirements. For instance, if we want to permit each patient to be visited by a limited number of care workers instead of just one, we can replace constraints (2) with $\sum_{c \in C | Q_p \subseteq Q_c} x_{c,p} \leq B\delta_p$, where B represents the allowed limit. We can also remove constraints (7) to provide flexibility in visit days scheduling for existing patients.

3.2 Subproblem (SP)

The subproblem generates a feasible schedule of visits that is compatible with time windows, maximum work time, and service level constraints for each care worker based on the assignment decisions made in the MP. The SP is modeled as a constraint satisfaction problem (CSP) with (nonlinear) service level constraints. To solve this complex scheduling problem, we use constraint programming for its ability to find feasible schedules quickly. The effectiveness of CP in solving a variety of scheduling problems has been demonstrated in recent studies (Hashemi Doulabi et al., 2016; Heching et al., 2019; Roshanaei et al., 2020; Martínez et al., 2022; Elçi and Hooker, 2022). For readers who are less familiar with constraint programming modeling, Table 5 presents the CP functions and global constraints used in our SP model, along with their definitions (IBM, 2025).

Table 5: Functions and global constraints used in SP model

Functions:	
StartOf	It returns the start of an interval variable.
LengthOf	It returns the length of an interval variable.
Global Constraints:	
NoOverlap	It constraints a set of interval variables not to overlap with each other.
FirstOf	It constraints an interval variable to be the first in the set of interval variables.
LastOf	It constraints an interval variable to be the last in the set of interval variables.
Span	It creates an interval variable that starts at the beginning of the earliest interval and ends at the end of the latest interval in a given set of interval variables.

Given a solution vector $\boldsymbol{y}$ determined by MP, we define a set of patients assigned to a care worker c on day d as $\tilde{P}_{c,d} = \{p|\, \tilde{y}_{c,p,d} = 1\}$ and a set of nodes in the route of care worker c on day d as $\tilde{Z}_{c,d} = \{z|z \in \{l_p : p \in \tilde{P}_{c,d}\} \cup \{l_c, l'_c\}\}$. As for scheduling, CP provides a particular variable type called the *interval variable*, which represents a time interval during which a task is performed. The start and end times of each interval variable are decision variables whose domain is a subset of $\{[\eta, \beta)|\eta, \beta \in \mathbb{Z}, \eta \leq \beta\}$, and their difference determines the length of the interval variable (Naderi et al., 2023). In this context, η and β represent the earliest possible start and latest possible end of the interval variable, respectively. While CP represents time points as integers, a sufficiently wide time horizon makes time effectively continuous (Laborie et al., 2018). We define our interval variable $\pi_{z,d}$ to represent the visit to node $z \in \tilde{Z}_{c,d}$ on day d. Throughout this paper, the terms *node* and *location* are used interchangeably. The domain of $\pi_{z,d}$ is determined by the time window associated with node z. Additionally, the length of each interval variable $\pi_{z,d}$ is characterized by its nominal service time $\bar{s}_z$, which is a predefined parameter indicating the duration of the visit at node z. If node z corresponds to the care worker's location, $\bar{s}_z$ is set to 0. For each patient p, the variable ν_p represents the appointment time. If the SP determines that there is no feasible solution, we add no-good cuts (9) to the master problem to eliminate the current solution from the feasible space of the MP. Note that the bar signs above parameters s_z indicate the expected values of these parameters. We use nominal values for travel and service times in the constraints and the domain of decision variables to ensure that our model satisfies the same time window constraints as the deterministic model (Heching et al., 2019) whereas the service level requirements are guaranteed by incorporating probabilistic constraints. The SP for care worker c can be defined as follows:

$$\text{NoOverlap}(\{\pi_{z,d}|z \in \tilde{Z}_{a,d}\}) \qquad \forall d \in D, \tag{11}$$

$$r_p \leq \nu_p \leq e_p - \bar{s}_p \qquad \forall p \in \tilde{P}_{c,d}, \tag{12}$$

$$\text{FirstOf}(\{\pi_{z,d}|z \in \tilde{Z}_{c,d}\}) = l_c \qquad \forall d \in D, \tag{13}$$

$$\text{LastOf}(\{\pi_{z,d}|z \in \tilde{Z}_{c,d}\}) = l'_c \qquad \forall d \in D, \tag{14}$$

$$\text{Startof}(\pi_{l_c,d}) = r_c, \text{Startof}(\pi_{l'_c,d}) = e_c \qquad \forall d \in D, \tag{15}$$

$$\text{StartOf}(\pi_{l_p,d}) = \nu_p \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \tag{16}$$

$$\lambda_d = \text{Span}(\{\pi_{l_p,d}|p \in \tilde{P}_{c,d}\}) \qquad \forall d \in D, \tag{17}$$

$$\sum_{d \in D} \text{LengthOf}(\lambda_d) \leq h_c \tag{18}$$

$$Pr\{\text{StartOf}(\pi_{l_p,d}) \leq \nu_p + \mathcal{L}\} \geq \alpha \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \tag{19}$$

$$\nu_p \in \mathbb{Z}, \pi_{z,d} \subseteq [r_z, e_z) \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall z \in \tilde{Z}_{c,d}. \tag{20}$$

Constraints (11) represent global constraints in CSP (Baptiste et al., 2001) and indicate that patient visits assigned to care worker c do not overlap while accounting for nominal travel time between nodes (i.e., $\bar{t}_{i,i'}$). Constraints (12) impose that the appointment time must fall within the patient's time window. Constraints (13) and (14)) state that the care worker always begins at the starting location l_c and returns to the ending location l'_c at the end of the day. Constraints (15) fix the start of the care worker's interval variable at the starting and ending locations, based on the care worker's time window. To ensure time consistency, constraints (16) require the start of patient service to align with their appointment time. Constraints (17) impose span constraints on interval variables (Laborie, 2009) to define that the care worker's workday on day d begins with the visit to the first patient in $\tilde{P}_{c,d}$ and ends with the departure from the last patient in $\tilde{P}_{c,d}$. Constraint (18) limits the maximum expected work time of the care worker. Service level constraints are defined in constraints (19) and will be explained in subsequent sections. Finally, constraints (20) define the domains of the variables.

For an existing patient p with an appointment time ι, we need to modify the time window $[r_p, e_p]$ to $[\iota, \iota + \bar{s}_p]$. However, we can leave the time window unchanged to allow flexibility. Note that constraints (19) take the form of probabilistic chance constraints, which limit the risk of delays at a given service level based on the decision-maker (Gendreau et al., 2014). More specifically, constraints (19) guarantee that, for each patient, care worker c will arrive on time with a given probability α. Therefore, $(1 - \alpha)$ is the maximum risk of arriving late that the decision-maker can accept. The arrival time at each patient location primarily depends on the travel times along the traversed arcs and the service durations of previously visited patients. To respect time consistency, care workers cannot start their service sooner than the appointment time, as patients may not be available before their appointment. As a result, the arrival time is also influenced by the start-service time at preceding patients. Under stochastic travel and service times, both arrival and start-service times become inherently random. Recall that $\mathcal{L}$ represents an acceptable buffer of delay. In other words, an actual arrival after $\nu_p + \mathcal{L}$ is considered a delay. We propose to model the chance constraint using scenario-based and EVT-based models, described in the next subsections.

3.2.1 Scenario-based model

In the scenario-based model, uncertainty is represented by a discrete set of scenarios, N, generated from travel and service time distributions through a Monte Carlo sampling approach. Since optimization is agnostic to the scenario generation approach in this model, one can incorporate complex scenario generation methods that also consider dependencies and correlations among uncertain parameters (i.e., travel and service times in our case). We substitute constraints (19) with a sample average approximation based on the samples generated using this method. To do so, we define new variables for the arrival time and start-service time for patient p in sample n on day d, that is, $a^n_{p,d}$ and $\tau^n_{p,d}$, respectively. These variables introduce new constraints to the problem, expressed as:

$$\tau^n_{p,d} \geq a^n_{p,d},\ \tau^n_{p,d} \geq \nu_p \quad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N.$$

Given the TST realization for each scenario, we can check whether the inequality $a^n_{p,d} \leq \nu_p + \mathcal{L}$ is satisfied. For a given route, we can calculate the arrival time at patient p based on the start-service time of the previous patient, the service time of the previous patient, and the travel time from the previous patient $p-1$ as follows: $a^n_{p,d} = \tau^n_{p-1,d} + s^n_{p-1} + t^n_{p-1,p}$. To keep track of the number of on-time arrivals for patient p, the binary variable $\gamma^n_{p,d}$ is set to 1 when we arrive on time; otherwise, it is set to 0. The sample average of $\gamma^n_{p,d}$ must be equal to or greater than α, that is, $\frac{1}{|N|}\sum_{n \in N} \gamma^n_{p,d} \geq \alpha \quad \forall p \in \tilde{P}_{c,d}, \forall d \in D$. Note that, in mixed-integer linear programming (MILP), the binary value of $\gamma^n_{p,d}$ is typically done through the use of disjunctive (big-M) constraints: $a^n_{p,d} \leq \nu_p + \mathcal{L} + M(1 - \gamma^n_{p,d})$. A large number of big-M constraints in MILP models can lead to weak LP relaxations (Naderi et al., 2023). However, with CP, we can avoid the need for big-M formulations. By leveraging CP's logical constraints and inference algorithms (Hooker, 2002), binary variables ($\boldsymbol{\gamma}$) can be readily determined as follows: $\gamma^n_{p,d} == (a^n_{p,d} \leq \nu_p + \mathcal{L}) \quad \forall p \in \tilde{P}_{c,d}, \forall d \in D$. That is, $\gamma^n_{p,d}$ is set to 1 if the condition inside parentheses holds; otherwise, it is set to 0. We present the MILP Formulation of the scenario-based model in Appendix A and compare its computational performance with that of CP-based model.

Despite its generalizability and flexibility, the scenario-based model can face scalability issues as the numbers of variables and constraints grow with the number of scenarios (Gendreau et al., 2016). To tackle large-scale instances more efficiently, we propose an alternative modeling framework that relies on a parametric approximation of the underlying probabilistic distributions of arrival times in the following subsection.

3.2.2 EVT-based model

In VRPs under stochastic travel and service times with hard time windows, service is not allowed to start before the beginning of each patient's time window. Unlike the case of soft time windows (e.g., see Hoogeboom et al. 2021), since the probability distributions of the arrival times at patients are truncated with the presence of waiting times, it is not directly possible to derive formulae for the convolution of the arrival time distributions (Gendreau et al., 2014). Indeed, the distribution of $\boldsymbol{a}_p$ (represents a vector including $a_{p,d}$ $\forall d \in D$) depends not only on travel times of the arcs traversed and service times of patients visited before arrival at the patient's location p but also on waiting times at prior patients. As a result, to find the mean and variance of $\boldsymbol{a}_p$, applying the convolution of distributions such as normal or gamma to sum the means and variances of the service times of the nodes and travel times of the arcs traversed before p is not directly applicable. By leveraging extreme value theory, Ehmke et al. (2015) proposed a method to approximate each customer's arrival time and start-service time distributions when waiting times must be considered under the hard time window constraints. We adapt this approximation for our modeling framework and integrate it into our stochastic optimization framework, referring to it as the extreme value theory-based model. The derivation by Ehmke et al. (2015) is based on the assumption that travel times are stochastic and follow normal distributions independently. However, the application of their approach may be limited in real situations in which significant dependencies may occur (Gendreau et al., 2016). In Section 3.2.2, we propose an extension of the EVT-based model to cases with dependent travel times to address this issue. Despite assuming normal distributions for travel times, Ehmke et al. (2015) demonstrated that the method also applies to instances with non-normal travel time distributions. We evaluate the impact of non-normal travel times in Section 4. In addition to travel times, our model also assumes that service times are random and follow a normal distribution.

In our problem, each patient is assigned a consistent visit time ν_p when scheduled with a care worker. If the care worker arrives at patient p before ν_p, they must wait until ν_p to begin service. Otherwise, if they arrive after ν_p, service starts immediately. Recall that $\boldsymbol{a_p}$ and $\boldsymbol{\tau_p}$ represent the arrival time and the start-service time for patient p, respectively.

To ensure that the probability of on-time arrival time at each patient p by $\nu_p + \mathcal{L}$ meets or exceeds the service level α, we need to derive the probabilistic distribution of $\boldsymbol{a_p}$ for each patient. This distribution depends on the likelihood of a care worker waiting at prior patients. Consequently, the arrival time at patient p is the sum of random travel times for arcs traversed before reaching p, random waiting times, and random service times for all prior patients. Figure 2 illustrates different arrival time distribution scenarios when the visit time window limits the start of service. Case (a) shows a scenario in which the arrival time is highly likely to fall within the patient's visit time window, resulting in a low probability of waiting. Case (b) depicts a situation in which the care worker arrives early and is likely to wait at the patient's location. Finally, case (c) represents a scenario when part of the arrival time distribution falls before the starting time window and the remaining part is within the window. Correspondingly, the arrival time distribution for the next patient, $\boldsymbol{a_{p+1}}$, can be calculated as the sum of the random parameters $\boldsymbol{\tau_p}$, s_p, and $t_{p,p+1}$.

We first describe how the EVT functions as presented in Ehmke et al. (2015) can be incorporated into our framework using the following example. To serve the first patient, the care worker requires

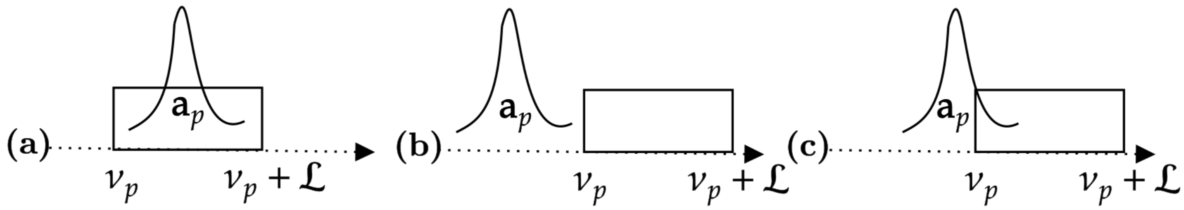

Figure 2: Arrival time distribution possibilities in presence of visit time window (adapted from Ehmke et al. (2015))
Note:(a) Arriving within the time window, (b) Arriving early, and (c) Arriving early or within the time window

$t_{0,1}$ time to travel from their address to the first patient. The start-service time for patient 1 is $\tau_1 = \max\{t_{0,1}, \nu_1\}$. This max operator complicates the mean and variance calculations of patients' start of service time. However, we can find the mean and variance of the maximum of two normally distributed variables using extreme value theory (see Nadarajah and Kotz 2008 for detailed derivations). Given the assumption that travel time is normally distributed, we model $t_{0,1} \sim N(\mu_{t_{0,1}}, \sigma^2_{t_{0,1}})$. The consistent appointment time ν_1 is an integer decision variable assigned to patient 1. Since it takes a single deterministic value, it can be represented as a normal distribution with zero variance, that is, $\nu_1 \sim N(\mu_{\nu_1}, 0)$ where $\mu_{\nu_1} = \nu_1$. The mean and variance of the start-service time for patient 1 can be expressed as:

$$\mu_{\tau_1} = \mu_{t_{0,1}} \Phi[(\mu_{t_{0,1}} - \nu_1)/\sigma_{t_{0,1}}] + \nu_1 \Phi[(\nu_1 - \mu_{t_{0,1}})/\sigma_{t_{0,1}}] + \sigma_{t_{0,1}} \phi[(\mu_{t_{0,1}} - \nu_1)/\sigma_{t_{0,1}}],$$
$$\sigma^2_{\tau_1} = (\mu^2_{t_{0,1}} + \sigma^2_{t_{0,1}}) \Phi[(\mu_{t_{0,1}} - \nu_1)/\sigma_{t_{0,1}}] + \nu_1^2 \Phi[(\nu_1 - \mu_{t_{0,1}})/\sigma_{t_{0,1}}] + (\mu_{t_{0,1}} + \nu_1)\sigma_{t_{0,1}} \phi[(\mu_{t_{0,1}} - \nu_1)/\sigma_{t_{0,1}}] - \mu^2_{\tau_1}$$

where $\Phi[...]$ and $\phi[...]$ denote the cumulative distribution function and the probability density function of the standard normal distribution, respectively. To calculate the arrival time distribution at the subsequent patient, patient 2, we first determine the distribution of the start-service time for patient 1, $\boldsymbol{\tau_1} \sim N(\mu_{\tau_1}, \sigma^2_{\tau_1})$, the service time for patient 1, $s_1 \sim N(\mu_{s_1}, \sigma^2_{s_1})$, and the travel time from patient 1 to patient 2, $t_{1,2} \sim N(\mu_{t_{1,2}}, \sigma^2_{t_{1,2}})$. Although $\boldsymbol{\tau_1}$ is initially not a normally distributed variable because it takes the maximum of two random variables, it is treated as a normal distribution to derive an approximation for arrival time distribution at patient 2; $\boldsymbol{a}_2 = \boldsymbol{\tau}_1 + s_1 + t_{1,2}$ (Clark, 1961; Ehmke et al., 2015).

More specifically, for any patient p, the mean and variance of the arrival time distribution are given by: $\mu_{\boldsymbol{a}_p} = \mu_{\boldsymbol{\tau}_{p-1}} + \mu_{s_{p-1}} + \mu_{t_{p-1,p}}$ and $\sigma^2_{\boldsymbol{a}_p} = \sigma^2_{\boldsymbol{\tau}_{p-1}} + \sigma^2_{s_{p-1}} + \sigma^2_{t_{p-1,p}}$. The approximate mean and variance of the start-service time for patient p can be determined using the following equations:

$$\mu_{\tau_p} = \mu_1 \Phi[(\mu_1 - \mu_2)/\sigma_1] + \mu_2 \Phi[(\mu_2 - \mu_1)/\sigma_1] + \sigma_1 \phi[(\mu_1 - \mu_2)/\sigma_1] \tag{21}$$
$$\sigma^2_{\tau_p} = (\mu_1^2 + \sigma_1^2) \Phi[(\mu_1 - \mu_2)/\sigma_1] + \mu_2^2 \Phi[(\mu_2 - \mu_1)/\sigma_1] + (\mu_1 + \mu_2)\sigma_1 \phi[(\mu_1 - \mu_2)/\sigma_1] - \mu^2_{\tau_p} \tag{22}$$

where $\mu_1 = \mu_{\boldsymbol{a_p}}$, $\sigma_1 = \sigma_{\boldsymbol{a_p}}$, and $\mu_2 = \nu_p$. An important distinction between our study and the work of Ehmke et al. (2015) is that, in their research, ν_p denotes a constant early time window, whereas in our problem, ν_p is represented as an integer variable. Consequently, Equations (21) and (22) introduce nonlinearity into the problem. Despite this, CP is well-suited for handling complex problems with nonlinear constraints, logical statements, or non-convex solution space (Wang et al., 2015). To effectively manage the complexity of $\Phi[...]$ and $\phi[...]$ functions, we can leverage callback functions and integrate them into CP solvers. For each route and schedule, we determine the α-percentile of the arrival time at each patient using the approximate mean and variance of the arrival time, as discussed earlier. Consequently, to enforce the chance constraint (19), we apply the following constraint: $\mu_{a_{p,d}} + \Phi^{-1}(\alpha)\sigma_{a_{p,d}} \leq \nu_p + \mathcal{L} \quad \forall p \in \tilde{P}_{c,d}, \forall d \in D$, where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard normal distribution.

EVT-based model with dependent travel times

Park and Rilett (1999) highlighted the significance of correlations between travel times, which can be influenced by factors such as weather and traffic conditions (Jaillet et al., 2016). Some studies used the covariance matrix of travel times to account for correlations (Bomboi et al., 2022; Bakach et al., 2021), but this approach is computationally prohibitive. One effective approach is to represent uncertain travel times as a weighted sum of independent factors, some of which would be common for all parameters (Gendreau et al., 2016). Following this approach, we assume that travel time between any node i and j can be represented by an affine regression function of Q independent variables, denoted by $\tilde{f}_1, ..., \tilde{f}_Q$, which may represent factors such as weather conditions, traffic incidents, and other relevant external influences (Caceres et al., 2016; Jaillet et al., 2016; Hoogeboom et al., 2021). Specifically, the travel time is modeled as:

$$\tilde{t}_{i,j} = t^0_{i,j} + \sum_{q=1}^{Q} \tilde{b}^q_{i,j} \tilde{f}_q$$

where $t^0_{i,j}$ is set to be the same as travel time when there is no correlation. The factor coefficients $\tilde{b}^q_{i,j}$ are predefined input parameters that can be obtained either from a regression model or expert knowledge. Thanks to its relatively simple structure, linear regression has been applied by several researchers to predict travel times using relevant covariates (Oh et al., 2015) (see, e.g., Lee et al. 2015; Kwon et al. 2000). Note that linear regression aligns well with the structure of our EVT-based model, where the mean and variance of the arrival time at each patient are approximated using the convolution of normal distributions. In particular, since linear regression yields a weighted sum of independent inputs, the resulting travel time remains normally distributed if the inputs are normal. This property preserves the additivity of means and variances, which is essential for efficiently computing arrival time distributions.

Instead of sampling travel time values from travel time distributions in our scenario-based model, we need to calculate them using the aforementioned affine functions. To do so, we incorporate independent factors as random values using their distribution functions. The EVT-based model requires mean and variance of travel times distributions. With the assumption that we have access to the distribution of independent factors in the affine function, we can find the mean and variance of travel times as follows: $\mu_{\tilde{t}_{i,j}} = \mu_{t^0_{i,j}} + \sum_{q=1}^{Q} \tilde{b}^q_{i,j} \mu_{\tilde{f}_q}$ and $\sigma^2_{\tilde{t}_{i,j}} = \sigma^2_{t^0_{i,j}} + \sum_{q=1}^{Q} (\tilde{b}^q_{i,j})^2 \sigma^2_{\tilde{f}_q}$.

3.3 Branch-and-check algorithm

We describe the B&Ch procedure, which is applied to solve both the scenario-based and EVT-based models. This approach is a variant of LBBD (Hooker and Ottosson, 2003), where Benders cuts are added in a branch-and-cut fashion throughout the branch-and-bound (B&B) procedure of the master problem. It has been shown that B&Ch can be more efficient than standard LBBD when the master problem is more complex to solve than the subproblem (Hooker, 2011; Heching et al., 2019). In the B&Ch approach, when an integer solution candidate for the MP is found at a B&B node, the SP is called to validate the feasibility of this candidate. If the subproblem finds the solution infeasible, no-good cuts (9) are added as a global constraint to the MP; otherwise, the B&B process continues. Infeasibility indicates that the subproblem could not find a schedule that satisfies all of its constraints, such as patient time windows, care worker shift limits, and service level requirements. While violation of the on-time arrival probability constraint is often a contributing factor, infeasibility may result from any subset of constraints. These cuts eliminate the current infeasible solution and prevent the master problem from assigning the same set of patients to care workers in subsequent B&B iterations. This procedure repeats at all nodes of the B&B tree when a feasible solution candidate for the MP is found until termination.

3.4 Time window inequalities

Inequalities derived from a relaxation of the subproblem could be added to the master problem from the outset, alongside constraints (2)–(8) and (10), to enhance computational performance. Heching et al. (2019) demonstrates that time window relaxation can significantly reduce computational time. These additional inequalities prevent the master problem from the over-assignment of patients to care workers, that is, assigning more patients than a care worker can feasibly serve. The underlying idea is that the total nominal service and travel times assigned to each care worker must fit within the time interval available for visits, as determined by the patients' availabilities and the care worker's shift duration (Heching and Hooker, 2016). This type of relaxation is based on forward and backward intervals. A forward interval for patient p begins with the start of their time window and ends with the termination of the care worker's time window. In addition, for patient p, a backward interval begins with the start of the care worker's time window and ends with the termination of the patient's time window. For each patient p assigned to the care worker c on day d, forward and backward augmented durations (p', p'') are calculated using following equations: $p'_{c,p,d} = \bar{s}_p + \min\{\min_{q \in P_{c,d}}\{\bar{t}_{p,q}\},\ \bar{t}_{p,l'_c}\}$ and $p''_{c,p,d} = \bar{s}_p + \min\{\bar{t}_{l_c,p},\ \min_{q \in P_{c,d}}\{\bar{t}_{q,p}\}\}$, where $P_{c,d}$ are the patients pre-assigned to care worker c on day d, as well as any new patient.

For patient $p \in P_{c,d}$, the sum of forward augmented durations of patients in $P[r_p, e_c]$, which is the set of patients whose time windows are in the forward interval of p, must observe the forward interval's width of p. It is the same for the patient's p backward interval. The following inequalities can be added to the MP:

$$\sum_{p \in P[r_p, e_c]} p'_{c,p,d} y_{c,p,d} \le e_c - r_p, \qquad \forall p \in P_{c,d}, \tag{23}$$

$$\sum_{p \in P[r_c, e_p]} p''_{c,p,d} y_{c,p,d} \le e_p - r_c, \qquad \forall p \in P_{c,d}. \tag{24}$$

4 Experimental results

This section presents our computational results obtained by three models, that is, deterministic, scenario-based, and EVT-based. The experiments were performed on benchmark instances in Heching et al. (2019) that are generated from real data provided by a home hospice care agency in the US. The code instances used in this paper and the detailed results are available on https://github.com/SHosseini94/Stochastic-Con-HHCRSP.

4.1 Experimental setting

The dataset comprises 60 patients and 20 care workers, with a planning horizon of 5 working days. In line with Heching et al. (2019), we initially designated only 8 out of 20 care workers to attend to new patients in addition to their current assignments (this is considered as Case A). In this case, the rolling schedule is employed, meaning that the schedule is reoptimized every week. During each rolling period, the master problem involves fixed variables (δ, x, and y) for existing patients, as specified by constraints 5–7. We generated several instances by considering the last k patients as new patients, where k ranges from 8 to 18 out of the total 60 patients. For existing patients, to ensure service consistency, the assigned care worker and visit days must remain the same, but visit times can be rescheduled. For each care worker, if there is an addition to the list of patients assigned to them, the schedule to visit patients is reoptimized. Otherwise, we keep their schedules unchanged. Thus, subproblems are solved only for care workers assigned new patients. In addition to Case A, we tested the models in more complex settings. First, we increase the number of new patients in the range of 20 to 40 (out of the 60-patient population), and make all 20 care workers available to serve them (instead of a subset of care workers as in the original case A). This is called Case B. Like Case A, Case B also employs a rolling schedule. Additionally, we consider Case C when the schedules must be generated from scratch (rather than on a rolling basis like in the original case). More specifically, we generated instances by considering the first k patients as new and the rest as discharged, where k ranges from 40 to 60 in this case. Indeed, both Cases B and C are much more difficult to solve than original one (Case A) in Heching et al. (2019). Table 6 presents a comparative summary of the three case configurations.

Table 6: Instance configurations for numerical experiments

Instance	#Care Worker	#New Patients	#Total Patients	#Visits	Rolling
Case A	8	[8-18]	25	[112,123]	Yes
Case B	20	[20-40]	60	269	Yes
Case C	20	[40-60]	[40-60]	[181-269]	No

In our instances, some patients require multiple visits per week. When they require two visits per week, there is a minimum gap of two days between consecutive visits, and when there are three weekly visits, there is a minimum gap of one day between consecutive visits. These requirements are incorporated as part of the feasibility constraints on the visit days (8), which can be expressed as follows (Heching et al., 2019):

$$y_{c,p,d} + y_{c,p,d+i} \leq 1 \qquad \forall c \in C, \forall p \in P \text{ with } w_p \in \{2,3\}, 1 \leq i \leq 4 - w_p, 1 \leq d \leq 5.$$

We implemented the models in Python 3.8.10 and performed the experiments on Compute Canada servers with 8 GB RAM and one core allocated for each instance. We used Gurobi 9.5.0 (Gurobi Optimization, LLC, 2022) to solve the master problem and the constraint programming solver of Cplex 22.1.0, CP Optimizer (IBM, 2022), to solve the subproblems. The maximum computation time was set to 7200 seconds per instance. Additionally, constraints (21) and (22) were implemented through a custom constraint handler (i.e., black-box expressions in CP Optimizer (IBM, 2022)) to evaluate the mean and variance of the start-service time for patients.

For each model, once the master problem identified an optimal solution, the corresponding subproblems were solved sequentially to determine schedules and routes. In the simulation phase, these final plans, including patient assignments, scheduled appointment times, and routing decisions, were held fixed. For each scenario, given the realized travel and service times, the actual arrival times were then computed. We evaluated the performance of the models using simulations with $10,000$ scenarios for service times and travel times. Travel times follow normal, shifted-gamma, and shifted-exponential distributions, with skewness values of 0, 1, and 2, respectively, to study the effect of skewness on the performance of the EVT-based model. The mean of travel time between each of the two locations equals the nominal value in the instance set. The CoV of travel times ($\mathrm{CoV_t}$) was set to 10, 25, or 50 percent of the mean travel times to represent different travel conditions. Service times follow a normal distribution in which the mean equals the nominal service times whereas the coefficient of variation (CoV) of the service time (denoted by $\mathrm{CoV_s}$) was set to 10 percent. The allowable delay parameter ($\mathcal{L}$) was set to 10, 20, or 30 minutes. This implies that we consider a late arrival a delay only when a care worker arrives later than the appointment time, plus the allowable delay. Additionally, we considered service levels α of 95% and 98% ($\Phi^{-1}(\alpha) = 1.64, 2.05$) to investigate the impact of different service levels. Finally, to study the effects of increasing the number of scenarios in the scenario-based model, we used the following different numbers of scenarios: 10, 50, 100, 300, and 500.

For the case of dependent travel times, we considered weather as an independent factor that impacts all travel routes (Caceres et al., 2016). We assumed two weather conditions: inclement (rain or snow) and clear. We set the probability of encountering inclement weather and clear weather at 40% and 60%, respectively. For clear weather, the factor coefficient was set to zero; for inclement weather, it was set to 30% of the base travel time.

4.2 Performance of the stochastic optimization approach for HHCRSP

We evaluate the performance of the models using several criteria. First, we compare the models in terms of computation times. As shown in Table 7, the deterministic model adapted from Heching et al. (2019) solves instances faster than the EVT-based and scenario-based models, which is to be expected. Indeed, service level constraints in stochastic models add complexity to the problem. For the EVT-based model, computation times range from less than 1 minute to around 5 minutes, depending on the number of new patients. This performance is acceptable for many home care agencies (Heching et al., 2019). When comparing the stochastic models, the EVT-based and scenario-based models are comparable when using 10 scenarios. However, with 50 or more scenarios, the EVT-based model consistently outperforms the scenario-based model. Increasing the number of scenarios to 100 or more, prevents the scenario-based model from reaching optimality within the time limit for some instances. To complement these results, Table 8 reports additional computational metrics, including the number of nodes explored and cuts added, which provide further insight into the computational behavior of the different models.

Table 7: Computation times in seconds (averaged over five runs) required to reach optimality for models in case A

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t = 25\%}$, $\mathbf{CoV_s = 10\%}$, **distribution = normal**							
New Patients	**Deterministic**	**EVT**	**Scenario**				
			10	**50**	**100**	**300**	**500**
8	2.5	15.3	3.6	45.5	90.4	341.3	692.4
10	3.1	13.9	7.6	50.1	204.5	905.6	1476.1
12	5.1	37.9	8.6	128.9	359.6	1569.2	3924.1
14	48.1	90.9	93	353.4	1504.8	6543	⋆
16	56.7	101.4	146.4	897.6	2935.5	⋆	⋆
18	77.8	378.5	505.5	3322.6	⋆	⋆	⋆

⋆ Computational time exceeded two hours

Table 8: Detailed computational metrics for models in case A

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t = 25\%}$, $\mathbf{CoV_s = 10\%}$, **distribution = normal**															
New Patients	**Deterministic**			**EVT**			**Scenario (10)**			**Scenario (50)**			**Scenario (100)**		
	Time (s)	**# Nodes**	**# Cuts**	**Time (s)**	**# Nodes**	**# Cuts**	**Time (s)**	**# Nodes**	**# Cuts**	**Time (s)**	**# Nodes**	**# Cuts**	**Time (s)**	**# Nodes**	**# Cuts**
8	2.5	1	7	15.3	30	42	3.6	1	13	45.5	3	26	90.4	8	26
10	3.1	1	14	13.9	3	50	7.6	1	22	50.1	10	32	204.5	86	60
12	5.1	32	18	37.9	286	104	8.6	22	30	128.9	73	61	359.6	151	84
14	48.1	1017	261	90.9	1057	396	93	1393	360	353.4	558	199	1504	4304	431
16	56.7	619	304	101.4	1066	470	146.4	3083	579	897.6	1021	484	2935.5	2805	775
18	77.8	1035	431	378.5	15189	1466	505.5	46542	1772	3322.6	9003	1735	⋆	⋆	⋆

⋆ Computational time exceeded two hours

The key difference between the deterministic and stochastic models concerns service level constraints embedded in the latter. To investigate the impact of these constraints on actual service levels for patients, we analyze simulation results using two criteria: minimum service level (MinS) and average service level (AvgS). As shown in Table 9, the deterministic model consistently fails to meet the service level of 98%, with both MinS and AvgS falling significantly below this target across all instances. In contrast, the EVT-based model successfully achieves the given service level in all instances. The performance of the scenario-based model varies depending on the number of scenarios used. Incorporating as few as 10 scenarios yields significantly better results than the deterministic model, with an average increase of 46.3% in MinS and 20.5% in AvgS, though it still fails to reach the target. Increasing the number of scenarios to 100 further improves performance, but it remains imperfect.

Table 9: Service level comparisons between models based on simulations in case A

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t = 25\%}$, $\mathbf{CoV_s = 10\%}$, **distribution = normal**														
New Patients	**Deterministic**		**EVT**		**Scenario (10)**		**Scenario (50)**		**Scenario (100)**		**Scenario (300)**		**Scenario (500)**	
	MinS	**AvgS**	**MinS**	**AvgS**	**MinS**	**AvgS**	**MinS**	**AvgS**	**MinS**	**AvgS**	**MinS**	**AvgS**	**MinS**	**AvgS**
8	0.45	0.76	0.98	0.99	0.67	0.91	0.92	0.98	0.97	0.99	0.98	0.99	0.97	0.99
10	0.44	0.75	0.98	0.99	0.67	0.92	0.92	0.98	0.98	0.99	0.99	0.99	0.99	0.99
12	0.44	0.78	0.98	0.99	0.67	0.93	0.89	0.98	0.97	0.99	0.98	0.99	0.98	0.99
14	0.49	0.79	0.99	0.99	0.67	0.93	0.93	0.98	0.97	0.99	0.97	0.99	⋆	⋆
16	0.45	0.74	0.98	0.99	0.67	0.92	0.95	0.99	0.96	0.99	⋆	⋆	⋆	⋆
18	0.48	0.74	0.98	0.99	0.67	0.92	0.91	0.99	⋆	⋆	⋆	⋆	⋆	⋆

⋆ Computational time exceeded two hours

In addition to service level metrics, we must consider the number of newly accepted patients in the objective function. However, considering only the number of newly accepted patients without considering the number of those who get service on time, at least α percentage of times, fails to provide a comprehensive view of the performance. In this regard, in Table 10, in addition to the number of newly accepted patients (NA), we also present the number of newly accepted patients whose on-time service level is satisfied (ON).

Table 10: Comparison of models based on number of newly accepted patients and newly accepted patients with satisfied on-time service level in case A simulations

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t = 25\%}$, $\mathbf{CoV_s = 10\%}$, **distribution = normal**														
New Patients	**Deterministic**		**EVT**		**Scenario (10)**		**Scenario (50)**		**Scenario (100)**		**Scenario (300)**		**Scenario (500)**	
	NA	ON	NA	ON	NA	ON	NA	ON	NA	ON	NA	ON	NA	ON
8	8	1	7	**7**	8	2	8	7	7	7	7	7	7	7
10	9	2	8	**8**	9	2	9	7	8	8	8	8	8	8
12	11	4	10	10	11	5	11	5	11	10	11	11	11	11
14	13	3	12	**12**	13	5	13	11	12	11	12	11	⋆	⋆
16	16	3	14	**14**	15	7	15	14	15	13	⋆	⋆	⋆	⋆
18	18	2	16	**16**	17	6	17	14	⋆	⋆	⋆	⋆	⋆	⋆

⋆ Computational time exceeded two hours

According to Table 10, under moderate variations in travel times and a 10-minute allowable delay, the deterministic model generally accepts more new patients compared to the EVT-based model. However, most of these patients fail to receive services on time. The maximum difference in the number of newly accepted patients between the two models is two. Despite this, the EVT-based model consistently achieves the given service level for all newly accepted patients. For the scenario-based model, the results depend on the number of scenarios considered. When using 100 scenarios, its performance is comparable to the EVT-based model.

In our problem definition, each care worker is subject to a maximum weekly work time. This may raise concerns about workload inconsistency; specifically, the possibility that a care worker might work significantly more on some days than others. Workload consistency can be viewed as another dimension of consistency in vehicle routing problems (VRPs), alongside personnel and time consistency. However, achieving high personnel or time consistency does not necessarily guarantee consistent workloads (Wu et al., 2025). Although workload consistency is not the primary objective of this work, it can be readily incorporated into our subproblem by adding the following constraint:

$$\lambda_d - \lambda_{d'} \leq L \ \ \forall d, d' \in D; d \neq d'$$

where in our SP, λ_d denotes the workload of a caregiver on day d, and L is a specified upper bound on the allowable difference in workload across days.

Moreover, in our benchmark instances, over 80% of patients require five visits per week (Heching et al., 2019). Under such circumstances, high personnel and time consistency are likely to increase the chance of achieving workload consistency. Figure 3 supports this observation: the majority of care workers experience less than 20% variation in their daily workload. However, a few exhibit higher variability, around 50%. This analysis reflects a single week of planning, so it is possible that the company assigns additional patients to these care workers in subsequent weeks.

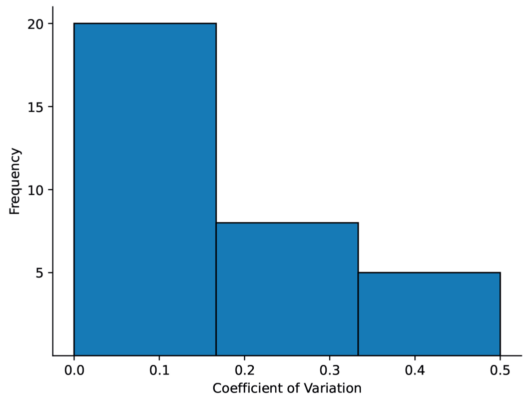

Figure 3: Daily workload variability across care workers in EVT-based model under case A

Case B: scheduling new patients with additional care workers We now consider the case when more care workers are available to be assigned to new patients. Table 11 compares the three models based on several criteria, including NA, ON, MinS, and AvgS. While the deterministic model quickly finds optimal solutions, its actual average service level is only 78%, falling well below the service level of 98%. In contrast, the EVT-based model, with the same number of newly accepted patients, consistently outperforms in service level metrics across all benchmark instances. It achieves the given service level in all cases while solving instances within 15 minutes. The scenario-based model, however, requires over 100 scenarios to approach comparable service levels, resulting in significantly higher computation times (see Figure 4). This highlights the computational cost associated with the scenario-based model relative to the EVT-based model.

Table 11: Comparison of models based on simulations in case B

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t = 25\%}$, $\mathbf{CoV_s = 10\%}$, **distribution = normal**																
New Patients	**Deterministic**				**EVT**				**Scenario (10)**				**Scenario (100)**			
	NA	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**
20	20	5	0.45	0.78	20	**20**	0.98	0.99	20	7	0.60	0.94	20	19	0.97	0.99
25	25	6	0.41	0.78	25	**25**	0.98	0.99	25	10	0.82	0.97	25	25	0.98	0.99
30	30	7	0.48	0.76	30	**30**	0.98	0.99	30	21	0.84	0.97	30	27	0.92	0.99
35	35	10	0.47	0.78	35	**35**	0.98	0.99	35	13	0.82	0.95	35	32	0.96	0.98
40	40	13	0.44	0.78	40	**40**	0.98	0.99	40	21	0.77	0.96	40	38	0.94	0.99

Figure 5 illustrates the trade-off between average travel time, idle time, and late time for each instance in Case B, based on the simulated scenarios. In terms of average lateness across all scenarios, both the EVT-based model and the scenario-based model with 100 scenarios demonstrate near-zero lateness, whereas the deterministic model exhibits poor reliability in patient arrival times. The low average lateness of the EVT-based model is achieved at the cost of higher idle time compared to the deterministic model; however, this is largely offset by its lower travel time. In addition, Figure 6 presents the worst-case lateness observed across all models. The substantial difference in maximum lateness between the stochastic models and the deterministic model may be particularly compelling for decision-makers, as it highlights the benefit of accepting some idle time in exchange for significantly improved reliability.

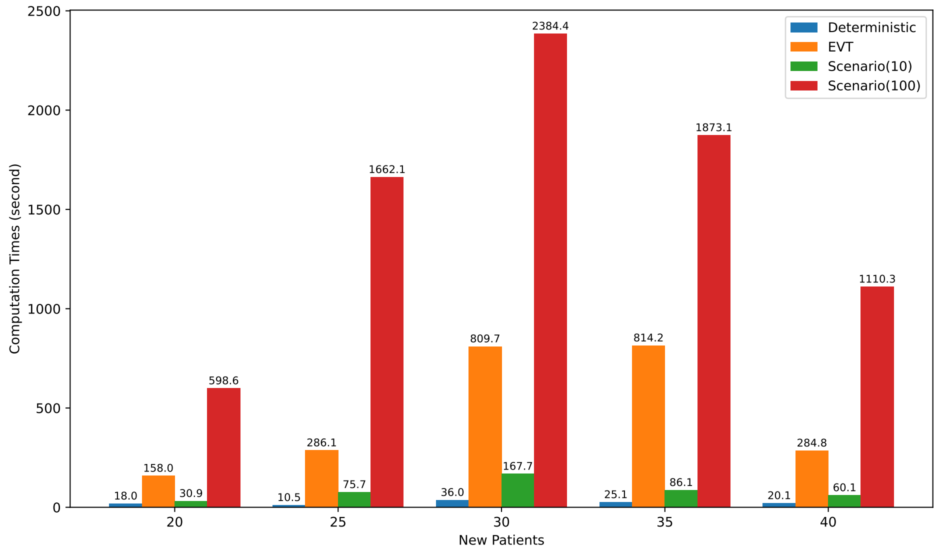

Figure 4: Average computation times in seconds over five runs for models in case B

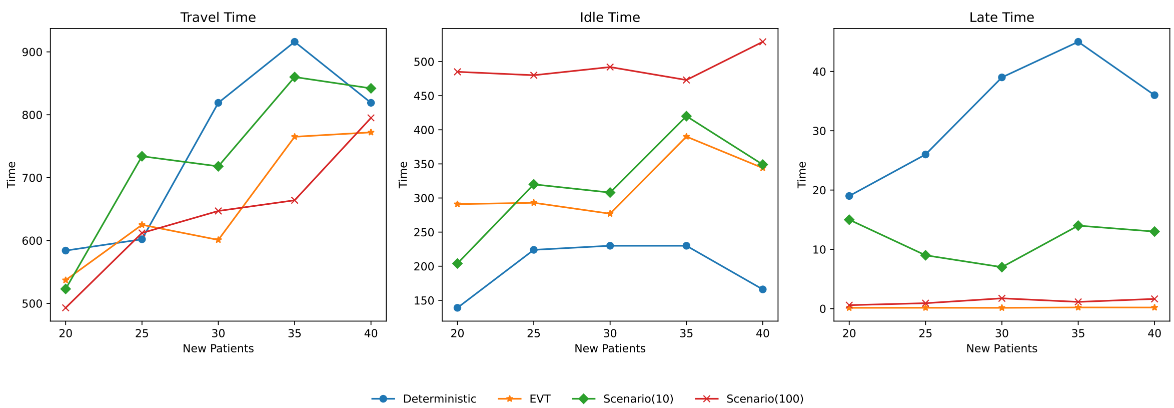

Figure 5: Average travel time, idle time, and late time under case B

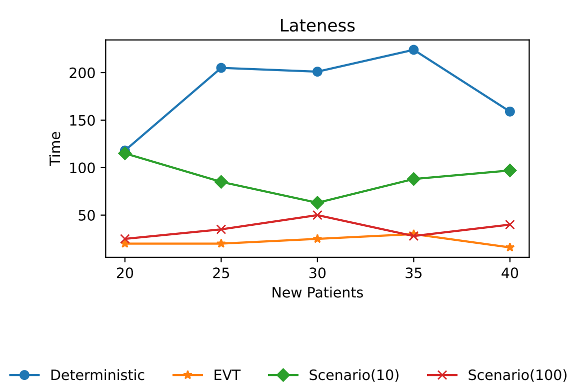

Figure 6: Maximum lateness under case B

Case C: scheduling from scratch In this case, all patients are considered new to study the performance of the models when instances are solved from scratch rather than on a rolling basis. A similar pattern to Table 11 is observed in Table 12. The EVT-based model demonstrates remarkable efficiency, solving a 40-patient instance in less than 1 minute—only slightly slower than the deterministic model. For larger instances, such as those with 60 patients, the EVT-based model performs well, solving the problem in about 10 minutes. In contrast, the scenario-based model with 100 scenarios fails to find an optimal solution for 60 patients within a two-hour time limit. Moreover, the EVT-based model schedules the same number of patients as the deterministic model but does so more efficiently, delivering higher actual service levels based on simulation results.

Table 12: Comparison of models based on simulations in case C

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t} = \mathbf{25\%}$, $\mathbf{CoV_s} = \mathbf{10\%}$, **distribution = normal**																				
Patients	**Deterministic**					**EVT**					**Scenario (10)**					**Scenario (100)**				
	$\mathbf{T^1}$	**NA**	**ON**	**MinS**	**AvgS**	$\mathbf{T^1}$	**NA**	**ON**	**MinS**	**AvgS**	$\mathbf{T^1}$	**NA**	**ON**	**MinS**	**AvgS**	$\mathbf{T^1}$	**NA**	**ON**	**MinS**	**AvgS**
40	27.7	40	21	0.57	0.89	38.1	40	**40**	0.98	0.99	19.9	40	32	0.83	0.97	174.6	40	38	0.96	0.99
50	37.5	50	20	0.52	0.84	87.4	50	**50**	0.98	0.99	21.1	50	28	0.80	0.96	1291	50	49	0.96	0.99
60	53.3	60	10	0.45	0.77	543.3	60	**60**	0.98	0.99	181.8	60	28	0.77	0.95	*	*	*	*	*

[1] Computation times in seconds (averaged over five runs)
⋆ Computational time exceeded two hours

Regarding the scalability of the proposed methods, it is important to note that applying Benders decomposition allows us to separate assignment decisions from sequencing decisions. As observed in practice, each care worker typically visits fewer than six patients per day on average (Yuan et al., 2015), which helps keep the number of variables in each subproblem limited. Consequently, in the largest instance solved in Case C, involving 60 patients, the problem does not correspond to a VRP with 60 nodes. Specifically, each subproblem includes at most six variables for appointment times, ν_p, and six variables for sequencing, $\pi_{z,d}$, per day (i.e., a total of $6 + 6 \times 5$ variables). In contrast, the master problem contains significantly more binary decision variables, that is, $|P||D||C| + |P||C| + |P|$. To identify the largest instances solvable within two hours over a five-day planning horizon, we examine the performance of the EVT-based model, which shows superior computational efficiency compared to the scenario-based model. The original hospice care dataset from Heching et al. (2019) consists of 20 care workers and 60 patients. We extended this dataset by proportionally increasing the numbers of care workers and patients. The computational results for these new instances are reported in Table 13. The computation times (averaged over five runs) increase with instance size, and the largest instance solved within the time limit by the EVT-based model consists of 72 patients and 24 care workers.

Table 13: Computational results for larger instances with EVT-based model

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t} = \mathbf{25\%}$, $\mathbf{CoV_s} = \mathbf{10\%}$, **distribution = normal**					
Patients	**Visits**	**Care Workers**	**Time (s)**	**# Nodes**	**# Cuts**
60	269	20	543.3	1561	1218
63	284	21	909.45	1962	2231
66	297	22	2166.9	6395	6198
69	312	23	2892.7	4487	7333
72	327	24	6043.1	17122	18909
75	342	25	*	*	*

⋆ Computational time exceeded two hours

4.3 Sensitivity analysis

We investigate the effects of parameter changes on the performance of the models. Sensitivity analyses were conducted to analyze changes in $\mathcal{L}$, CoV_v, types of probability distributions for travel times, and

α under Case A. Figures 7–8 demonstrate the impact of increasing the allowable delay in the models. Both the EVT-based model and scenario-based model with 100 scenarios leverage a larger $\mathcal{L}$ to accept more new patients. Specifically, when $\mathcal{L}$ is set to 30 minutes, these models match the deterministic model in the number of newly accepted patients across all instances. However, they significantly outperform the deterministic model in key metrics such as ON, MinS, and AvgS. Although a larger allowable delay improves the deterministic model's performance in terms of MinS and AvgS, the ON value often remains below 50% of NA.

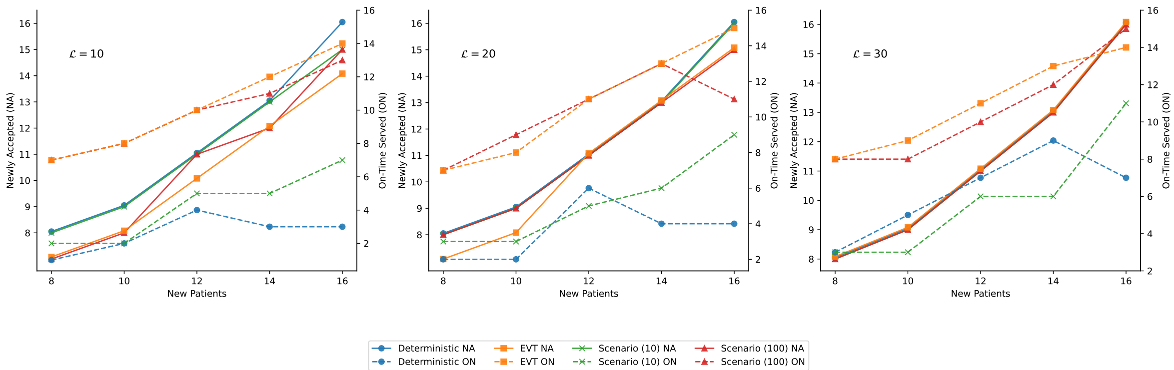

Figure 7: Sensitivity analysis of allowable delay parameter for NA and ON

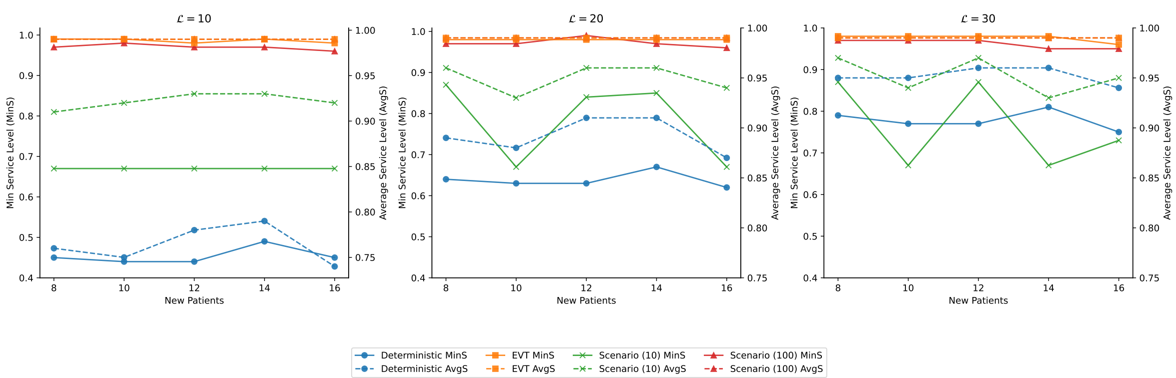

Figure 8: Sensitivity analysis of allowable delay parameter for MinS and AvgS

According to Figures 9–10, lower uncertainty in travel times generally leads to better service levels for patients. The EVT-based model consistently meets service level requirements for all newly accepted patients and can accept more new patients as $\mathrm{CoV_t}$ decreases. Both stochastic models adapt their solutions to changes in $\mathrm{CoV_t}$. In contrast, the deterministic model cannot adjust to varying levels of travel time uncertainty, resulting in decreased performance as travel time uncertainty increases.

In the EVT-based model, since estimated CoV values of travel times between nodes are input parameters, to investigate the effect of mispecifications of estimated CoV values based on actual data, we evalute the model's performance using CoV values that are smaller than, equal to, or larger than the actual CoV (see Table 14). When we use a smaller value for CoV in the model, we underestimate the actual variation in travel times. Therefore, more new patients may be accepted, but not all can be served at the given service level. In contrast, an overestimation happens when we consider a larger value for CoV in the model. As a result, we might accept fewer new patients than we should since the solution can be overly conservative. Thus, it is essential to properly estimate the CoV values used in the EVT-based model.

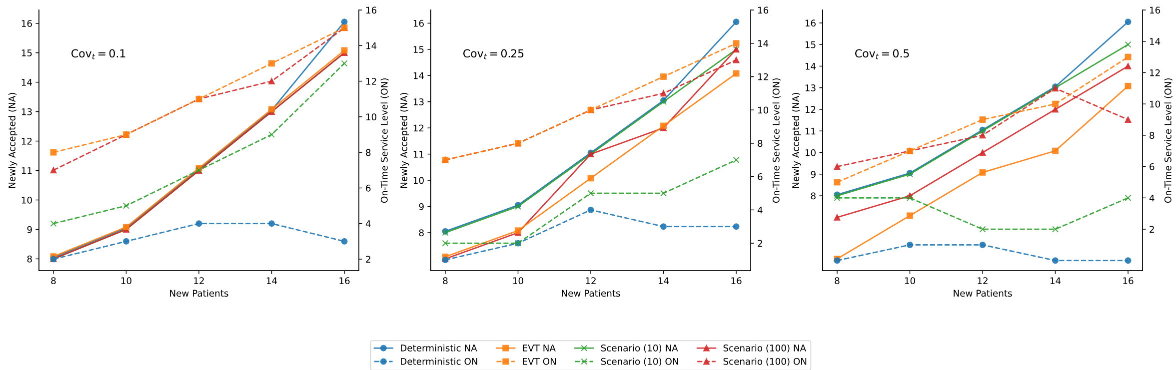

Figure 9: Effects of Coefficient of Variation of Travel Times on NA and ON

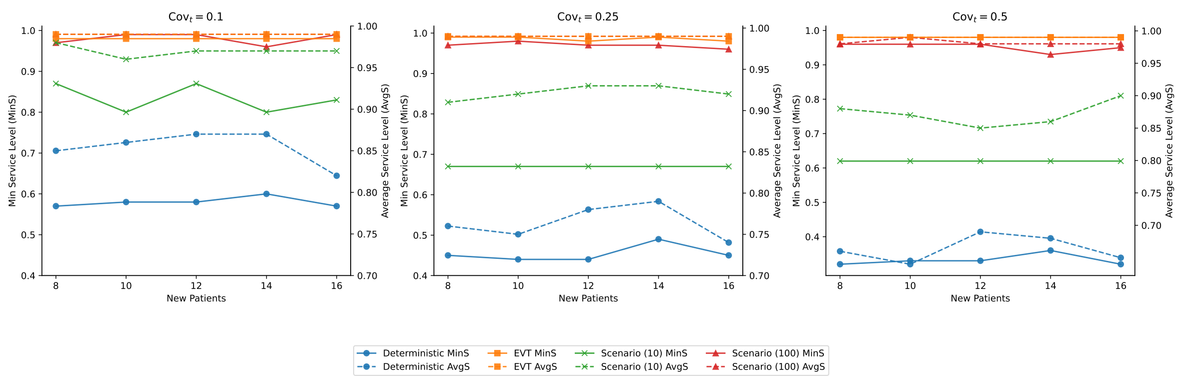

Figure 10: Effects of Coefficient of Variation of Travel Times on MinS and AvgS

Table 14: Proper Travel CoV Value for EVT-Based Model

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t = 25\%}$, $\mathbf{CoV_s = 10\%}$, **distribution = normal**												
New Patients	**EVT CoV (0.15)**				**EVT CoV (0.25)**				**EVT CoV (0.35)**			
	NA	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**
8	7	6	0.94	0.98	7	**7**	0.98	0.99	7	**7**	0.98	0.99
10	8	6	0.91	0.98	8	**8**	0.98	0.99	8	**8**	0.98	0.99
12	11	7	0.94	0.98	10	**10**	0.98	0.99	10	**10**	0.98	0.99
14	13	10	0.94	0.98	12	**12**	0.99	0.99	11	11	0.98	0.99
16	15	10	0.91	0.98	14	**14**	0.98	0.99	14	**14**	0.98	0.99
18	17	9	0.91	0.97	16	**16**	0.98	0.99	15	15	0.99	0.99

We also performed a sensitivity analysis to investigate whether a non-normal probability distribution of travel times affects the solution quality of the EVT-based model. The results, shown in Table 15, indicate that the model remains robust across different distributions. When travel times follow a shifted-gamma or shifted-exponential distribution, AvgS still meets or exceeds the service level of 98%. However, in the worst-case scenario, MinS reaches 97% for the shifted-gamma distribution and 96% for the shifted-exponential distribution. This slight decline in MinS is expected as the skewness of the travel time distribution increases. MinS tends to be more sensitive to deviations from normality, but the proposed methodology continues to provide accurate estimates for arrival time distributions.

Finally, the effect of service level α on the performance of the models is investigated in Figures 11–12. For both the deterministic model and the scenario-based model with 10 scenarios, α does not affect the MinS and AvgS metrics, though lowering α can increase the ON value in some cases. Interestingly, the EVT-based model is sensitive to changes in α; reducing the service level requirement from 98% to 95% sometimes allows for the acceptance of one additional patient. The MinS metric for both the EVT-based model and scenario-based model with 100 scenarios aligns with the given α, ensuring that the solution adjusts to meet the given service level.

Table 15: Performance of EVT-Based model with different distribution types

	$\alpha = 0.98$, $\mathcal{L} = 10$, $\mathrm{CoV_t} = 25\%$, $\mathrm{CoV_s} = 10\%$											
New Patients	**Normal**				**Shifted-gamma**				**Shifted-exponential**			
	NA	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**
8	7	**7**	0.98	0.99	7	**7**	0.97	0.99	7	5	0.97	0.98
10	8	**8**	0.98	0.99	8	7	0.97	0.99	8	5	0.97	0.98
12	10	**10**	0.98	0.99	10	**10**	0.97	0.99	10	8	0.96	0.98
14	12	**12**	0.99	0.99	12	**12**	0.97	0.99	12	9	0.97	0.98
16	14	**14**	0.98	0.99	14	**14**	0.97	0.99	14	12	0.96	0.98
18	16	**16**	0.98	0.99	16	13	0.97	0.98	16	10	0.96	0.98

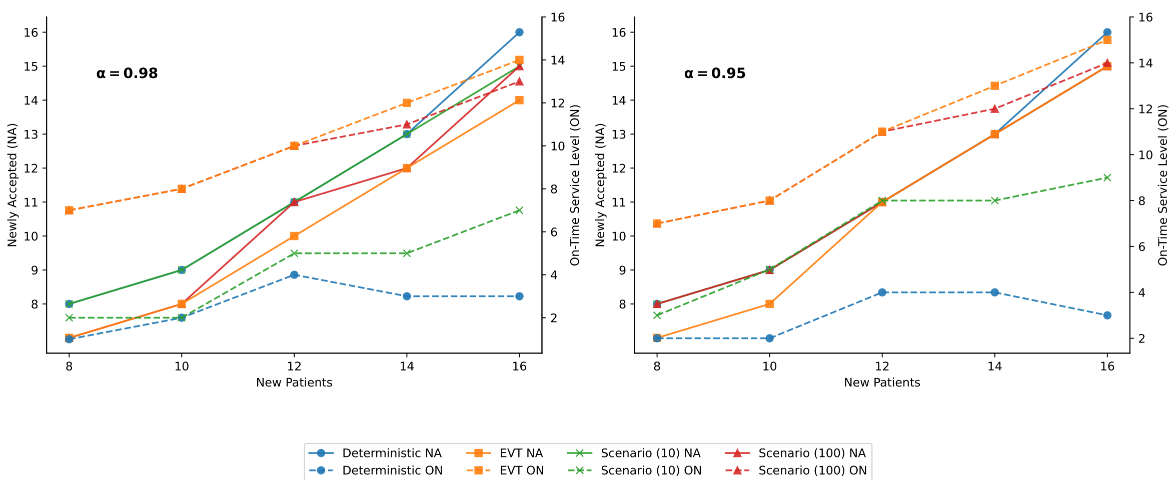

Figure 11: Effects of required service levels on NA and ON

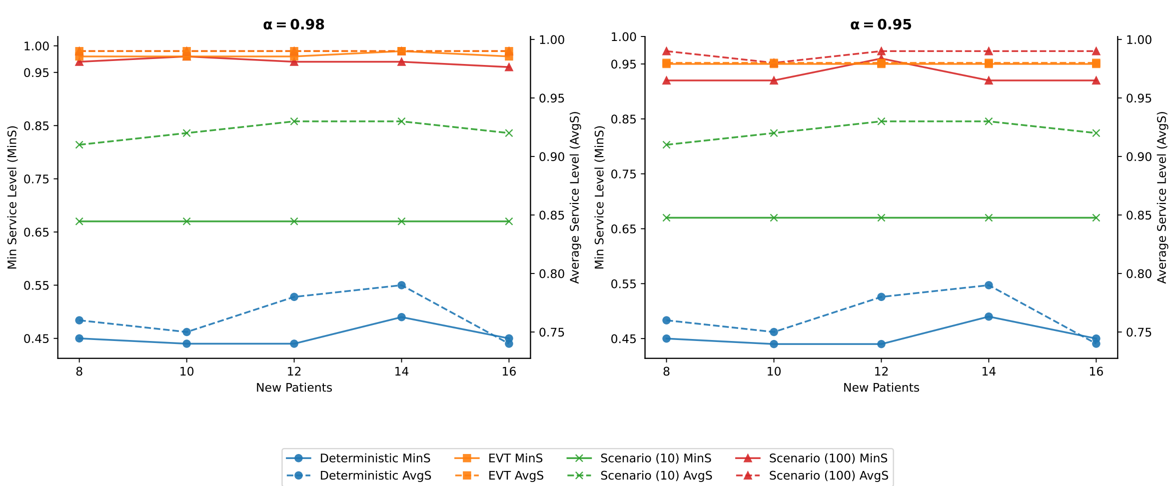

Figure 12: Effects of required service levels on MinS and AvgS

4.4 Performance for the case with dependent travel times

Table 16 demonstrates the performance of the models when travel times are dependent. The deterministic model shows a significant decrease in performance, with MinS dropping by at least 14% and AvgS by at least 11%. Additionally, in 4 out of 6 instances, it fails to meet the given service level for all newly accepted patients. However, in the EVT-based model, AvgS still meets the given service level across all instances, while MinS drops from 98% to 96% in some cases. With the introduction of the affine regression function, the scenario-based model better estimates the variance of arrival times at patients. Specifically, with 10 scenarios, we observe improvement in MinS ranging from 11% to 20% and up to 5% in AvgS. This improvement occasionally comes at the cost of accepting one fewer patient. Additionally, with 100 scenarios, the ON value matches or exceeds that of the EVT-based model in 5 out of 6 instances. In conclusion, when travel times are dependent, the integration of an affine regression function in the EVT-based and scenario-based models yields significantly better results in terms of service level than the deterministic model, without increasing algorithmic complexity.

Table 16: Performance of models under dependent travel times

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t} = \mathbf{25}\%$, $\mathbf{CoV_s} = \mathbf{10}\%$, **distribution = normal**																	
New Patients	**Dependent**	**Deterministic**				**EVT**				**Scenario (10)**				**Scenario (100)**			
		NA	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**	**NA**	**ON**	**MinS**	**AvgS**
8	no	8	1	0.45	0.76	7	**7**	0.98	0.99	8	2	0.67	0.91	7	**7**	0.97	0.99
8	yes	8	0	0.29	0.64	7	6	0.97	0.98	7	4	0.87	0.96	7	**7**	0.95	0.99
10	no	9	2	0.44	0.75	8	**8**	0.98	0.99	9	2	0.67	0.92	8	**8**	0.98	0.99
10	yes	9	1	0.30	0.62	8	**7**	0.97	0.98	8	4	0.87	0.96	8	**7**	0.95	0.98
12	no	11	4	0.44	0.78	10	**10**	0.98	0.99	11	5	0.67	0.93	11	**10**	0.97	0.99
12	yes	11	1	0.30	0.67	10	7	0.96	0.98	11	3	0.79	0.94	10	**10**	0.98	0.99
14	no	13	3	0.49	0.79	12	**12**	0.99	0.99	13	5	0.67	0.93	12	11	0.97	0.99
14	yes	13	0	0.31	0.66	11	8	0.96	0.98	13	4	0.87	0.95	12	**10**	0.97	0.99
16	no	16	3	0.45	0.74	14	**14**	0.98	0.99	15	7	0.67	0.92	15	13	0.96	0.99
16	yes	16	0	0.29	0.62	14	10	0.96	0.98	15	7	0.79	0.95	14	**12**	0.95	0.99
18	no	18	2	0.48	0.74	16	**16**	0.98	0.99	17	6	0.67	0.92	*	*	*	*
18	yes	18	0	0.32	0.62	15	**11**	0.96	0.98	16	5	0.80	0.96	16	10	0.94	0.98

5 Conclusions and future research

In this paper, we present a chance-constrained modeling framework that extends upon the deterministic home health care routing and scheduling problem of Heching et al. (2019). Our framework takes into account uncertainty in travel and service times, along with personnel and time consistency constraints, and can be adapted to handle dependent travel times. We proposed two formulation schemes: an EVT-based and a scenario-based model. The EVT-based model relies on a parametric approximation of the start-service time and arrival time distributions for patients. To solve the problem, we implemented an exact method called branch-and-check, which decomposes the problem into a MILP master problem and CP subproblems. Through simulations, we evaluated various metrics to compare the efficiency of our stochastic models with the deterministic model of Heching et al. (2019). Our experiments demonstrate that the deterministic model fails to achieve the given service level, whereas the stochastic models consistently meet this criterion. Moreover, the number of newly accepted patients in the stochastic models is generally comparable to that of the deterministic model. Additionally, the EVT-based model offers a significant computational advantage by replacing multiple scenarios with a nonlinear EVT-based function, which leads to faster solution times compared to the scenario-based model. This paper can be extended in different directions. For example, given that we have studied the problem in a static setting, its dynamic variants could be promising in future studies. Additionally, although hard service constraints are more difficult to solve exactly, it would be interesting to develop new models and methods for a soft constraint version of service consistency, where personnel and time consistency constraints can be relaxed and penalized in the objective function.

A Mixed-Integer Linear formulation of scenario-based model

This appendix presents the MILP formulation of the scenario-based model. Table A1 shows decision variables used in this model.

Table A1: Description of notation used in MILP scenario-based model

Decision Variable:	
$u_{i,i',d}$	1 if care worker travels from location i to location i' on day d; and 0 otherwise
$\gamma^n_{p,d}$	1 if care worker arrives on time to patient p on day d in sample n; and 0 otherwise
$q^n_{p,d}$	1 if care worker arrives sooner to patient p on day d in sample n; and 0 otherwise
ν_p	Integer variable for appointment time of patient p across all visit
$a^n_{p,d}$	Continuous variable for arrival time at patient p on day d in sample n
$\tau^n_{p,d}$	Continuous variable for start-service time for patient p on day d in sample n
b^n_d	Continuous variable for start-work care worker on day d
f^n_d	Continuous variable for end-work care work on day d
λ^n_d	Continuous variable for work time of care worker on day d in sample n
Set:	
D	Set of days
N	Set of scenarios
$\tilde{P}_{c,d}$	Set of patients assigned to care worker c on day d, determined by master problem
I	Set of all nodes, including care workers and patients
O_d	Set of arcs on day d
Parameter:	
l_p	Location of patient p
$[r_p, e_p]$	Time window of patient p
l_c	Starting location of care worker c
l'_c	Ending location of care worker c
$[r_c, e_c]$	Time window of care worker c
h_c	Maximum work time over the week of care worker c
s^n_p	Service time patient p in sample n
$t^n_{i,i'}$	Travel time between node i and i' in sample n
$\|N\|$	Number of scenarios
$\mathcal{L}$	Maximum allowable delay
α	Minimum acceptable service level

$$\sum_{(i',i)\in O_d} u_{i',i,d} - \sum_{(i,i')\in O_d} u_{i,i',d} = 0 \qquad \forall i \in \tilde{P}_{c,d}, \forall d \in D, \tag{A1}$$

$$\sum_{(i',i)} u_{i',i,d} = 1 \qquad \forall i \in \tilde{P}_{c,d}, \forall d \in D, \tag{A2}$$

$$\sum_{i\in\tilde{P}_{c,d}} u_{l_c,i,d} = 1 \qquad \forall d \in D, \tag{A3}$$

$$\sum_{i\in\tilde{P}_{c,d}} u_{i,l'_c,d} = 1 \qquad \forall d \in D, \tag{A4}$$

$$\tau^n_{p,d} + s^n_p + t^n_{l_p,l'_p} - M(1 - u_{p,p',d}) \le a^n_{p',d} \qquad \forall p, p' \in \tilde{P}_{c,d} : p! = p', \forall d \in D, \forall n \in N \tag{A5}$$

$$\tau^n_{p,d} + s^n_p + t^n_{l_p,l'_p} + M(1 - u_{p,p',d}) \ge a^n_{p',d} \qquad \forall p, p' \in \tilde{P}_{c,d} : p! = p', \forall d \in D, \forall n \in N \tag{A6}$$

$$a^n_{p,d} \le \nu_p + \mathcal{L} + M(1 - \gamma^n_{p,d}) \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N \tag{A7}$$

$$\frac{1}{|N|}\sum_{n\in N} \gamma^n_{p,d} \ge \alpha \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \tag{A8}$$

$$\tau^n_{p,d} \geq a^n_{p,d} \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A9}$$
$$\tau^n_{p,d} \leq a^n_{p,d} + M(q^n_{p,d}) \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A10}$$
$$\tau^n_{p,d} \geq \nu_p \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A11}$$
$$a^n_{p,d} \leq \nu_p + M(1 - q^n_{p,d}) \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A12}$$
$$a^n_{p,d} \geq r_a + t^n_{l_a,l_p} \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A13}$$
$$\tau^n_{p,d} \geq b^n_d \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A14}$$
$$\tau^n_{p,d} \leq b^n_d + M(1 - u_{l_c,p,d}) \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A15}$$
$$\tau^n_{p,d} + s^n_p \leq f^n_d \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A16}$$
$$\tau^n_{p,d} + s^n_p + M(1 - u_{p,l'_c,d}) \geq f^n_d \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N, \tag{A17}$$
$$\lambda^n_d = f^n_d - b^n_d \qquad \forall d \in D, \forall n \in N, \tag{A18}$$
$$\sum_{d \in D} \lambda^n_d \leq h_c \qquad \forall n \in N \tag{A19}$$
$$\nu_p \in \{k | k \in \mathbb{Z}, k \in [r_p, e_p - \bar{s}_p)\},\ a^n_{p,d} \geq r_c, \tau^n_{p,d} \geq r_c, \tag{A20}$$
$$b^n_d \in [r_c, e_c), f^n_d \in [r_c, e_c) \qquad \forall p \in \tilde{P}_{c,d}, \forall d \in D, \forall n \in N.$$

Constraints (A1)–(A4) represent flow conservation between the assigned patient nodes and the care worker's start and end locations. Constraints (A5)–(A6) ensure that when a care worker visits two patients consecutively, the arrival time at the latter equals the sum of the start-service time, service duration, and travel time. Otherwise, a big-M term is used to relax the constraint when the corresponding arc is not traversed. These constraints, often referred to as commodity flow constraints, work in conjunction with flow constraints to eliminate subtours. Service level constraints are defined in constraints (A7)–(A8). Constraints (A9)–(A12) guarantee that the care worker waits at the patient's location until the scheduled appointment time and begins service immediately upon on time or delayed arrival. Constraints (A13) limits the earliest possible arrival at a patient's location by the start of the care worker's time window plus the travel time from their departure point. Constraints (A14)–(A15) enforce that the care worker starts their shift at the first visited patient, while constraints (A16)–(A17) ensure that their working day ends upon departure from the last patient. Constraints (A18)–(A19) limit the maximum work time of the care worker. Finally, constraints (A20) define the domains of the variables.

Table A2 compares the scenario-based models in terms of computation time. The only difference between the two approaches lies in how the subproblem is modeled and solved, that is, the CP model uses constraint programming, while the MILP model uses mixed-integer programming. As both formulations solve the same problem, they yield identical optimal solutions but differ in computational efficiency. On average, the CP-based model reduces computation time by 29.4% and outperforms the MILP model in 9 out of 10 solved instances.

Table A2: Computation times in seconds (averaged over five runs) required to reach optimality for models in case A

$\alpha = \mathbf{0.98}$, $\mathcal{L} = \mathbf{10}$, $\mathbf{CoV_t = 25\%}$, $\mathbf{CoV_s = 10\%}$, **distribution = normal**						
New Patients	**10-Scenarios**			**100-Scenarios**		
	CP	**MILP**	**% Improvement**	**CP**	**MILP**	**%Improvement**
8	3.6	6.1	-41%	90.4	278.3	-67.5%
10	7.6	9.2	-17.4%	204.5	155.5	31.5%
12	8.6	13.6	-36.8%	359.6	427.3	-15.8%
14	93.0	135.0	-31.3%	1504.8	3241.1	-53.5%
16	146.4	219.5	-33.3%	2935.5	⋆	

⋆ Computational time exceeded two hours

References

AlayaCare, 2023. 3 ways to improve care management in home care. Available at https://alayacare.com/blog/3-ways-to-improve-care-management-in-home-care/. (Accessed 20 March 2025).

Alvarez, A., Cordeau, J.F., Jans, R., 2024. The consistent vehicle routing problem with stochastic customers and demands. Transportation Research Part B: Methodological 186, 102968. doi:https://doi.org/10.1016/j.trb.2024.102968.

Bakach, I., Campbell, A.M., Ehmke, J.F., Urban, T.L., 2021. Solving vehicle routing problems with stochastic and correlated travel times and makespan objectives. EURO Journal on Transportation and Logistics 10, 100029. doi:https://doi.org/10.1016/j.ejtl.2021.100029.

Baptiste, P., Le Pape, C., Nuijten, W., 2001. Constraint-based scheduling: applying constraint programming to scheduling problems. volume 39. Springer Science & Business Media.

Bennett, A.R., Erera, A.L., 2011. Dynamic periodic fixed appointment scheduling for home health. IIE Transactions on Healthcare Systems Engineering 1, 6–19. doi:https://doi.org/10.1080/19488300.2010.549818.

Bertels, S., Fahle, T., 2006. A hybrid setup for a hybrid scenario: combining heuristics for the home health care problem. Computers & Operations Research 33, 2866–2890. doi:https://doi.org/10.1016/j.cor.2005.01.015. part Special Issue: Constraint Programming.

Bomboi, F., Buchheim, C., Pruente, J., 2022. On the stochastic vehicle routing problem with time windows, correlated travel times, and time dependency. 4OR 20, 217–239. doi:https://doi.org/10.1007/s10288-021-00476-z.

Borsani, V., Matta, A., Beschi, G., Sommaruga, F., 2006. A home care scheduling model for human resources, in: 2006 International Conference on Service Systems and Service Management, pp. 449–454.

Caceres, H., Hwang, H., He, Q., 2016. Estimating freeway route travel time distributions with consideration to time-of-day, inclement weather, and traffic incidents. Journal of Advanced Transportation 50, 967–987. doi:https://doi.org/10.1002/atr.1384.

Cappanera, P., Scutellà, M.G., 2022. Addressing consistency and demand uncertainty in the home care planning problem. Flexible Services and Manufacturing Journal 34, 1–39. doi:https://doi.org/10.1007/s10696-021-09412-z.

Cissé, M., Yalçındağ, S., Kergosien, Y., Şahin, E., Lenté, C., Matta, A., 2017. Or problems related to home health care: A review of relevant routing and scheduling problems. Operations Research for Health Care 13-14, 1–22. doi:https://doi.org/10.1016/j.orhc.2017.06.001.

Clark, C.E., 1961. The greatest of a finite set of random variables. Operations Research 9, 145–162. doi:https://doi.org/10.1287/opre.9.2.145.

Coelho, L.C., Cordeau, J.F., Laporte, G., 2012. Consistency in multi-vehicle inventory-routing. Transportation Research Part C: Emerging Technologies 24, 270–287. doi:https://doi.org/10.1016/j.trc.2012.03.007.

Deloitte, 2021. Canada's elder care crisis: Addressing the doubling demand. Available at https://www.cma.ca/latest-stories/demand-elder-care-set-double-10-years-new-report-commissioned-cma. (Accessed 20 March 2025).

Demirbilek, M., Branke, J., Strauss, A.K., 2021. Home healthcare routing and scheduling of multiple nurses in a dynamic environment. Flexible Services and Manufacturing Journal 33, 253–280. doi:https://doi.org/10.1007/s10696-019-09350-x.

Di Mascolo, M., Espinouse, M.L., Hajri, Z.E., 2017. Planning in home health care structures: A literature review. IFAC-PapersOnLine 50, 4654–4659. doi:https://doi.org/10.1016/j.ifacol.2017.08.689. 20th IFAC World Congress.

Di Mascolo, M., Martinez, C., Espinouse, M.L., 2021. Routing and scheduling in home health care: A literature survey and bibliometric analysis. Computers & Industrial Engineering 158, 107255. doi:https://doi.org/10.1016/j.cie.2021.107255.

Diabat, A., Archetti, C., Najy, W., 2021. The fixed-partition policy inventory routing problem. Transportation Science 55, 353–370. doi:https://doi.org/10.1287/trsc.2020.1019.

Ehmke, J.F., Campbell, A.M., Urban, T.L., 2015. Ensuring service levels in routing problems with time windows and stochastic travel times. European Journal of Operational Research 240, 539–550. doi:https://doi.org/10.1016/j.ejor.2014.06.045.

Elçi, Ö., Hooker, J., 2022. Stochastic planning and scheduling with logic-based benders decomposition. INFORMS Journal on Computing 34, 2428–2442.

Eveborn, P., Flisberg, P., Rönnqvist, M., 2006. Laps care—an operational system for staff planning of home care. European journal of operational research 171, 962–976. doi:https://doi.org/10.1016/j.ejor.2005.01.011.

Eveborn, P., Rönnqvist, M., Einarsdóttir, H., Eklund, M., Lidén, K., Almroth, M., 2009. Operations research improves quality and efficiency in home care. Interfaces 39, 18–34. doi:https://doi.org/10.1287/inte.1080.0411.

Fikar, C., Hirsch, P., 2017. Home health care routing and scheduling: A review. Computers & Operations Research 77, 86–95. doi:https://doi.org/10.1016/j.cor.2016.07.019.

Freeman, G., Hughes, J., 2010. Continuity of care and the patient experience. The King's Fund, Available at https://archive.kingsfund.org.uk/concern/published_works/000094996?locale=en#?cv=0. (Accessed 20 March 2025).

Gendreau, M., Jabali, O., Rei, W., 2014. Chapter 8: Stochastic vehicle routing problems, in: Vehicle Routing: Problems, Methods, and Applications, Second Edition. SIAM, pp. 213–239.

Gendreau, M., Jabali, O., Rei, W., 2016. 50th anniversary invited article—future research directions in stochastic vehicle routing. Transportation Science 50, 1163–1173. doi:https://doi.org/10.1287/trsc.2016.0709.

Goeke, D., Roberti, R., Schneider, M., 2019. Exact and heuristic solution of the consistent vehicle-routing problem. Transportation Science 53, 1023–1042. doi:https://doi.org/10.1287/trsc.2018.0864.

Grenouilleau, F., Lahrichi, N., Rousseau, L.M., 2020. New decomposition methods for home care scheduling with predefined visits. Computers & Operations Research 115, 104855. doi:https://doi.org/10.1016/j.cor.2019.104855.

Groër, C., Golden, B., Wasil, E., 2009. The consistent vehicle routing problem. Manufacturing & Service Operations Management 11, 630–643. doi:https://doi.org/10.1287/msom.1080.0243.

Gurobi Optimization, LLC, 2022. Gurobi Optimizer Reference Manual. URL: https://www.gurobi.com/wp-content/plugins/hd_documentations/documentation/9.5/refman.pdf. version 9.5.

Güven-Koçak, Ş., Heching, A., Keskinocak, P., Toriello, A., 2024. Continuity of care in home health care scheduling: a rolling horizon approach. Journal of Scheduling 27, 375–392. doi:https://doi.org/10.1007/s10951-023-00796-4.

Hashemi Doulabi, S.H., Rousseau, L.M., Pesant, G., 2016. A constraint-programming-based branch-and-price-and-cut approach for operating room planning and scheduling. INFORMS Journal on Computing 28, 432–448.

Heching, A., Hooker, J.N., 2016. Scheduling home hospice care with logic-based benders decomposition, in: International conference on AI and OR techniques in constriant programming for combinatorial optimization problems, Springer. pp. 187–197. doi:https://doi.org/10.1007/978-3-319-33954-2_14.

Heching, A., Hooker, J.N., Kimura, R., 2019. A logic-based benders approach to home healthcare delivery. Transportation Science 53, 510–522. doi:https://doi.org/10.1287/trsc.2018.0830.

Hewitt, M., Nowak, M., Nataraj, N., 2016. Planning strategies for home health care delivery. Asia-Pacific Journal of Operational Research 33, 1650041. doi:https://doi.org/10.1142/S021759591650041X.

Hoogeboom, M., Adulyasak, Y., Dullaert, W., Jaillet, P., 2021. The robust vehicle routing problem with time window assignments. Transportation Science 55, 395–413. doi:https://doi.org/10.1287/trsc.2020.1013.

Hooker, J., 2002. Logic, optimization, and constraint programming. INFORMS Journal on Computing 14, 295–321. doi:https://doi.org/10.1287/ijoc.14.4.295.2828.

Hooker, J., 2011. Logic-based methods for optimization: combining optimization and constraint satisfaction. John Wiley & Sons.

Hooker, J., Ottosson, G., 2003. Logic-based Benders decomposition. Mathematical Programming, Series A 96, 33–60. doi:https://doi.org/10.1007/s10107-003-0375-9.

IBM, 2022. IBM ILOG CPLEX Optimization Studio Documentation. URL: https://www.ibm.com/docs/en/icos/22.1.0. version 22.1.0.

IBM, 2025. Constraint programming modeling for python v2.30 documentation. Available at https://ibmdecisionoptimization.github.io/docplex-doc/cp/docplex.cp.modeler.py.html. (Accessed 28 July 2025).

Jabali, O., Leus, R., Woensel, T.V., Kok, T.D., 2015. Self-imposed time windows in vehicle routing problems. OR Spectrum 37, 331–352. doi:https://doi.org/10.1007/s00291-013-0348-1.

Jaillet, P., Qi, J., Sim, M., 2016. Routing optimization under uncertainty. Operations Research 64, 186–200. doi:https://doi.org/10.1287/opre.2015.1462.

Kovacs, A.A., Golden, B.L., Hartl, R.F., Parragh, S.N., 2014a. Vehicle routing problems in which consistency considerations are important: A survey. Networks 64, 192–213. doi:https://doi.org/10.1002/net.21565.

Kovacs, A.A., Golden, B.L., Hartl, R.F., Parragh, S.N., 2015a. The generalized consistent vehicle routing problem. Transportation Science 49, 796–816. doi:https://doi.org/10.1287/trsc.2014.0529.

Kovacs, A.A., Parragh, S.N., Hartl, R.F., 2014b. A template-based adaptive large neighborhood search for the consistent vehicle routing problem. Networks 63, 60–81. doi:https://doi.org/10.1002/net.21522.

Kovacs, A.A., Parragh, S.N., Hartl, R.F., 2015b. The multi-objective generalized consistent vehicle routing problem. European Journal of Operational Research 247, 441–458. doi:https://doi.org/10.1016/j.ejor.2015.06.030.

Kwon, J., Coifman, B., Bickel, P., 2000. Day-to-day travel-time trends and travel-time prediction from loop-detector data. Transportation Research Record 1717, 120–129. doi:https://doi.org/10.3141/1717.

Laborie, P., 2009. IBM ILOG CP Optimizer for detailed scheduling illustrated on three problems, in: In Proceedings of the 6th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CPAIOR 2009), pp. 148–162.

Laborie, P., Rogerie, J., Shaw, P., Vilím, P., 2018. Ibm ilog cp optimizer for scheduling: 20+ years of scheduling with constraints at ibm/ilog. Constraints 23, 210–250. doi:https://doi.org/10.1007/s10601-018-9281-x.

Lanzarone, E., Matta, A., 2014. Robust nurse-to-patient assignment in home care services to minimize overtimes under continuity of care. Operations Research for Health Care 3, 48–58. doi:https://doi.org/10.1016/j.orhc.2014.01.003.

Lee, J., Hong, B., Lee, K., Jang, Y.J., 2015. A prediction model of traffic congestion using weather data, in: 2015 IEEE International Conference on Data Science and Data Intensive Systems, IEEE. pp. 81–88. doi:10.1109/DSDIS.2015.96.

Liu, C., Lin, S., Shen, Z.J.M., Zhang, J., 2023. Stochastic service network design: The value of fixed routes. Transportation Research Part E: Logistics and Transportation Review 174, 103118. doi:https://doi.org/10.1016/j.tre.2023.103118.

Liu, M., Zhao, Y., Xie, X., 2024. Continuity-skill-restricted scheduling and routing problem: Formulation, optimization and implications. IISE Transactions 56, 201–220. doi:https://doi.org/10.1080/24725854.2023.2215843.

Liu, M., Zhao, Y., Xie, X., 2025. Express: Last-mile attended home healthcare delivery: A robust strategy to mitigate cascading delays and ensure punctual services. Production and Operations Management , 10591478251318914doi:https://doi.org/10.1177/10591478251318914.

Liu, R., Yuan, B., Jiang, Z., 2019. A branch-and-price algorithm for the home-caregiver scheduling and routing problem with stochastic travel and service times. Flexible Services and Manufacturing Journal 31, 989–1011. doi:https://doi.org/10.1007/s10696-018-9328-8.

Mancini, S., Gansterer, M., 2021. Vehicle scheduling for rental-with-driver services. Transportation Research Part E: Logistics and Transportation Review 156, 102530. doi:https://doi.org/10.1016/j.tre.2021.102530.

Martínez, K.P., Adulyasak, Y., Jans, R., 2022. Logic-based Benders decomposition for integrated process configuration and production planning problems. INFORMS Journal on Computing 34, 2177–2191.

Matta, A., Chahed, S., Sahin, E., Dallery, Y., 2014. Modelling home care organisations from an operations management perspective. Flexible Services and Manufacturing Journal 26, 295–319. doi:https://doi.org/10.1007/s10696-012-9157-0.

Nadarajah, S., Kotz, S., 2008. Exact distribution of the max/min of two gaussian random variables. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 16, 210–212. doi:10.1109/TVLSI.2007.912191.

Naderi, B., Ruiz, R., Roshanaei, V., 2023. Mixed-integer programming vs. constraint programming for shop scheduling problems: new results and outlook. INFORMS Journal on Computing 35, 817–843. doi:https://doi.org/10.1287/ijoc.2023.1287.

Nickel, S., Schröder, M., Steeg, J., 2012. Mid-term and short-term planning support for home health care services. European Journal of Operational Research 219, 574–587. URL: https://www.sciencedirect.com/science/article/pii/S0377221711009829, doi:https://doi.org/10.1016/j.ejor.2011.10.042. feature Clusters.

Oh, S., Byon, Y.J., Jang, K., Yeo, H., 2015. Short-term travel-time prediction on highway: a review of the data-driven approach. Transport Reviews 35, 4–32. doi:https://doi.org/10.1080/01441647.2014.992496.

Park, D., Rilett, L.R., 1999. Forecasting freeway link travel times with a multilayer feedforward neural network. Computer-Aided Civil and Infrastructure Engineering 14, 357–367. doi:https://doi.org/10.1111/0885-9507.00154.

Redd, J., 2022. Top 10 complaints clients have about home care agencies. Available at https://activatedinsights.com/articles/top-10-complaints-clients-have-about-home-care-agencies/. (Accessed 27 July 2025).

Restrepo, M.I., Rousseau, L.M., Vallée, J., 2020. Home healthcare integrated staffing and scheduling. Omega 95, 102057. doi:https://doi.org/10.1016/j.omega.2019.03.015.

Roshanaei, V., Booth, K.E., Aleman, D.M., Urbach, D.R., Beck, J.C., 2020. Branch-and-check methods for multi-level operating room planning and scheduling. International Journal of Production Economics 220, 107433.

Shi, Y., Boudouh, T., Grunder, O., 2019. A robust optimization for a home health care routing and scheduling problem with consideration of uncertain travel and service times. Transportation Research Part E: Logistics and Transportation Review 128, 52–95. doi:https://doi.org/10.1016/j.tre.2019.05.015.

Shi, Y., Boudouh, T., Grunder, O., Wang, D., 2018. Modeling and solving simultaneous delivery and pick-up problem with stochastic travel and service times in home health care. Expert Systems with Applications 102, 218–233. doi:https://doi.org/10.1016/j.eswa.2018.02.025.

Song, Y., Ulmer, M.W., Thomas, B.W., Wallace, S.W., 2020. Building trust in home services—stochastic team-orienteering with consistency constraints. Transportation Science 54, 823–838. doi:https://doi.org/10.1287/trsc.2019.0927.

Spliet, R., Gabor, A.F., 2015. The time window assignment vehicle routing problem. Transportation Science 49, 721–731. doi:https://doi.org/10.1287/trsc.2013.0510.

Subramanyam, A., Gounaris, C.E., 2017. A decomposition algorithm for the consistent traveling salesman problem with vehicle idling. Transportation Science 52, 386–401. doi:https://doi.org/10.1287/trsc.2017.0741.

Subramanyam, A., Wang, A., Gounaris, C.E., 2018. A scenario decomposition algorithm for strategic time window assignment vehicle routing problems. Transportation Research Part B: Methodological 117, 296–317. doi:https://doi.org/10.1016/j.trb.2018.09.008.

Sungur, I., Ren, Y., Ordóñez, F., Dessouky, M., Zhong, H., 2010. A model and algorithm for the courier delivery problem with uncertainty. Transportation Science 44, 193–205. doi:https://doi.org/10.1287/trsc.1090.0303.

Tarantilis, C.D., Stavropoulou, F., Repoussis, P.P., 2012. A template-based tabu search algorithm for the consistent vehicle routing problem. Expert Systems with Applications 39, 4233–4239. doi:https://doi.org/10.1016/j.eswa.2011.09.111.

Tsang, M.Y., Shehadeh, K.S., 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, 48–63. doi:https://doi.org/10.1016/j.ejor.2022.09.020.

Wang, K., Zhen, L., Xia, J., Baldacci, R., Wang, S., 2021. Routing optimization with generalized consistency requirements. Transportation Science 56, 223–244. doi:https://doi.org/10.1287/trsc.2021.1072.

Wang, R., Shehadeh, K.S., Xie, X., Li, L., 2022. Integrated home care staffing and capacity planning: Stochastic optimization approaches. arXiv preprint arXiv:2203.14430 doi:https://doi.org/10.48550/arXiv.2203.14430.

Wang, T., Meskens, N., Duvivier, D., 2015. Scheduling operating theatres: Mixed integer programming vs. constraint programming. European Journal of Operational Research 247, 401–413. doi:https://doi.org/10.1016/j.ejor.2015.06.008.

Woodward, C.A., Abelson, J., Tedford, S., Hutchison, B., 2004. What is important to continuity in home care?: Perspectives of key stakeholders. Social Science & Medicine 58, 177–192. doi:https://doi.org/10.1016/S0277-9536(03)00161-8.

Wu, S., Jin, C., Bo, H., 2025. Exact solution of workload consistent vehicle routing problem with priority distribution and demand uncertainty. Computers & Industrial Engineering 202, 110940. doi:https://doi.org/10.1016/j.cie.2025.110940.

Xie, W., Liu, T., Li, X., Zheng, C., 2023. Robust homecare service capacity planning. Computers & Operations Research 154, 106155. doi:https://doi.org/10.1016/j.cor.2023.106155.

Yang, M., Ni, Y., Song, Q., 2022. Optimizing driver consistency in the vehicle routing problem under uncertain environment. Transportation Research Part E: Logistics and Transportation Review 164, 102785. doi:https://doi.org/10.1016/j.tre.2022.102785.

Yang, M., Ni, Y., Yang, L., 2021. A multi-objective consistent home healthcare routing and scheduling problem in an uncertain environment. Computers & Industrial Engineering 160, 107560. doi:https://doi.org/10.1016/j.cie.2021.107560.

Yao, Y., Van Woensel, T., Veelenturf, L.P., Mo, P., 2021. The consistent vehicle routing problem considering path consistency in a road network. Transportation Research Part B: Methodological 153, 21–44. doi:https://doi.org/10.1016/j.trb.2021.09.005.

Yu, X.P., Hu, Y.S., Wu, P., 2024. The consistent vehicle routing problem considering driver equity and flexible route consistency. Computers & Industrial Engineering 187, 109803. doi:https://doi.org/10.1016/j.cie.2023.109803.

Yuan, B., Liu, R., Jiang, Z., 2015. A branch-and-price algorithm for the home health care scheduling and routing problem with stochastic service times and skill requirements. International Journal of Production Research 53, 7450–7464. doi:https://doi.org/10.1080/00207543.2015.1082041.

Zhan, Y., Wang, Z., Wan, G., 2021. Home service routing and appointment scheduling with stochastic service times. European Journal of Operational Research 288, 98–110. doi:https://doi.org/10.1016/j.ejor.2020.05.037.

Zhang, Y., Baldacci, R., Sim, M., Tang, J., 2019. Routing optimization with time windows under uncertainty. Mathematical Programming 175, 263–305. doi:https://doi.org/10.1007/s10107-018-1243-y.