Tackling Large-scale Home Health Care Delivery Problem with Uncertainty

Cen Chen **Zachary B. Rubinstein** **Stephen F. Smith** **Hoong Chuin Lau**

School of Information Systems
Singapore Management University
80 Stamford Road
Singapore 178902
cenchen.2012@smu.edu.sg

The Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{zbr, sfs}@cs.cmu.edu

School of Information Systems
Singapore Management University
80 Stamford Road
Singapore 178902
hclau@smu.edu.sg

Abstract

In this work, we investigate a multi-period Home Health Care Scheduling Problem (HHCSP) under stochastic service and travel times. We first model the deterministic problem as an integer linear programming model that incorporates real-world requirements, such as time windows, continuity of care, workload fairness, inter-visit temporal dependencies. We then extend the model to cope with uncertainty in durations, by introducing chance constraints into the formulation. We propose efficient solution approaches, which provide quantifiable near-optimal solutions and further handle the uncertainties by employing a sampling-based strategy. We demonstrate the effectiveness of our proposed approaches on instances synthetically generated by real-world dataset for both deterministic and stochastic scenarios.

Introduction

Home health care provides a wide range of health care services that are delivered at patients' homes. These medical services range from cleaning, personal hygiene to administering some medical treatments, such as blood pressure tests, prescriptions, injections and so on. Over the past decade, an increasing number of people subscribe to home health care services, especially for patients with chronic conditions to minimize cost and maximize the quality of life. According to the US National Association for Home Care & Hospice (2010), roughly 12 million people received home health care services from 33,000 providers in 2008. This number is set to grow rapidly with an increasing aging population: Population Reference Bureau (2016) predicts that "the number of Americans above 65 will increase from 46 million in 2016 to over 98 million by 2060". Against the manpower crunch in health-care professionals, in-home service providers are under increasing pressure to provide high-quality service at a low cost to an ever growing demand.

Home health care services comprise a complex set of processes requiring significant coordination among health care providers. An important task for health care providers is to achieve a high service level, i.e, satisfy as many patients' needs as much as possible, in a cost efficient manner. However, many home health care companies still rely on health care providers to individually coordinate with patients, which is very inefficient, and often myopic as it is not well coordinated among health care providers. With the increase service demand, more and more time is being spent in the coordination processes, resulting in less time available for services. For example, the Palliative Care Department at West Pennsylvania Hospital reports that it takes an hour a day to schedule that day's services, significantly reducing the time health care providers can spend addressing patients' needs. Thus, it is critical for health care providers to increase the efficiency of the scheduling processes, more precisely, to design a better automated mechanism, that schedules visits in optimized fashion subject to a complex set of constraints.

In the literature, a considerable amount of problem specific systems have been developed for home health care settings across the world. For example, Begur, Miller, and Weaver (1997) first developed a spatial decision support system for Visiting Nursing Association in the United States to minimize total travel time, taking into consideration of routing, provider availabilities, and fixed visitation frequencies. Eveborn, Flisberg, and Rönnqvist (2006) presented a single day scheduling system, called LAPS CARE, for Swedish health care system that maximizes the number of served requests and considers time windows, skill requirements, and breaks. We believe the existing systems/practices fell short in the following two key aspects.

- **Failure to address the uncertainties**: Due to the varying health conditions of the patients as well as the experience of healthcare providers, the service durations can have quite substantial variance. Figure 1 plots the means and standard deviations of actual service durations over one month of *actual visit records* grouped by different service disciplines. We can see that service duration uncertainty is clearly exhibited, where standard deviations under all disciplines are larger than 10 minutes up to 30 minutes. In addition, unforeseen traffic conditions, such as congestion, accidents, and breakdowns, often result in varying travel durations, thus further complicates the prediction of service start times. Faced with both types of duration uncertainties, providers responsible for scheduling are confronted with the dilemma of either over-estimating duration times so as to guarantee services and travel and, hence, under-utilizing their resources, or under-estimating those times and, thus, short-changing patients by making them wait for services. In any case, duration uncertainties can potentially deteriorate patient satisfaction and result in additional costs for the company.

- **Insufficient business considerations**: Existing automated scheduling systems differ significantly, as problems originate from different regions with various requirements and regulations. Fikar and Hirsch (2017) recently presented a comprehensive survey on the home health care problem. To the best of our knowledge, none of the existing works completely addresses all the business requirements raised by our problem.

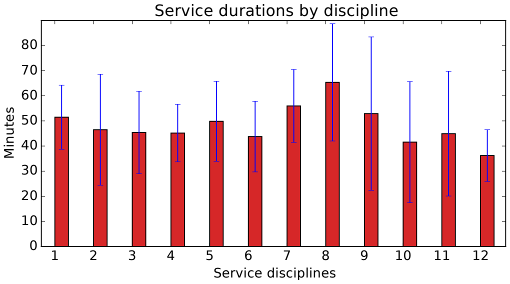

Figure 1: Means and standard deviations of service durations over one-month visits grouped by different service disciplines. Note, the statistics are summarized using one month of *actual visit* records of Sept. 2015. Discipline specifies the type of service required, e.g., speech therapy, skilled nursing, physical therapy, to name a few.

To address these challenges, we propose to develop a decision-support system for the efficient delivery of services to patients at home, while tightly integrating both patients' and providers' preferences. In this paper, we focus on investigating a large-scale scheduling framework that generates the start-of-the-day schedules, considering realistic scenarios driven by our real-world needs, with a comprehensive set of considerations for home health care settings, e.g., time windows, continuity of care, workloads, inter-visit temporal dependencies, and especially the *duration uncertainties*. Our goal is to generate a set of routes for health care providers to service visit that maximize the *patients' satisfaction*.

We make the following key contributions in our paper. First, we begin with developing a mathematical formulation for the deterministic version of the problem as a baseline, to handle various business requirements in home health care, and ultimately present the robust version of the formulation to cope with uncertainty in service/travel durations. Second, we present scalable solution approaches to solve both the deterministic and stochastic problems.

Related Work

The Home Health Care Scheduling Problem (HHCSP) is the problem of scheduling and routing service providers to visit patients at home. It was first mentioned by (Fernandez et al. 1974), where community nurses are scheduled to visits patients. Essentially, HHCSP instantiates some form of *workforce scheduling* and *routing* problems. Castillo-Salazar, Landa-Silva, and Qu (2016) review the workforce scheduling and routing problem, not limited to the health care industry and view the home health care as one of the application domains. Other similar applications can also be found in the technician scheduling, emergency services, transportation systems, or call centers (Ernst et al. 2004; Castillo-Salazar, Landa-Silva, and Qu 2016). Problems from different domains share some commonality, and also differs substantially in problem settings. Given a large amount of literature, we focus on the research more related to our problem and heath care domain.

On the workforce scheduling aspect, literature on workforce scheduling in health care domain mainly deals with staff rostering requirements, such as skills, shifts, time-related constraints, work regulations and ect (Burke et al. 2004; Rais and Viana 2011). On the routing aspect, vehicle routing (Toth and Vigo 2014) and orienteering problems (Gunawan, Lau, and Vansteenwegen 2016) have long been extensively studied to model various problems from transportation and logistics. In literature, HHCSP is often modeled as the vehicle routing problem (Fikar and Hirsch 2017), with assumption that the workforce is sufficient and the goal is to cover all the requests with the minimal travel costs or manpower. In our problem, however, the company faces an oversubscribed situation, where the number of requests received often exceeds its service capacity, especially under the case when we care about patient preferences and continuity of care. The current practice for the company is to first schedule requests for full-time providers. Any uncovered requests are then made up by a combination of either extending the visits for full-time providers or outsourcing to part-time providers. We are concerned with scheduling full-time workforce with the objective to maximize patients' satisfaction, measured in terms of rewards. Hence, the underlying problem at stake is an *orienteering problem* rather than a vehicle routing problem. Orienteering problem is motivated by scheduling score-orienteering events in which participants get rewards from visiting a selected subset of nodes within the time budget. We model the problem as a variant of the team orienteering problem with time windows.

There is also a thread of research dedicated to home health care. Fikar and Hirsch (2017) recently present a comprehensive survey on HHCSP. The majority of the related works focuses on single-period problems, i.e., a single day as scheduling horizon, and only few papers consider multi-period ones, i.e., over multiple days. Our problem can be categorized as a *multi-period HHCSP*. The problem becomes much more challenging and complicated when going into a longer scheduling horizon, as the scheduling process involve more complex assignments, regulations, and continuity of care. Problems differ substantially due to different business considerations. Time windows, qualifications, and provider availabilities are commonly addressed in multi-period problems, while other aspects, such as workload fairness, continuity of care, inter-visit temporal dependencies are less incorporated, especially uncertainties (Fikar and Hirsch 2017). Next, we focus on the multi-period HHCSP literature.

In multi-period HHCSP literature, there have been works addressing the workload fairness as the objective function, to minimize the workload difference among providers or optimize utilization factors (Hertz and Lahrichi 2009; Barrera, Velasco, and Amaya 2012; Cappanera and Scutellà ; Errarhout, Kharraja, and Corbier 2016). Yuan and Fügenschuh (2015) handle the workload in the objective to minimize daily working hours. In our problem, the company desires a provider to service a minimum workload if possible and overworking is not preferred. We model the minimum workload as soft constraints as the penalty in the objective function and maximum workload as hard constraints.

Considering the continuity of care, works by (Bachouch, Guinet, and Hajri-Gabouj 2011; Carello and Lanzarone 2014) model it as hard constraints that require patients to be visited strictly by the same providers over the scheduling horizon, which is not flexible to a certain extent. Nickel, Schröder, and Steeg (2012) try to minimize the sum of, over all the providers, the different providers serving the same the patient, while Rodriguez et al. (2015) allow a patient to be visited by a maximum number of different service providers. However, these providers are treated equally without priorities. Lin et al. (2016) use five cases of hard weight allocation criteria to enforce care continuity and incorporate priority. Instead, we take a softer and flexible approach. Our objective is to maximize patients' satisfaction, measured in terms of rewards collected for serving the patients. The provider-request dependent rewards give us the flexibility to reflect the continuity of care with prioritized provider candidates and to further capture requests' information, such as type, emergency or request priority.

Inter-visit dependency is another aspect to consider. Rasmussen et al. (2012) study a single-day problem with five types of temporal precedences. Existing research on the multi-period problem focuses on day-level temporal dependencies. Some impose fixed visiting days (Begur, Miller, and Weaver 1997; Trautsamwieser and Hirsch 2014), while some assume service frequency, often handled by using predefined service combinations as the input and let the model decide the visiting days (Bennett and Erera 2011; Shao, Bard, and Jarrah 2012; Nickel, Schröder, and Steeg 2012; Yuan and Fügenschuh 2015). In our problem, we handle both fixed visiting days, as well as visiting service frequencies for patients with flexible availabilities.

With regards to handling uncertainties, we notice that most of the existing works with uncertainties consider the uncertain demands (Koeleman, Bhulai, and van Meersbergen 2012; Carello and Lanzarone 2014; Rodriguez et al. 2015; Bowers et al. 2015), whereas stochastic service or travel times are rarely studied in HHCSP. Yuan, Liu, and Jiang (2015) present an exact branch-and-price algorithm to tackle a single-day problem with uncertain service times, which solves small instances up to 50 patients. It focuses on the penalty for late arrivals in the objective, instead of enforcing the time windows. Errarhout, Kharraja, and Corbier (2016) propose a two-stage model in a multi-period setting to cater uncertain service times and solve the model by CPLEX with instances up to 11 nurses and 75 patients. However, time-windows and inter-visit temporal dependencies are not respected. The introduction of uncertain durations makes the problems with time windows even harder to solve. In this problem, we incorporate duration uncertainties by enforcing a set of time window chance constraints.

In summary, to the best our knowledge, none of the existing works can be readily extended to handle our problem.

Home Health Care Scheduling Problem

In this section, we provide the formal definition for our multi-period home health care scheduling problem, motivated by the requirements from a leading home health care and hospice company in Pittsburgh. The problem is defined as the following tuple:

$$\langle D, N, T, R, K\rangle$$

D denotes the set of days d for the scheduling horizon, i.e., $d \in D$. Each day is discretized into minutes, indexed from 1 to 1440. *N* represents the set of all nodes, $N = N_t \cup N_k$, where N_t and N_k are the set of patient' requested visit locations and health care providers' start and end locations, respectively. T is the pairwise travel time matrix and $t_{ij} \in T$ denotes the travel time between node i and j.

R represents the set of patients' requests, for the given scheduling horizon, e.g., the next 7 days. Each request represents a service task to be specifically performed at a patient's home. A request belongs to only one patient, and a patient may specify several requests over the scheduling horizon. Each request i is characterized by a tuple $\langle n_i, a_i, w_i, q_i^{req}, \{[o_{id}, c_{id}]\}\rangle$ where:

- $n_i \in N_t,\ a_i$ and w_i are the service location, service duration, units of work required by the request, respectively.
- q_i^{req} is the service discipline specifying the type of service that the request needs, e.g., speech therapy, skilled nursing, physical therapy, just to name a few.
- $\{[o_{id}, c_{id}]\}$ refers to the set of available **time windows**, during which a patient wish servicing of a request i to be started. It is possible that a patient indicates several available time windows for the same request i over different days d – e.g., Alice is free on Monday morning and Thursday 2pm-4pm for a physical therapy treatment, from which health care providers have to decide the best time slot to allocate. More specifically, o_{id} and c_{id} are the earliest and latest start-times for a time window $[o_{id}, c_{id}]$. If a provider arrives *earlier* than o_{id}, he will *wait* until the time window opens. While arriving *later* than c_{id} will lead to the violation of the time window constraint.

K represents the set of health care providers. Each health care provider $k \in$ *K* has a set of qualifications that he holds. Typically, qualifications are flat and distinct, e.g., nurse, nutritionist, and therapist. Each provider is constrained by an availability time window $[T_{kd}^1, T_{kd}^2]$ on each day *d*, i.e., he will leave his start node N_k^1 at time T_{kd}^1, and return to the end node N_k^2 before time T_{kd}^2.

A request is considered as *completed* if and only if a **qualified and available** provider starts the service within the patient's available time window and stays with the patient for the whole service duration.

The goal is to schedule and route service providers for home health care visits on a weekly basis that considers the requirements from both patients and service providers. The problem is further subject to the following considerations:

- **Inter-visit Temporal Dependency**: A patient may subscribe several visits/requests over the same week. These requests can be temporally dependent such that request j has to be fulfilled at least D_{ij}^{-} days after and no more than D_{ij}^{+} days after servicing request i. Such inter-visits dependencies are often seen in practice.
- **Workload Fairness**: A provider is paid a fixed salary as long as he is working on a day. The company desires a provider to service at least W^{-} units of work on a daily basis, if possible. At the same time, a maximum working unit W^{+} is imposed, as overworking is not preferred.
- **Continuity of Care**: Each patient has a set of prioritized provider candidates and is preferred to be visited by his primary provider. This is important for patient satisfaction, especially for patients with chronic conditions.
- **Uncertain Durations**: In real-world scenarios, travel and service durations are usually uncertain. We assume travel times t_{ij} and service times a_i are random variables following certain distributions.

Mathematical Model

In this section, we first propose an ILP model for the deterministic problem, followed by incorporating the duration uncertainties. Decision variables are summarized as follows:

Variables	Descriptions
$x_{ijk}^d \in \{0,1\}$	set to 1 if provider k serves request j right after i on day d.
$y_{ik}^d \in \{0,1\}$	set to 1 if request i is assigned to provider k on day d.
$v_i \in \{0,1\}$	set to 1 if request i is not assigned to any provider.
$e_{ik}^d \in \{1,...,T\}$	service start time of request i for provider k on day d.
$f_{kd} \in \{0,1\}$	set to 1 if provider k is assigned with requests on day d.
$p_{kd} \in \{0,1\}$	set to 1 if the route for provider k on day d is penalized.

Intuitively, reward will be collected if a request is completed by a provider. Let r_{ik} be the provider-dependent rewards, defined based on the units of work (w_i) the request needs and whether this provider is the patient's primary provider under this discipline(l_{ik}). Thus, we have:

$$r_{ik} = r \cdot w_i + r^{+} \cdot l_{ik},$$

where r is a constant base reward and r^{+} is the additional reward assigned for the primary provider. The reward structure helps the *continuity of care*, where there is an incentive to assign primary providers to patients, and promotes the productivity, where higher reward will be collected for request requiring more units of work. The bigger the r^{+}, the stronger the enforcement of care continuity.

The objective of this model is to generate a sequence of requests to visit for each provider on each day that maximizes the expected total rewards collected for the whole team considering the route penalties. γ is the amount of penalty incurred if a route does not meet the minimum workload.

Constraints (2) ensure each request is assigned at most one provider over the entire scheduling horizon. As K_{id} denotes the set of providers who are qualified and available to serve request i on day d, constraints (3) make sure that requests will not be assigned to any unavailable or unqualified providers. Constraints (4) reflect requests' inter-visit temporal dependencies. Constraints (5) specify that a route exists when it contains any request. Constraints (6) enforce the minimum and maximum workload requirements on every route. M is a large positive number. If a provider works less than W^{-} units on a day, the route for that provider on that day will be penalized (i.e., $p_{kd} = 1$). Note, if a provider is not assigned any requests on a day, it will not be penalized, as there is no cost incurred by the health care company. Constraints (7) bind the decision variables y_{ik}^d with decision variables x_{ijk}^d, which ensure that requests are assigned only when they can be visited.

$$\max \sum_{i \in N_t} \sum_{d \in D} \sum_{k \in K} r_{ik} \cdot y_{ik}^d - \gamma \cdot \sum_{d \in D} \sum_{k \in K} p_{kd}, \tag{1}$$

$$v_i + \sum_{d \in D} \sum_{k \in K} y_{ik}^d = 1, \qquad \forall i \in N_t, \tag{2}$$

$$y_{ik}^d = 0 \qquad \forall i \in N_t;\ d \in D;\ k \in \{K \setminus K_{id}\}, \tag{3}$$

$$\begin{cases} D_{ij}^{-} - M(v_i + v_j) \leqslant \sum_{k \in K} \sum_{d \in D} d \cdot (y_{jk}^d - y_{ik}^d) \\ \sum_{k \in K} \sum_{d \in D} d \cdot (y_{jk}^d - y_{ik}^d) \leqslant D_{ij}^{+} + M(v_i + v_j), \end{cases} \qquad \forall i, j \in N_t, \tag{4}$$

$$y_{ik}^d \leqslant f_{kd}, \qquad \forall i \in N_t; k \in K; d \in D, \tag{5}$$

$$\begin{cases} W^{-} \cdot f_{kd} - \sum_{i \in N_t} y_{ik}^d \cdot w_i \leqslant M \cdot p_{kd}, \\ \sum_{i \in N_t} y_{ik}^d \cdot w_i \leqslant W^{+} \qquad \forall k \in K; d \in D, \end{cases} \tag{6}$$

$$y_{ik}^d \leqslant \sum_{j \in N} x_{ijk}^d \qquad \forall i \in N_t; k \in K; d \in D, \tag{7}$$

Table 1: The Deterministic Model-Part1

The rest of the constraints (8) - (12) are provider-day (k, d) level routing constraints. For each provider $k \in K$ on each day $d \in D$, the same set of constraints applies.

$$\begin{cases} \sum_{j \in N} x_{ijk}^d = \sum_{j \in N} x_{jik}^d, & \forall i \in N \setminus \{N_k^1, N_k^2\}, \\ \sum_{j \in N} x_{ijk}^d - \sum_{j \in N} x_{jik}^d = 1, & i = N_k^1, \\ \sum_{j \in N} x_{jik}^d - \sum_{j \in N} x_{ijk}^d = 1, & i = N_k^2, \end{cases} \tag{8}$$

$$e_{ik}^d = T_{kd}^1, \qquad i = N_k^1, \tag{9}$$

$$e_{ik}^d + a_i + t_{ij} - e_{jk}^d \leqslant M(1 - x_{ijk}^d), \quad \forall i, j \in N, \tag{10}$$

$$o_{id} \leqslant e_{ik}^d \leqslant c_{id}, \qquad \forall i \in N_t, \tag{11}$$

$$e_{ik}^d \leqslant T_{kd}^2, \qquad i = N_k^2. \tag{12}$$

Table 2: The Deterministic Model-Part2

Constraints (8) ensure that every provider k starts and ends at his specified nodes N_k^1 and N_k^2 and the connectivity of the nodes. Note that, e_{ik}^d refers to the start service time of node i. Early service or late service will cause the violation of the request time window constraints. Constraints (9) initialize the start service time of the start node for provider k. Timing consistency constraints are enforced in constraints (10). For every provider k, if node j is visited immediately after node i (i.e., $x^d_{ijk} = 1$), the start service time e^d_{jk} of node j should be at least $(a_i + t_{ij})$ time units later than the start service time e^d_{ik} of node i (also eliminates the subtours). Constraints (11) make sure that requests' time windows are respected. Finally, constraints (12) limit the time budget.

Modelling Duration Uncertainty

To incorporate the uncertain durations, we then extend the deterministic model. We assume travel times t_{ij} and service times a_i are random variables following certain distributions. We assume these distributions are given as problem input, which can be derived from historical records obtained from the home health care domain. Note that in the deterministic model, these durations only affect the time window and time budget constraints. Thus, to model duration uncertainty, we replace the time window constraints (11) and time budget constraints (12) by a set of chance constraints (13) and (14), while the rest of the constraints remain the same as the deterministic formulation. Thus, we have:

$$\max \;\; Objective\ (1)$$
$$s.t.\;\; Constraints\ (2)-(10)$$
$$P(o_{id} \leqslant e^d_{ik} \leqslant c_{id}) \geqslant 1-\alpha, \forall i \in N_t; k \in K; d \in D \quad (13)$$
$$P(e^d_{ik} \leqslant T^2_{kd}) \geqslant 1-\alpha, \qquad \forall i = N^2_k; k \in K; d \in D. \quad (14)$$

Table 3: The Chance Constrained Model

The chance constraints (13) and (14) enforce the probability of satisfying the respective constraints are at least $1-\alpha$, $\alpha \in [0,1]$. α is the *risk level*, indicating the decision maker's level of conservativeness.

Solution Approach

The deterministic model of Table (1-2) can be solved by commercial solvers such as CPLEX. However, it is not scalable with increasing number of providers and requests with longer scheduling horizon. In this section, we propose to use Lagrangian relaxation and dual decomposition to improve its scalability.

Lagrangian Relaxation (LR) is a widely used technique in combinatorial optimization, where we approximate the original difficult primal problem by a simpler dual problem (Fisher 1981). The basic idea is to relax the complicating constraints into the objective function and penalize violations of the constraints using Lagrangian multipliers.

We relax constraints (7), which couple all the assignment and routing decision variables together, into the objective function by applying Lagrangian relaxation.

$$\min\ L(\lambda) = -\sum_{i\in N_t}\sum_{d\in D}\sum_{k\in K} y^d_{ik}\cdot r_{ik} + \gamma\cdot\sum_{d\in D}\sum_{k\in K} p_{kd} + \sum_{i\in N_t}\sum_{d\in D}\sum_{k\in K}\lambda^d_{ik}(y^d_{ik} - \sum_{j\in N} x^d_{ijk}). \quad (15)$$

We then *decompose* the relaxed dual problem $L(\ \lambda\ )$ into one assignment sub-problem, which assigns the requests to the providers, and provider-day-level sub-problems, which find the routes for the providers.

Assignment Sub-problem: The assignment sub-problem is defined as:

$$\min \sum_{i\in N_t}\sum_{d\in D}\sum_{k\in K} y^d_{ik}\cdot(\lambda^d_{ik} - r_{ik}) + \gamma\cdot\sum_{d\in D}\sum_{k\in K} p_{kd},$$
$$s.t.\ Constraints\ (2)-(6)$$

The assignment ILP model can be solved exactly by CPLEX. To further scale up the sub-problem, we can apply linear relaxation. The idea is to drop integer constraints for decision variables, which transforms hard ILP problem into an easier polynomial solvable linear problem (LP). In minimization problem, the objective value achieved by the LP is always smaller or equal to that of the original ILP, thus it can serve as a lower bound for the assignment sub-problem.

Routing Sub-problems: There are K · D independent provider-day-level routing sub-problems for each provider k and on each day d:

$$\min -\sum_{i\in N_t}\sum_{j\in N}\lambda^d_{ik}\cdot x_{ij}, \quad (16)$$
$$s.t.\, Constraints\ (8)-(12)$$
$$\sum_{j\in N} x_{ij} \leqslant 1, \qquad \forall i \in N_t, \quad (17)$$
$$\sum_{i\in N_t}\sum_{j\in N} x_{ij}\cdot w_i \leqslant W^+. \quad (18)$$

Constraints (17) and (18) are included to further tighten the sub-problems, such that one node is visited at most once in a route and the maximum workload is enforced (i.e., no more than W^+ units of work in a route). This routing sub-problem can be viewed as an orienteering problem, which is NP-hard. Instead of solving the routing ILP, we develop a search algorithm (1), that exploits the routing problem structure. During the search phase, we systematically extend the routes and discard unpromising dominated routes, which guarantees to find the optimal routing solution. Before going into the algorithm details, we first present the following observation.

Observation - *Route Comparability: two feasible routes are comparable if and only if they contain exactly the same set of nodes and end at the same node, i.e.,* $|r_1| = |r_2|$ *and* $r^k_1 = r^k_2$ *(assume k is the last node in the route).*
Observation 1 holds because two feasible routes having the same set of nodes with different visiting sequences and the same ending node leads to the same objective value, and provides a fair starting point for further route extension.

For two comparable routes, route r_1 *dominates* route r_2 if they are comparable and the total time for r_1 is less than the total time for r_2. The intuition is that the route with the shorter total time will have more room for future insertion. In algorithm (1), we insert only nodes with positive Lagrangian multipliers(λ^d_{ik}) into the route to improve the objective value. A node can be inserted into the route if and only if all the time windows, time budget, and maximum workload constraints are satisfied (i.e., GETFEASIBLEREQUEST()). We start with a route with only one node, i.e., the provider's start location. At each iteration, *non-dominated* routes from the last iteration are expanded by one more *feasible* node. A node is considered as feasible if the resulting extended route satisfies the time window and time budget requirements. By doing so, routes of longer length are generated. Only *non-dominated* routes will be stored and dominated routes will be pruned. Note, with the dominance, if there are several comparable routes with the same shortest total time, we will keep just the first one. The search procedure stops when no route can be further expanded.

Algorithm 1: BFS with Dominance Prunning

```
Input: $(\lambda^d_{ik}, N)$
$R \leftarrow$ INITIALIZE(), CONTINUE$\leftarrow$TRUE
while CONTINUE do
    for $r \in R$ do
        $N_{feasible} \leftarrow$ GETFEASIBLEREQUEST$(r, N, \lambda^d_{ik})$
        for $n \in N_{feasible}$ do
            $r^{'} \leftarrow$ EXTENDROUTE$(r, n, \lambda^d_{ik})$
            $R \leftarrow$ UPDATENONDOMINATEDSET$(r^{'}, R)$
        end
    end
    CONTINUE $\leftarrow$ CHECKCONTINUE()
end
$r^* \leftarrow$ UPDATEBEST$(R)$
```

Since all the sub-problems are independent from each other, our decomposition based approach allows further parallelization for the sub-problems in the system implementation, which would largely speed up the solution approach. Note, in the experiment section, sub-problems are run sequentially, without parallelization.

Solving the Lagrangian Dual

After dual decomposition, we can iteratively update the Lagrangian multipliers and solve the Lagrangian dual problem through projected sub-gradient descent algorithm (Fisher 1981). At each iteration, sub-problems are solved given the Lagrangian multipliers λ_t. Lagrangian multipliers are updated through the master function:

$$\lambda^d_{ik,t+1} := \lambda^d_{ik,t} + \alpha_t(y^d_{ik} - \textstyle\sum_{j \in N} x^d_{ijk}) \quad \forall i, k, d.$$

where the Lagrangian multiplier is increased if a request assigned is not fulfilled in the routing. The duality gap is evaluated and the program is stopped if the termination condition is met. We adopt a commonly used adaptive step function:

$$\alpha_t = \mu_t \cdot \frac{primal^* - dual(\lambda_t)}{\sum_{i \in N_t} \sum_{k \in K} \sum_{m=1}^{M_k} \left(y^d_{ik} - \sum_{j \in N} x^d_{ijk}\right)^2_{w.r.t.\ \lambda_t}}$$

where $dual(\lambda_t)$ is the summation of all sub-problem objectives and $primal^*$ is the upper bound for the dual problem, which is usually obtained by applying a heuristic to the primal problem(P). In this case, we use the best primal solution obtained so far, as the upper bound for the dual problem.

In order to iteratively move towards the optimal solution and determine the convergence, we need the best primal solution with the dual solutions at each iteration. However, dual solutions may not always result in a feasible primal solution, as a request may appear in several routing solutions. Let $z^d_{ik} = \sum_{j \in N} x^d_{ijk}$ where x^d_{ijk} are the decisions from routing sub-problems. To restore a good feasible *primal solution* from the routing solutions, we solve the following ILP:

$$\min - \sum_{i \in N_t} \sum_{d \in D} \sum_{k \in K} r_{ik} \cdot y^d_{ik} + \gamma \cdot \sum_{d \in D} \sum_{k \in K} p_{kd}, \qquad (19)$$

$$s.t.\ Constraints\ (2) - (6)$$

$$y^d_{ik} \leqslant z^d_{ik} \qquad \forall i \in N_t; k \in K; d \in D. \qquad (20)$$

To improve the scalability of primal extraction, we also developed a *greedy local search* heuristic. Given routing solutions as the base, we try to greedily resolve the conflicts of same requests appearing in several routes and insert unassigned requests using local search heuristic (several local search operations are used, such as insert, replace and etc).

Handling Duration Uncertainty

To solve the chance constrained model of Table (3), we apply Sample Average Approximation (SAA) (Pagnoncelli, Ahmed, and Shapiro 2009) to reduce realization set to manageable size and convert the stochastic formulation into a deterministic one. We randomly generate a set of independent and identically distributed samples, $S = \{\xi_1, \xi_2, ..., \xi_s\}$, for all t_{ij} and a_i from the known distributions, and check whether time window and time budget constraints are satisfied. Note, these duration distributions can be derived based on the domain knowledge and historical data. We approximate the probabilities as:

$$P\left(o_{id} \leqslant e^d_{ik} \leqslant c_{id}\right) \approx |S^+|/|S|$$
$$P\left(e^d_{ik} \leqslant T^2_{kd}\right) \approx |S^+|/|S| \qquad (21)$$

where $|S^+|$ are the number of samples under which the corresponding constraints are satisfied. Note that, when approximating chance constraints on a discrete set of samples, it is important to identify a smaller risk threshold $\alpha^{'}$, where $\alpha^{'} < \alpha$. Since SAA replaces the original distributions with empirical distributions obtained from the samples, a smaller $\alpha^{'}$ is used to hedge against the under-representation of the limited samples.

The resulting formulation is an ILP, that still maximizes the total collected rewards but also considers the constraints over all the samples (the detailed formulation is omitted here, due to the space limit). Similarly, the chance constrained criteria can be incorporated into our solution approach by modifying the routing sub-problems, i.e., specifically the GETFEASIBLEREQUEST() Function in our specialized search routine. Now, feasible requests are generated by checking the chance constraints over all the duration samples (against $\alpha^{'}$), instead of just the deterministic durations.

Sample selection heuristic: The scalability and quality of the SAA method depend on the size and representativeness of the sample set. So instead of use a fixed large sample set S', we use a small representative subset $S \in S'$, where we try to select samples that are as different as possible. Intuitively, the smaller the durations, the lesser chance of constraint violation. We first generate a large amount of samples from the duration distributions, say $|S'| = 1000$.

We then sort the samples according to certain metric, i.e., $d_s = \sum_{i \in N} \sum_{j \in N} t_{ij} + \sum_{i \in N_t} a_i$. We then uniformly select $|S|$ samples from $|S'|$ based on the distances d_s.

Experiments

In this section, we empirically demonstrate the efficiency and effectiveness of our approaches on instances adapted from a real world dataset.

Instance Generation: The problem instances considered in our experiments are adapted from a real-world dataset from a leading home health care and hospice company in Pittsburgh, USA. The data contains one month of visit-related information for September 2015. Each service provider is characterized by start geo-coordinates and his qualifications. Also, each patient is associated with a home geo-coordinates and provider preferences. A patient may have several visit records over the week. Each visit record contains information such as a visit datetime, actual service duration, type, discipline, and provider assigned. As the dataset contains the actual visits information and not the input requests for the scheduling, we need to generate the requests from the dataset.

To broaden the analysis, we also synthetically generate an additional set of problem instances with time windows, inter-visit dependencies and duration uncertainties[1]. We first retrieve the visits within the specified scheduling horizon and the subregions. From the selected visits, we retrieve the corresponding patients' and providers' information. Deterministic pair-wise Haversine distances (i.e., great-circle distances) are computed based on their actual geo-information and converted into travel times a priori (with a travel speed of 30mph). We then synthetically generate some additional patients' request-related information, such as available time windows and inter-visit dependencies. We assume visit i's start service time s_i is the mid point for request i's time window. Two types of request time windows are then specified as $[s_i - tw/2, s_i + tw/2]$, with time-window width tw set to 2 hours and 6 hours, denoted as *ins-tight* and *ins-loose* respectively. 2-hour time-window is rather realistic while 6-hour time-window provides more scheduling flexibility. To reflect the inter-visit temporal dependencies, we randomly select 40% of the patients and filter out the patients with only one request. For each selected patient, we then synthetically generate $[D^-_{ij}, D^+_{ij}]$ for all his request pairs $(i, j) \in R_k$. Additionally, we allow those selected requests to have several available days, while the rest of the requests have to be visited on the same days as the actual visits. Lastly, we assume providers are available on the days of the visits, and they work from 7am to 5pm. Provider-dependent rewards are generated based on r and r^+. Here, we assume the base reward r is a fixed value of 100.

For stochastic instances, in lieu of having distributions based on historical data, we assume durations $t \in T$ (both travel and service times) are normally distributed $N\left(\mu_t, \sigma_t^2\right)$ with μ_t equal deterministic durations and σ_t as 10 minutes.

Algorithms Compared: The performances and runtime of our proposed methods depend on how sub-problems are solved. To investigate the trade-off between the optimality and the time performance, in the experiment section, we evaluate the following four algorithms. The routing sub-problems are solved by the search algorithm (1). These algorithms differ from each other on how the assignment sub-problem is solved and how the primal solution is extracted at each iteration. More specifically, we have:

- DLR-E: Both assignment and primal extraction are solved exactly by ILP formulation.
- DLR-H: This is the relaxed version of DLR-E, that solves the assignment sub-problem by linear relaxation and primal extraction by the greedy local search heuristics.
- SLR-E/ SLR-H: Each extends DLR-E and DLR-H respectively to handle duration uncertainties, by applying SAA in the routing search procedure.

In all the experiments, the route penalty γ is set to 80. The cut-off running time for all approaches is set to **10 minutes**, if not specified in experiments. Experiments were conducted on a machine with i7-4790 CPU@3.6GHz and 32 GB RAM.

Experimental Results

We first compare the scalability of exact ILP with DLR. After running CPLEX for exact ILP on small instances (e.g., (D, R, K)=(7, 359, 44)) for 2 hours, it turns out that optimality cannot be reached and running out of memory. Our approaches can return good solutions within 10 minutes even for large-scale instances up to size (D, R, K)=(7, 4203, 273).

Results on Deterministic HHCSP: We first compare the quality of the schedules produced by our LR-based approaches against the actual schedules derived from company-provided visit data. This instance is of size (D, R, K)=(7, 2062,199), where all the requests have fixed visiting days and no time windows. To compare the trade-off between generating more *valid* routes and assigning more *primary* providers, we test the instance with different $r^+ \in \{100, 50\}$. Results are summarized in Table (4). Metrics are measured in percentage of relative difference, normalized by those of the actual visits. Table (4) shows that LR-based approaches improve the objective value by at least 10%. The number of route metric is measured in terms of route reduction. DLR-E generates *fewer routes* compared to the others, i.e., less *manpower* required from the company. In terms of *workload fairness*, we compare the number of *valid* routes generated by each approach. A route is valid if it meets the minimum and maximum workloads. The more valid routes generated, the better. Again, DLR-E outperforms the rest. For *continuity of care*, we compare the total number of primary providers assigned. Both LR-based approaches substantially assign more primary providers compared to actual visits, especially DLR-H. Results also show that increasing r^+ biases the solutions towards more primary providers assigned, a smaller r^+ tends to generate solutions with fewer routes. This demonstrates the flexibility of our methods on generating solutions for different focuses.

[1] Due to patient privacy concerns, it is not possible for us to make actual problem data provided by the HHC Company publicly available. However, the additional synthetically generated problems are accessible: sites.google.com/site/homehealthcarewebsite/

r^+	Objective		No. of Routes		No. of Valid Routes		Primary Assigned	
	DLR-E	DLR-H	DLR-E	DLR-H	DLR-E	DLR-H	DLR-E	DLR-H
100	20.32%	19.72%	-2.59%	0.00%	17.65%	0.74%	46.12%	47.25%
50	14.79%	12.27%	-13.45%	0.17%	52.21%	-3.68%	37.23%	46.65%

Table 4: Comparison of DLR-E and DLR-H against the actual schedules on one instance of size (D, R, K)=(7, 2062, 199) with different r^+.

Instance	DLR-E		DLR-H		
type	gap%	tPerItr(s)	gap%	nGap%	tPerItr(s)
ins-tight	4.46%	137.01	10.52%	6.03%	53.72
ins-loose	0.27%	132.97	7.71%	3.08%	54.26

Table 5: Comparison on synthetic instances with time-windows, temporal dependencies and (D, R, K)=(7, 2062, 199).

Table (5) describes the key results on the performance comparison between the DLR-E and DLR-H on both *ins-tight* and *ins-loose* instances with time-windows and inter-visit dependencies. Results are averaged over 10 random instances of each instance type. The gap here refers to the gap between the best primal and the best dual solution found. It clearly shows that both DLR-E and DLR-H are able to get provably near-optimal solutions, with optimality gaps less than 4.46% for DLR-E and 10.52% for DLR-H. nGap represents the normalized gap, which is calculated as the percentage difference between the best primal solution found by DLR-H and the best dual solution found by DLR-E (tighter lower bound). We can see that the actual optimality gaps for DLR-H should be smaller than nGap, i.e., less than 6% on both instances. tPerItr represents the runtime per iteration. DLR-E provides better solution quality while DLR-H provides a trade-off between solution quality and runtime, which finds good solutions within a shorter time.

Results on Stochastic Extension: Due to durational uncertainties, there is a probability that a request cannot be served within the time window or that a route may exceed its time budget. In this section, we examine the solution quality based on 1000 random duration realizations. Results are averaged over 10 random instances of each instance type and an instance is evaluated over all the realizations. In the experiments, we set our risk level α as 0.3. The smaller the α' we set, the more conservative we are against the representativeness of the selected samples. Meanwhile, increasing the sample size leads to better approximations for the real distributions, but less efficient. Based on our initial set of parameter experiments (omitted here due to space limit), we set α' as 0.15 and sample size as 60. Samples are generated using our sample selection heuristic.

We focus on two evaluation metrics: 1) Chance constraint violation ratio, the number of chance constraints that are violated, normalized by the total number; 2) Expected objective, the average objective value achieved by the solution. For these results, a chance constraint is considered as violated if more than $\alpha \cdot 1000$ times over the 1000 realizations ($\alpha = 30\%$ here), this constraint is not satisfied. The expected objective is calculated as the sum of provider-dependent rewards of all requests, whose time window chance constraints are not violated, averaged over the random instances.

	Expected Obj	TW Violations	TB Violations
DLR-E: mean	382600	76.7/1987.2	14/508.2
DLR-H: mean	387470	65.9/1991.3	11.6/531
DLR-E: max	319780	0/1597.4	0/512.2
DLR-H: max	340320	0/1708.6	0/551.9
SLR-E	388350	0.2/1938	0 /511
SLR-H	394090	0.1/1960	0 /534

Table 6: Results on synthetic instances of *ins-tight* with (D, R, K)=(7, 2062, 199). *TW* denotes time-window chance constraints, while *TB* refers to the time budget chance constraints.

We compare the stochastic extensions with both deterministic approaches, DLR-E and DLR-H. We evaluate the deterministic approaches with two set of duration input from the distributions: mean and maximum durations. We use the durations that are 2σ upper away from the means, i.e., 97.75% percentile, to approximate the maximum values.

We can observe from Table (6), our deterministic approaches with mean-duration are sensitive to duration uncertainties on *ins-tight* instances, where time window chance constraints and time budget chance constraints can be easily violated. On the other hand, deterministic approaches with max-duration can guarantee services and travel. However, they suffer from worse expected objectives, where providers are underutilized. While both SLR-E and SLR-H, are robust towards the stochasticity, achieving almost no violations of the chance constraints. Stochastic approaches slightly outperform the deterministic counterparts with mean-duration on expected objective values and they substantially outperform the deterministic counterparts with max-duration on this metric. Thus, more sophisticated approaches, i.e., our stochastic approaches, to model uncertainty are warranted.

Conclusion

We investigate the multi-period HHCSP problem driven by the real-world needs. An integer linear programming model is proposed to formulate the deterministic problem. We further extend it with chance constraints to handle stochastic travel and service times, which is useful for real-world situations. Subsequently, to develop a solution that can scale to city-scale scenarios, we apply the Lagrangian relaxation and exploit the separable problem structure to decompose the formulation into smaller sub-problems. Finally, we solve the chance constrained problem by applying sample average approximation. With this sampling based approach incorporated, as long as there is a distribution simulator driven by the historical duration data, we can provide proactive solutions, which react well to potential uncertainties.

References

Bachouch, R. B.; Guinet, A.; and Hajri-Gabouj, S. 2011. A decision-making tool for home health care nurses planning. In *Supply Chain Forum: an International Journal*, volume 12, 14–20. Taylor & Francis.

Barrera, D.; Velasco, N.; and Amaya, C.-A. 2012. A network-based approach to the multi-activity combined timetabling and crew scheduling problem: Workforce scheduling for public health policy implementation. *Computers & Industrial Engineering* 63(4):802–812.

Begur, S. V.; Miller, D. M.; and Weaver, J. R. 1997. An integrated spatial dss for scheduling and routing home-healthcare nurses. *Interfaces* 27(4):35–48.

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

Bowers, J.; Cheyne, H.; Mould, G.; and Page, M. 2015. Continuity of care in community midwifery. *Health care management science* 18(2):195–204.

Burke, E. K.; De Causmaecker, P.; Berghe, G. V.; and Van Landeghem, H. 2004. The state of the art of nurse rostering. *Journal of scheduling* 7(6):441–499.

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

Carello, G., and Lanzarone, E. 2014. A cardinality-constrained robust model for the assignment problem in home care services. *European Journal of Operational Research* 236(2):748–762.

Castillo-Salazar, J. A.; Landa-Silva, D.; and Qu, R. 2016. Workforce scheduling and routing problems: literature survey and computational study. *Annals of Operations Research* 239(1):39–67.

Ernst, A. T.; Jiang, H.; Krishnamoorthy, M.; and Sier, D. 2004. Staff scheduling and rostering: A review of applications, methods and models. *European journal of operational research* 153(1):3–27.

Errarhout, A.; Kharraja, S.; and Corbier, C. 2016. Two-stage stochastic assignment problem in the home health care. *IFAC-PapersOnLine* 49(12):1152–1157.

Eveborn, P.; Flisberg, P.; and Rönnqvist, M. 2006. Laps care—an operational system for staff planning of home care. *European Journal of Operational Research* 171(3):962–976.

Fernandez, A.; Gregory, G.; Hindle, A.; and Lee, A. 1974. A model for community nursing in a rural county. *Journal of the Operational Research Society* 25(2):231–239.

Fikar, C., and Hirsch, P. 2017. Home Health Care Routing and Scheduling: A Review. *Computers & Operations Research* 77:86–95.

Fisher, M. L. 1981. The lagrangian relaxation method for solving integer programming problems. *Management science* 27(1):1–18.

Gunawan, A.; Lau, H. C.; and Vansteenwegen, P. 2016. Orienteering problem: A survey of recent variants, solution approaches and applications. *European Journal of Operational Research.*

Hertz, A., and Lahrichi, N. 2009. A patient assignment algorithm for home care services. *Journal of the Operational Research Society* 60(4):481–495.

Koeleman, P. M.; Bhulai, S.; and van Meersbergen, M. 2012. Optimal patient and personnel scheduling policies for care-at-home service facilities. *European Journal of Operational Research* 219(3):557–563.

Lin, M.; Chin, K. S.; Wang, X.; and Tsui, K. L. 2016. The therapist assignment problem in home healthcare structures. *Expert Systems with Applications* 62:44–62.

National Association for Home Care & Hospice. 2010. Basic statistics about home care. *Washington, DC: National Association for Home Care & Hospice* 1–14.

Nickel, S.; Schröder, M.; and Steeg, J. 2012. Mid-term and short-term planning support for home health care services. *European Journal of Operational Research* 219(3):574–587.

Pagnoncelli, B.; Ahmed, S.; and Shapiro, A. 2009. Sample average approximation method for chance constrained programming: theory and applications. *Journal of optimization theory and applications* 142(2):399–416.

Population Reference Bureau. 2016. Fact Sheet: Aging in the United States. http://www.prb.org/Publications/Media-Guides/2016/aging-unitedstates-fact-sheet.aspx.

Rais, A., and Viana, A. 2011. Operations research in healthcare: a survey. *International transactions in operational research* 18(1):1–31.

Rasmussen, M. S.; Justesen, T.; Dohn, A.; and Larsen, J. 2012. The home care crew scheduling problem: Preference-based visit clustering and temporal dependencies. *European Journal of Operational Research* 219(3):598–610.

Rodriguez, C.; Garaix, T.; Xie, X.; and Augusto, V. 2015. Staff dimensioning in homecare services with uncertain demands. *International Journal of Production Research* 53(24):7396–7410.

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

Toth, P., and Vigo, D. 2014. *Vehicle routing: problems, methods, and applications*. SIAM.

Trautsamwieser, A., and Hirsch, P. 2014. A branch-price-and-cut approach for solving the medium-term home health care planning problem. *Networks* 64(3):143–159.

Yuan, Z., and Fügenschuh, A. 2015. Home health care scheduling: a case study. In *proceedings of the 7th Multidisciplinary International Conference on Scheduling : Theory and Applications (MISTA 2015), 25 - 28 Aug 2015, Prague, Czech Republic*, 555–569.

Yuan, B.; Liu, R.; and 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(24):7450–7464.