Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Clustering-based iterative heuristic framework for a non-emergency patients transportation problem

Jamal Abdul Nasir a, Yong-Hong Kuo a, b, *, Reynold Cheng b, c, d

a Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong
b HKU Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong
c Department of Computer Science, The University of Hong Kong, Hong Kong
d Guangdong-Hong Kong-Macau Joint Laboratory for Smart Cities, Hong Kong

ARTICLE INFO

Keywords:
Shared healthcare mobility
Dial-a-ride problems
Non-emergency patients transportation
Clustering
Vehicle routing

ABSTRACT

Introduction: Non-emergency patient transportation (NEPT) services are particularly important nowadays due to the aging population and contagious disease outbreaks (e.g., Covid-19 and SARS). In this work, we study a NEPT problem with a case study of patient transportation services in Hong Kong. The purpose of this work is to study the discomfort and inconvenience measures (e.g., waiting time and extra ride time) associated with the transportation of non-emergency patients while optimizing the operational costs and utilization of NEPT ambulances.

Methods: A mixed-integer linear programming (MILP) formulation is developed to model the NEPT problem. This MILP model contributes to the existing literature by not only including the patient inconvenience measures in the objective function but also illustrating a better trade-off among different performance measures through its specially customized formulation and real-life characteristics. CPLEX is used to find the optimal solutions for the test instances. To overcome the computational complexity of the problem, a clustering-based iterative heuristic framework is designed to solve problems of practical sizes. The proposed framework distinctively exploits the problem-specific structure of the considered NEPT problem in a novel way to enhance and improve the clustering mechanism by repeatedly updating cluster centers.

Results: The computational experiments on 19 realistic problem instances show the effective execution of the solution method and demonstrate the applicability of our approach. Our heuristic framework observes an optimality gap of less than 5% for all those instances where CPLEX delivered the result. The weighted objective function of the proposed model supports the analysis of different performance measures by setting different preferences for these measures. An extensive sensitivity analysis performed to observe the behavior of the MILP model shows that when operating costs are given a weightage of 0.05 in the objective function, the penalty value for user inconvenience measures is the lowest; when the weightage value for operating costs varies between 0.8 and 1.0, the penalty value for the same measures is the highest.

Conclusions: This research can assist decision-makers in improving service quality by balancing operational costs and patient discomfort during transportation.

* Corresponding author. Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong.
E-mail address: yhkuo@hku.hk (Y.-H. Kuo).

https://doi.org/10.1016/j.jth.2022.101411
Received 6 September 2020; Received in revised form 9 June 2022; Accepted 10 June 2022
Available online 5 July 2022
2214-1405/© 2022 Elsevier Ltd. All rights reserved.
1. Introduction

Patient transportation facilities to reach hospitals/clinics are an integral aspect of any health care system and are necessary for the convenience of the elderly and citizens with disabilities. Bellamy et al. (2003) describe the non-emergency patient transportation (NEPT) requirements and related situations in which patients are unable to access or leave the health care facilities themselves. Almost 3.6 million elderly American residents are unable to receive health care because of inadequate availability of transport services (Wallace et al., 2005). Hence, access to health care facilities and associated transportation services is critical for sustaining the highest rank of fitness and also beneficial for the middle-aged population and senior citizens. NEPT services are also critical in mitigating the risk of contagious disease outbreaks, such as Covid-19 (Li et al., 2020; Choi, 2020) and SARS (Ksiazek et al., 2003; Peiris et al., 2003). Some NEPT ambulances have configurable vehicle capacity, facilitating social distancing. To ensure patient safety during transportation, the partition of vehicles, usage of personal protective equipment, and decontamination of vehicles can be carried out effectively in NEPT ambulances. Such services provide an alternative to public transportation systems for accessing healthcare facilities, thereby reducing the risk of spreading the diseases in the community.

Health care establishments arrange the necessary ambulances and staff to take elderly people or patients with disabilities from home to hospitals. Patients’ transportation requests and needs vary across geographic locations, different clinics, and in relation to their different types of impairment. Some patients who cannot independently undertake medical treatments may be escorted by some attendants or family members. Therefore, additional space would be required to support such transportation requests. Besides, some of those patients using a wheelchair or a stretcher may require specialized support throughout their transportation. To satisfy the patients’ transportation requests, the planning staff takes into account the following issues simultaneously: (i) deployment of enough ambulances to meet the demand considering the available fleet of non-emergency ambulances; and (ii) developing a routing plan for each ambulance under different constraints and restrictions. This planning exercise usually aims to address the following issues: (i) maximizing the number of patients’ transportation requests served under the given resources and constraints; (ii) minimizing the daily operational expenses to run these ambulances; and (iii) minimizing the penalty expenses owing to the poor service quality or user inconvenience issues. To replace manual planning with better computerized scheduling and routing plans, efficient and effective solution methods are crucial. The problem discussed in this paper explores the non-emergency ambulance transfer service (NEATS) in Hong Kong. NEATS operates as a part of the Health Department in Hong Kong and offers transportation services to three categories of patients: (i) elderly patients who need to attend appointments at hospitals/clinics; (ii) patients requiring transfers among hospitals; and (iii) the patients who require return transports while leaving the hospital. Taking into account the non-emergency characteristics of these services, patients’ transportation requests are typically submitted prior to the given appointment dates. Multiple patients can share each ambulance route. Thus, group transfers are desired to improve utilization and operational cost efficiency. Hence, rather than a first-come, first-served basis, a patient’s tour is arranged with other patients going to the same destination or nearby location. While this approach increases the utilization rate of ambulances and, in some cases, also helps curtail transportation expenses, waiting times for patients increase significantly. Consequently, there is growing concern about how NEPT waiting times negatively impact service quality. The increase in the actual commute time and the higher gap between the actual and ideal travel times decrease the satisfaction level of commuters (Humagain and Singleton, 2020). To overcome these challenges and to develop a balanced approach for the NEPT, we explore the trade-off between patient inconvenience measures (waiting time and extra ride time) and operating costs. The impact of changes to these performance measures (operating costs, user inconvenience measures) on utilization rate is examined, and a dial-a-ride problem (DARP) based MILP formulation is proposed. The competing priorities need to be addressed in the context of economic viability and service quality for patient transportation services. Thus, the objective function of the proposed mathematical model seeks to minimize the weighted sum of four elements that represent different performance metrics considering non-urgent patient transportation service: (i) the total travelling cost by all the ambulances and related ambulance route assignment costs; (ii) the total patients’ waiting time at the pick-up locations; (iii) the total extra ride time observed by the patients; and (iv) the total underutilized capacity of ambulances concerning the toured nodes. Moreover, an enhanced K-means based iterative heuristic (EKMIH) framework is developed to solve medium and large problem instances by implementing a cluster-first, route-second approach. Patients with similar transportation needs are grouped together in this study using the clustering method.

This research makes a scientific contribution by attempting to answer the following research questions: (1) Is there any trade-off between user inconvenience measures and operating costs in the context of the NEPT problem under consideration?, (2) If there is a trade-off, how do these performance measures respond to changes in the values of some specific parameters and when different preferences are assigned to them via the objective function of the proposed MILP model? To accomplish this research, we incorporate several original features in our MILP model, solution framework and computational experiments.

The main contributions of our realistic NEPT features based mathematical model and heuristic framework to the existing NEPT literature are as follows: 1) Compared to other studies that usually minimize the travelling distance/cost, the objective function of our formulation minimizes the user inconvenience measures (the travelling distance/cost and ambulance route assignment cost), underutilized capacity and user inconvenience measures (e.g., waiting time and extra ride time). Hence, the weighted objective function assists in determining the better trade-off between operating costs and user inconvenience measures; 2) Taking into account the discomfort of elderly

---

1 While we use people-first disability language in this paper, we recognize that there is ongoing debate about using this language versus identity-first disability language. For more on this language debate, see Titchkosky (2001) and Ross (2013).

2 Excess travel time for kids with disabilities is concerning since it can result in missing class time and hinder peer contact (Buliung et al., 2021). Hence extra ride time can result in both discomfort and a missed appointment.
patients associated with the waiting for the ambulance and travelling time, our model incorporates the patient availability start time as a hard constraint and enforces penalties for the waiting time beyond this availability start time and extra ride time between origin-destination pair; 3) Most of the existing cluster-first routing-second approaches are straightforward, but the approach utilized in this work attempts to exploit the characteristics of the studied problem in a novel way. The heuristic framework, EKMIH, enhances the performance of the K-means algorithm by introducing priority rules to select initial centroids. Then the resulting centroids after each convergence of the K-means algorithm are updated iteratively by taking into account the problem-specific characteristics (filter out distant members, shared destination-based insertion, etc.) to filter the clusters and insert new patients in a cluster; 4) Considering the case of NEPT services in Hong Kong, the MILP model and the heuristic framework are employed to demonstrate the application of our approach. Computational experiments are conducted to examine the impacts of various cluster sizes and the number of clusters on the cost components of the objective function; 5) This study conducts a sensitivity analysis to observe the potential trade-off between the operating costs and user inconvenience measures. We analyze different performance measures by giving preference to these measures through the weighted objective function.

This paper is an extended work of our research presented at the International Conference on Health Care Systems Engineering (HCSE), Montréal, Canada, May 30 - June 1, 2019 (Nasir et al., 2020). The new work expands the prior work (Nasir et al., 2020) in three major dimensions. To begin with, because of the importance of travel time in determining the quality of a transportation service, the NEPT problem has been expanded to include the ride time and extra ride time. To integrate these changes, the objective function and constraints of MILP formulation are enhanced, and additional decision variables are added to the model. Furthermore, we exclusively used CPLEX to solve the MILP model in our earlier work (Nasir et al., 2020), while our current work proposes a brand-new element, the EKMIH framework-based heuristic approach, to solve the large-sized instances of the NEPT problem. Lastly, the completely new computational experiments in our current work involve a larger realistic data set and a greater number of instances. These new experiments also study the effect of variation in the number of clusters and size of clusters on the NEPT performance measures. Above all else, compared to the prior work, the thorough sensitive analysis conducted in the current work helps to analyze the behavior of the MILP model against varying preferences given to performance measures. This analysis also helps to develop managerial insights.

The remainder of this article is organized as follows. In Section 2, we review the relevant literature. In Section 3, we describe the mathematical model for our problem. Section 4 presents the solution methodology. Section 5 reports the computational experiments and results. Section 6 provides the conclusion and future work.

2. Literature review

The related literature considering the health care for non-emergency patients is presented with respect to three aspects: (1) home-based health care for non-emergency patients; (2) availability of non-emergency medical transport for patients; (3) mathematical models to solve non-emergency patients’ transportation problems. In the context of home healthcare, healthcare staff visits patients in their homes to provide necessary care sessions to the patients or elderly in a friendly home environment. Hence assignment, scheduling, and routing decisions for the home healthcare staff have also been examined (Lanzarone et al., 2012; Yal and et al., 2014). Eveborn et al. (2006) investigated a home health care staff scheduling problem in the Swedish home healthcare system. To make schedules for the visiting staff, a set partitioning model was formulated, whereas repeated matching was employed to solve the problem. Gutierrez et al. (2009) studied a combined staff scheduling and routing problem to provide home care services. The authors formulated a MILP model to solve the problem and demonstrated the high complexity of the computational problem. Nasir and Dang (2020) assessed the effectiveness of quantitative thresholds based decisions for the selection of home health care patients and care staff. The authors explore how the solution of the mathematical model and logistic regression-based approaches can be utilized to define decision rules for selection decisions in the future. Taking into account the continuity of service, dynamic patient arrivals and fast patient acceptance decisions, a home health care routing and scheduling problem was investigated by Demirbilek et al. (2019). The authors seek to maximize the daily visits of nurses and proposed a scenario-based heuristic approach to determine the assignment, scheduling, and patient acceptance decisions. The care sessions for non-emergency patients through home healthcare-based scheduling and routing models have been studied by different researchers (Rasmussen et al., 2012; Shao et al., 2012; Lanzarone and Matta, 2012; Rest and Hirsch, 2016; Nasir and Dang, 2018).

The availability and accessibility to healthcare facilities are fundamental aspects of healthcare provision. Minimizing transportation times and enhancing the supply of medical facilities are the two major approaches to strengthen the availability of healthcare services for elderly people and decrease the inequalities (Chen et al., 2020). Boisjoly et al. (2020) investigated the accessibility to general medical and surgical treatment hospitals by public transport across eight metropolitan areas in Canada. The authors observed that large urban areas underachieve considering the equitable distribution of facilities and average approachability. Wallace et al. (2005) investigated the availability of non-emergency medical transport (NEMT) to American citizens. The authors indicated that lack of access to NEMT is a barrier to quality healthcare and it disproportionately affects the segments of the society associated with females, poorer, elderly, less qualified citizens and members of minority groups. They emphasized the importance of improving access to NEMT in order to increase the quality of life while also benefiting society. Smith et al. (2017) compared the patterns of service usage and related costs of NEMT for the middle-aged population and elderly in rural versus urban areas in Delaware, USA. The authors analyzed the data of 39,194 patients and concluded that rural trips cover longer distances and incur higher costs; 50% of the visits were for dialysis, thus highlighting the need for a better dialysis facility in the vicinity. Chaiyachat et al. (2018) evaluated the effect of the availability of ride share-based transportation for the patients who are insured by Medicaid in West Philadelphia, USA. The results obtained from this pilot study suggested that the availability of rideshare-based transportation may give rise to the show rates for
primary care sessions.

Mathematical models to solve NEPT problems have been studied extensively in the literature. The DARP serves as a typical model to formulate the point-to-point non-emergency patients’ transportation. Although DARP shows similarities to the vehicle routing problem (VRP) with pickups and deliveries, from the modeling perspective, a clear difference can be established between the DARP and other routing problems when considering that the patient discomfort must be managed against minimizing the operational costs (Cordeau and Laporte, 2007). Our study presents the related DARP literature considering the static or dynamic nature of patient requests, the characteristics of the objective function and constraints, and solution methods. An outstanding overview of DARP classification, variants, and available solution methods can be found in Cordeau and Laporte (2007), Ho et al. (2018), and Parragh (2011). Separating the transport requests arrivals into static and dynamic categories, Cordeau and Laporte (2007) defined two different models for DARPs. The static category (Parragh, 2011; Lim et al., 2017; Zhang et al., 2015; Ritzinger et al., 2016) only covers the requests for transport received before the specified time restriction, while the dynamic case (Souza et al., 2020; Van Den Berg and Van Essen, 2019; Kergosien et al., 2011, 2014) accepts requests over the day and adjusts the scheduling plans correspondingly.

From the perspective of dynamic arrivals over the day, Kergosien et al. (2014) addressed both emergency patient requests and NEPT services. The authors employed a discrete event simulation-based method to study the efficiency and effectiveness of the proposed ambulance fleet management approaches. Their research examined three management approaches. It was noted that the proactive, integrated approach to fleet management produced more positive outcomes and was easier to incorporate into real-life scenarios. To meet the dynamic demands for patient transportation between the clinics, Kergosien et al. (2011) developed a Tabu Search (TS) based heuristic algorithm to find solutions. The capacity of every vehicle was restricted to one patient only. Subcontracting could be used to meet unserved requests. The authors incorporated three types of patient transport requests and proposed a solution approach that keeps the routes using the adaptive memory, and thereafter the initial solutions are enhanced by repeatedly implementing the TS algorithm. Considering the stochastic information on potential return transports, Schilde et al. (2011) studied the dynamic DARP in the setting of the Austrian Red Cross’s regular operations. Four different metaheuristic methods were tested to obtain the solutions. It was found that their look-ahead approach to the potential transportation of patients tends to raise the quality of the solution if the number of patients’ transportation requests needing the return transportation service is small compared to the total requests for transport. Van Den Berg and Van Essen (2019) investigated a procedure for optimizing NEPT routes by using a portion of the capacity while maximizing the unused capacity to accommodate emergency patients’ transportation requests. Two approaches were proposed to solve their problem. An integer linear programming formulation was solved with the enforcement of a time restriction in the first approach, while the second approach proposed a substitute formulation under discretized time characteristics to attain efficient solutions. The dynamic transfer requests are handled through an online model throughout the day.

Considering the static mode of DARP, a multi-trip ambulance routes and manpower planning issues-based NEPT problem was examined by Lim et al. (2017). The authors considered a hierarchical objective function in their mathematical model and proposed a metaheuristic that involves a local search-based mechanism to find the solutions. Zhang et al. (2015) presented a mathematical model for the NEPT problem that aims to minimize the unserved patient requests and travelling costs. They also developed a memetic algorithm to solve the studied problem. Diana and Dessouky (2004) proposed a parallel regret insertion heuristic-based solution method to solve the DARP with time windows for the provision of transportation services to the elderly and citizens with disabilities. Their solution method initializes the routes by taking into account both spatial and temporal constraints, whereas a regret insertion heuristic is then implemented to schedule the pending requests. The authors adopted a flexible approach to adjust dynamic requests if needed. Oliveira et al. (2015) modeled the NEPT problem as a team orienteering problem, which enables patient selection and route assignment against the objective of cost minimization. The authors proposed a MILP formulation and an iterative heuristic solution algorithm to obtain the solutions. A NEPT problem considering the manpower allocation and routing of ambulances was investigated by Zhang et al. (2017). The authors formulated the problem as an integer programming model and proposed variable neighborhood search (VNS) based algorithms to solve the problem. The mathematical model developed in our study follows the static mode. However, in contrast to the studies mentioned above, we use a number of service quality-oriented characteristics in our objective

### Table 1

**Characteristics of the related NEPT literature.**

| Article                  | Objective Constraints | Constraints |
|-------------------------|-----------------------|-------------|
|                         | RC/TT     | UU | WT | ERT | USR | TW | VL | UU | WT | ERT | MRT | MRL/MTT | CBHS | CS |
| Lim et al. (2017)       | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | -       | X    |   |
| Zhang et al. (2015)     | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | -       | X    |   |
| Ritzinger et al. (2016) | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | -       | X    |   |
| Schilde et al. (2011)   | X         | X  | -  | X   | X   | X  | X  | -  | -  | -   | X   | -       | X    |   |
| Parragh et al. (2009)   | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | -       | -    |   |
| Parragh (2011)          | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | X       | -    |   |
| Qu and Bard (2013)      | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | -       | X    |   |
| Oliveira et al. (2015)  | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | -       | X    |   |
| Souza et al. (2020)     | X         | -  | -  | X   | X   | X  | X  | -  | -  | -   | X   | X       | -    |   |
| This study              | X         | X  | X  | X   | X   | X  | X  | -  | -  | -   | X   | X       | X    | X |

| RC, Routing cost; TT, Travel time; UU, Underutilization; WT, Waiting time; ERT, Extra ride time; USR, Unserved requests. |
| TW, Time windows; VL, Vehicle load; MRT, Maximum ride time; MRL, Maximum route length; MRT, Maximum trip time; CBHS, Clustering based heuristic solution; CS, Case study. |
function and provide a tradeoff analysis through several computational experiments against the considered performance measures.

The issue of transferring non-emergency patients or the elderly has been also explored in the literature considering the routing costs and service quality characteristics in the objective function. Most of the existing research revolves around minimizing the routing costs. Only a few researchers have considered patients' inconvenience in terms of extra ride (travel) time in the objective function, whereas, the waiting time at the patient pickup location has been rarely included as an inconvenience measure in any study. To the best of our knowledge, our study is the first attempt that extends NEPT problem literature by including both extra ride time and pickup location waiting time in the objective function. Moreover, we also cover the underutilization penalty as a performance measure in the objective function. We compare our study against the most relevant literature in Table 1. Compared to our research where customer discomfort is the primary focus, Lim et al. (2017) and Zhang et al. (2015) focused more on the routing costs, workload balance among staff and minimization of the number of unserved requests. Ritzinger et al. (2016) examined a DARP where traveler discomfort was taken into consideration in transporting aged and patients with disabilities. The authors employed exact and heuristic algorithms based on dynamic programming (DP) to find solutions for their problem. Ritzinger et al. (2016) only addressed patient inconvenience through a hard constraint that limits patient ride time beyond a certain value. To find the efficient dispatch plans for the ambulance service of the Austrian Red Cross, Parragh et al. (2009) proposed a two-phase heuristic solution method for a multi-objective DARP. The proposed mathematical formulation included patient discomfort and cost of travel in the objective function. Pareto optimal solutions were found by engaging a path relinking module, whereas VNS was used to deliver the initial solutions. Parragh (2011) explored a DARP with heterogeneous vehicles and clients. The author considered various modes of use criteria (seats, wheelchair and stretcher) and suggested two mathematical model formulations. Considering the patient discomfort, the multi-objective approach suggested by Parragh et al. (2009) attempted to minimize the routing cost and average ride time of patients, whereas Parragh (2011) also aimed to minimize the extra ride time of the ambulance by penalizing it in the objective function. However, both of these studies were based on artificial data sets and did not provide any trade-off analysis for the real settings. The optimization of non-emergency patient transport operations was studied by Oberscheider and Hirsch (2016) for three different service levels, namely time windows, ride time and exclusive transports. Statistical analysis based on non-static service times and a matheuristic solution approach involving the TS algorithm were used in their study. To meet the specific patient demands considering capacity adjustment with the reconfigurable ambulances, Qu and Bard (2013) introduced a mixed-integer model with side constraints, and the problem was solved through a two-phase heuristic solution method. Qu and Bard (2013) also imposed penalty in the objective function to minimize the extra ride times, but the main focus of their study was to investigate the configurable capacity option for the pickup and delivery problem. Luo et al. (2019) investigated a practical DARP variant and developed a trip-based mathematical formulation. The authors proposed a two-phase branch-and-price-and-cut algorithm to get the exact solution for the studied problem, and maximum ride time constraint was used to avoid extra travel times. Schilde et al. (2011) considered patients’ discomfort in the objective function through a penalty for waiting time (tardiness), and maximum ride time is limited through a hard constraint. Recent work by Souza et al. (2020) contains a single term user inconvenience measure in the objective function. Oliveira et al. (2015) also considered a maximum ride time constraint for their team orienting-based NEPT problem model. It is pertinent to mention that in comparison to our work, most of the studies mentioned above rarely included extra ride time cost or waiting time at pick up location in the objective function, as evident from the Table 1. The constraints in our mathematical formulation define the values for extra ride time, waiting time and underutilization. Additionally, most studies only addressed patient inconvenience by introducing a maximum ride time constraint in the model.

None of the studies listed in Table 1 used clustering-based heuristic approaches. Since computational complexity is very high for DARPs, both exact methods and heuristic or metaheuristic approaches have been used to solve the DARP and its variants. The most popular exact methods used to solve DARPs include branch-and-cut and branch-and-price algorithms. Cordeau (2006), proposed the first branch-and-cut algorithm to solve the three-index MILP formulation. He also introduced families of valid inequalities to further strengthen the mathematical model. The studies that employ branch-and-cut algorithm based solutions for DARPs with different features include a two-index mixed-integer programming model with families of valid inequalities (Ropke et al., 2007), heterogeneous clients and vehicles (Parragh, 2011), heterogeneous vehicles and multi-depot (Brakens et al., 2014), multiple trips, different request types, and configurable vehicle capacity (Liu et al., 2015). On the other hand, a branch-and-price algorithm with the aim to maximize the passenger occupancy rate (Garaix et al., 2011), branch-and-price-and-cut algorithm considering maximum riding constraints (Gschwind and Irnich, 2015) and two-phase branch-and-price-and-cut algorithm by taking into account manpower planning and staff lunch breaks (Luo et al., 2019) have been proposed in the literature. Considering the application of metaheuristic methods, TS (Cordeau and Laporte, 2003; Kirchler and Calvo, 2013), VNS-based approaches (Parragh et al., 2010; Zhang et al., 2017), large neighborhood search (Qu and Bard, 2013; Brakens and Kovacs, 2016) and genetic algorithm-based methods (Masmoudi et al., 2017; Cubillos et al., 2009) were used to solve DARPs. From our literature review, clustering-based iterative heuristics approaches have not been used in the literature to solve the NEPT problem.

Some studies have applied clustering-based heuristic methods but in areas other than NEPT. Clustering algorithms have also been adopted to solve VRPs and DARPs. Ferrandez et al. (2016) used K-means clustering to determine the locations for a truck-drone network to deliver packages to customers. A genetic algorithm-based heuristic procedure was employed to find the truck route among the launching locations. Geetha et al. (2012) studied a multi-depot VRP in a setting of distribution logistics. A cluster first, route second methodology was adapted along with metaheuristics to solve the problem. Czioska et al. (2019) investigated the DARP in the context of shared demand-responsive transportation services. The authors grouped the customers by using a clustering method and calculated the meeting points for each cluster. A neighborhood search-based heuristic approach was used to create the routing plan for vehicles. Cinar et al. (2016) developed a two-phase heuristic method that utilizes a K-means clustering algorithm to enhance the computational performance of a cumulative VRP with limited duration. Our clustering-based iterative heuristic approach is different from the above-mentioned studies because it effectively exploits the characteristics of the studied NEPT problem to determine the
proper clusters for patients that will share the same ambulance. Similarly, the MILP model in our research is more focused on the patient inconvenience measures and assists in evaluating different performance indicators through a weighted objective function. Consequently, the consideration of practical characteristics in this work helps decision-makers to identify better and more reasonable transportation plans in real-life situations.

3. Problem description

This section presents the mathematical model for the studied NEPT problem. The MILP model includes the parameters (time windows, vehicle capacity, maximum route length, distance between locations and penalty values) and decision variables. The solution to this model will deliver the scheduling and routing decisions for the ambulances as well as the values for the decision variables while minimizing the total cost. The patient’s discomfort (extra ride time and waiting time) and vehicle underutilization-related variables contribute the most to making our formulation different from the existing literature. All these parameters and decision variables are mathematically represented in this section in order to use them in the formulation.

In the NEPT problem we study, each patient’s transportation request is marked by a pickup node and a respective delivery node. Let \( P = \{1, \ldots, n\} \) be the set of pickup nodes and \( D = \{n+1, \ldots, 2n\} \) be the set of patient delivery nodes, where \( n + j \) is the delivery node of the corresponding pickup node \( j \). Let \( 0 \) and \( 2n + 1 \) denote the origin node and depot (final destination node), respectively. Hence, the route for each ambulance will begin at the origin node and end at the depot. In this work, each route consists of a single trip for each ambulance. Let \( N_r = P \cup D \) be the set of all the nodes to be travelled by an ambulance after leaving the origin node. \( N^d = \{0\} \cup N_r \) denotes the set of the origin node and all the pickup and delivery nodes, whereas \( N^d = N^r \cup (2n + 1) \) represents the set of all the pickup and delivery nodes and the depot. The NEPT problem is defined on a graph \( G = (N, A) \), where \( N = \{0\} \cup P \cup D \cup (2n + 1) \) is the set of all nodes and \( A = \{(j, k) | j \in N, k \in N\} \) is the set of arcs. \( \alpha_1, \alpha_2, \alpha_3 \) and \( \alpha_4 \) are the allocated weights for corresponding performance measures.

The weighted objective function (1) consists of five terms and minimizes the total cost. The first two terms represent the operating cost, the third term shows the underutilization penalty, and the last two terms express the penalty for the user inconvenience (waiting time and extra ride time, respectively). Table 2 presents a summary of associated parameters and decision variables used in the mathematical formulation of the model.

### 3.1. Mathematical model

\[
\begin{align*}
\text{min } & \quad \alpha_1 \left( b \sum_{i \in V} \sum_{(j,k) \in A} d_{ik} x_{ijk} + c \sum_{i \in V} z_i \right) + \alpha_2 \left( k \sum_{i \in V} \sum_{j \in N^r} u_{ij} \right) \\
\end{align*}
\]

| Table 2 Notations. |
|--------------------|
| **Indices** |
| \( P \) & Set of patient pickup nodes |
| \( D \) & Set of patient delivery nodes |
| \( V \) & Set of non-emergency ambulances |
| \( N \) & Set of all nodes |
| **Parameters** |
| \( G \) & Total capacity of each ambulance \( i \in V \) |
| \( L \) & Route length limit for each ambulance \( i \in V \) |
| \( M \) & A very large number |
| \( T \) & Time duration for each stop for pickup/delivery service |
| \( t_{jk} \) & Travelling time needed to reach from node \( j \) to \( k \) |
| \( d_{jk} \) & Distance between two locations \( j \) and \( k \) |
| \( p_{jk} \) & Total passengers to be picked at pickup node \( j \in P \) |
| \( d_{ij} \) & Total passengers to be dropped at delivery node \( j \in D \) |
| \( E_{ij} \) & Patient availability start time for pickup at pickup node \( j \in P \) |
| \( b \) & Cost for travelling per unit of distance |
| \( c \) & Cost for route allocation to an ambulance |
| \( f \) & Penalty cost for waiting per minute at pickup nodes |
| \( g \) & Penalty cost for underutilization of ambulance at each visited node |
| \( h \) & Penalty cost for extra ride time per minute in an ambulance |
| **Decision Variables** |
| \( x_{ijk} \) & 1 if ambulance \( i \in V \) moves from node \( j \) to \( k \), 0 otherwise |
| \( z_i \) & 1 if ambulance \( i \in V \) has been allocated a route, 0 otherwise |
| \( q_{ij} \) & Start time of ambulance visit \( i \in V \) at node \( j \in N \) |
| \( V_{ij} \) & Vehicle load for ambulance \( i \in V \) at node \( j \in N \) |
| \( w_{ij} \) & Waiting time for ambulance \( i \in V \) at pickup node \( j \in P \) |
| \( u_{ij} \) & Underutilization of ambulance \( i \in V \) at node \( j \in N \) |
| \( r_{ij} \) & Ride time observed in ambulance \( i \in V \) for the patient at pickup node \( j \in N \) beyond the direct travel time \( t_{jk} \) |
| \( e_{ij} \) & Extra ride time in ambulance \( i \in V \) for the patient at pickup node \( j \in N \) beyond the direct travel time \( t_{jk} \) |
\[ + \alpha \left( \sum_{i \in V} \sum_{j \in P} w_{ij} \right) + \alpha_i \left( \sum_{i \in V} \sum_{j \in P} e_{ij} \right) \]  

\text{(1)}

Subject to:

\[ \sum_{k \in N(k \neq j)} x_{ik} = 1 \forall j \in P \]  

\text{(2)}

\[ \sum_{k \in N(k \neq j)} x_{ik} - \sum_{k \in N(k \neq j)} x_{i,j+1,k} = 0 \forall j \in P, \forall i \in V \]  

\text{(3)}

\[ \sum_{k \in N} x_{ik} = z_i \forall i \in V \]  

\text{(4)}

\[ \sum_{j \in N(j \neq h)} x_{ih} - \sum_{k \in N(k \neq h)} x_{ih} = 0 \forall i \in \text{Nr}, \forall i \in V \]  

\text{(5)}

\[ q_j + (T + t_k) x_{ik} \leq q_j + M(1 - x_{ik}) \forall i \in V, \forall j, k \in \text{Nr}(j \neq k) \]  

\text{(7)}

\[ q_j + \sum_{k \in N(k \neq j)} (T + t_k) x_{ik} \leq q_{j(j+1)} \forall i \in V, \forall j \in P \]  

\text{(8)}

\[ E \sum_{(j,k) \in A} x_{ik} \leq q_j \forall j \in P, \forall i \in V \]  

\text{(9)}

\[ E \sum_{(j,k) \in A} x_{ik} \geq q_j - w_{ij} + M \left( 1 - \sum_{(j,k) \in A} x_{ik} \right) \forall j \in P, \forall i \in V \]  

\text{(10)}

\[ VL_{ij} + (pi_i - dr_i)x_{ik} \leq VL_{ij} + M(1 - x_{ik}) \forall j, k \in N(j \neq k), \forall i \in V \]  

\text{(11)}

\[ VL_{ij} + u_i = Gc \sum_{k \in N(k \neq j)} x_{ik} \forall j \in \text{Nr}, \forall i \in V \]  

\text{(12)}

\[ \sum_{(j,k) \in A} d_{jk} x_{ik} \leq L \forall i \in V \]  

\text{(13)}

\[ q_{j(j+1)} - (q_j + T z_i) = r_j \forall j \in P, \forall i \in V \]  

\text{(14)}

\[ e_{ij} = r_j - \sum_{(j,k) \in A} t_{j(k+1)} x_{ik} \forall j \in P, \forall i \in V \]  

\text{(15)}

\[ x_{ik}, z_i \in \{0, 1\} \forall i \in V, (j, k) \in A \]  

\text{(16)}

\[ q_j, VL_{ij}, w_{ij}, u_i, r_j, e_{ij} \geq 0 \forall i \in V, j \in N \]  

\text{(17)}

The objective function (1) seeks to minimize the operating cost (travelling cost and route assignment cost), underutilization penalty cost, penalty cost for patients’ waiting times, and penalty cost for extra ride time. Constraints (2) guarantee that each transportation request is served by only one ambulance. Constraints (3) ensure that the same ambulance visits the pickup node and the respective delivery node. Constraints (4–6) guarantee that each route assigned ambulance starts its route at the origin node and returns to the final destination node at the end of the tour. Constraints (7) set the consistency for the ambulance arrival time variables at pickup/delivery nodes. Constraints (8) define the precedence relationship and assure that an ambulance visits the delivery node only after touring the corresponding pickup node. Constraints (9) ensure that the ambulance reaches a pickup node after the start of patient availability time for pickup as stated by the patient. Constraints (10) determine the waiting time found at each patient node. This waiting time represents the gap between the ambulance arrival time and the earliest start time for pickup as stated by the patient.

Constraints (11) calculate the ambulance load at each node toured by the NEPT ambulance. Constraints (12) ensure that the load does not surpass the ambulance capacity; they also determine the underutilization level of an ambulance at each toured node prior to arriving at the last delivery node. Constraints (13) guarantee that the ambulances do not violate the maximum route length restriction. Constraints (14) determine the total ride time of each patient during the ride to the destination node. Constraints (15) calculate the
extra ride time for each patient. Extra ride time would be the difference between the total ride time of a patient and the direct travelling time to the destination if the patient did not share the ride with other patients. Constraints (16) and (17) denote the binary and positive continuous decision variables, respectively.

4. Solution method

The NEPT problem presented in Section 3.1 is a DARP, which is NP-hard. Accordingly, medium to large instances of the problem are hard to solve through exact solution methods. Exact methods can, however, be used to evaluate the quality of the solutions produced by heuristic algorithms or meta-heuristic algorithms on small problem instances. Furthermore, exact solution methods can also be advantageous once good clusters are already available to optimize the individual routes using the cluster first, route second approach. In order to solve medium and large-sized instances of the NEPT problem, a heuristic framework, EKMIH, is proposed in this paper. In the first phase, the clustering method is used to partition the set of patients and aggregate them into proper clusters. Once the clusters are formed, each cluster is assigned a non-emergency ambulance from the fleet of available ambulances at the depot. Then, the MILP model is solved through CPLEX to optimize the route for patients of each cluster in the second phase. Clustering is the method of grouping patients with similar transportation requests together. The patients within a cluster have more similar features as compared

![EKMIH framework for NEPT.](image-url)
to the patients who belong to other clusters. In this study, the K-means clustering method is used to group patients. The K-means algorithm (Hartigan and Wong, 1979) is an important technique for solving clustering problems. The K-means algorithm is used to construct clusters of patients in a way that the patients in the same cluster are geographically close enough and can be served by the same ambulance more efficiently. This iterative clustering algorithm can be implemented as follows:

- Randomly select \( K \) number of patients out of a total \( n \) patients. These \( K \) patients will initially serve as centroids for the \( K \) clusters.
- Compute the Euclidean distance of each patient from every cluster center and assign the patient to the nearest cluster center.
- When all patients have been assigned, define the new cluster centers by recomputing the average Euclidian distance between patients allocated to each cluster and the cluster center.
- Repeat the previous two steps until there is no change in the allocated patients to each cluster center over a number of specified iterations.

It is well known that the K-means algorithm is sensitive to the initial centroids, and any changes to the initial centroids may deliver different results from the previous observations. Similarly, the features of data points used to compute the Euclidean distance may also influence the formation of clusters. Moreover, one more problem is that the K-means type algorithms handle all the features equally while making assignments to the clusters. In several real-life applications, some features need to be given more weightage owing to the nature of the problem and the diversity of the associated features. Therefore, we develop an enhanced version of the K-means algorithm. EKMIH improves the simple K-means through three different algorithmic enhancements: (i) instead of randomly selecting the \( K \)-patients as initial centroids, it selects priority-based initial centroids; (ii) after each convergence of K-means algorithm, the EKMIH filters the current clusters to identify distant cluster members; (iii) it inserts the distant members into other clusters following shared destination-based rules and updates the current centroids. This iterative framework assists in finding better clusters by regularly updating the centroids through the exploitation of problem-specific characteristics. The functioning of EKMIH is explained in Section 4.1.

4.1. EKMIH framework for NEPT

Subsection 4.1 presents the implementation of the EKMIH framework. All the necessary steps required to execute the EKMIH framework are written in sequential order under Subsection 4.1. Subsection 4.1.1 provides a method to select the priority-based centroids. Subsection 4.1.2 demonstrates the working of the K-means algorithm in the context of the EKMIH framework. Subsection 4.1.3 explains the strategy to identify distant members in each cluster. Subsection 4.1.4 shows how the distant members of a cluster can be relocated to other clusters based on the shared destination method. Subsection 4.1.5 describes the method to update centroids and clusters until there is no change in the clusters. Subsection 4.2 describes the overall operation of the EKMIH framework with the help of Fig. 1.

Given the number of clusters \( K \), EKMIH is implemented through the following steps:

4.1.1. Priority-based initial centroids

- Divide the patient availability start time for pickup of each pickup node \( j \in P \) by the number of persons to be picked at that pickup node \( (Es/p)j \) to get a priority list \( F \).
- Sort this list \( F \) in ascending order of the priority so that pickup nodes with an earlier pick-up time or more patients to be picked can be at earlier positions.
- Pick the first \( K \) patients from the list as the priority-based initial cluster centers.

4.1.2. K-means algorithm

After receiving the new centroids each time, the K-means algorithm is activated again to form new clusters until it converges. Let \( CP = \{CP_1, CP_2, CP_3, \ldots, CP_K\} \) be the set of clusters of patients, where \( CP_i \) is the set of patients in cluster \( i \). Note that (i) \( CP_i \neq \emptyset \), (ii) \( CP_i \cap CP_j = \emptyset \) \( \forall i, j = 1, 2, \ldots, K (i \neq j) \), and (iii) \( \bigcup_1^K CP_i = P \).

Each patient’s transportation request \( pr_i \) contains a set of attributes. To form more effective clusters, we incorporate all the important features of a request to compute the Euclidean distance. In our problem, coordinates of the origin and destination, number of riders and seats required, and preferred departure time are classified as features of each patient’s transportation request. The Euclidean distance between pickup node \( j \) and cluster \( k \), denoted as \( d_{jk} \), can be represented as a cost function defined as:

\[
d_{jk} = \sqrt{\sum_{i=1}^{m} (pr_{ij} - CP_{i,k})^2}
\]

where \( pr_{ij} \) is the \( i \)-th feature of the patient’s transportation request \( pr_j \), \( m \) is the total number of features characterized with each request, and \( CP_{i,k} \) is the \( i \)-th feature of the patient centroid \( CP_k \).

Let \( y_{jk} = 1 \) if a patient’s transportation request \( j \in P \) is assigned to a cluster center (centroid) \( k \in CP \), 0 otherwise. The maximum cluster size, i.e., the number of patient transportation requests which can be allocated to a cluster, is represented as \( Q \). One ambulance is assigned to one cluster. It is pertinent to mention that the value of \( Q \) is kept close to the vehicle capacity to avoid any infeasible solutions. However, the assignment of some extra patients in excess of the capacity does not create infeasibility in the context of our problem, as the assigned patients are continuously picked up and dropped off at different locations during the whole length of the
route. Hence, this process of clustering and assigning patients to clusters aims to solve the optimization problem as follows:

\[
min \sum_{i=1}^{n} \sum_{k=1}^{K} d_{jk}y_{jk} \tag{18}
\]

Subject to:

\[
\sum_{k=1}^{K} y_{jk} = 1 \forall j \in P \tag{19}
\]

\[
\sum_{j=1}^{n} y_{jk} \leq Q_{k} \forall k = 1, 2, \ldots, K \tag{20}
\]

\[
d_{jk} = \sqrt{\sum_{i=1}^{n} (p_{r_{ij}} - CP_{i,k})^2} \quad \forall j \in P, k = 1, 2, \ldots, K \tag{21}
\]

\[
y_{jk} \in \{0, 1\} \quad \forall j \in P, k = 1, 2, \ldots, K \tag{22}
\]

Objective function (18) minimizes the total assignment cost of allocating \( n \) patients/pickup nodes to \( K \) clusters. Constraints (19) ensure that a patient is assigned to exactly one cluster. Constraints (20) guarantee that the total number of assignments to a cluster does not exceed the maximum cluster size. Constraints (21) calculate the distances between all pairs of patients’ transportation requests and cluster centers. Constraints (22) impose the binary condition on the decision variables \( \{y_{jk}\} \).

4.1.3. Filtering the individual clusters

Based on the K-means algorithm’s solution, new clusters are filtered as follows:

- Calculate the individual weights for all the patients by computing the distance of each patient assigned to a specific cluster from the current centroid of that cluster using Equation (21).
- For each cluster \( CP_{i} \), sort the patient’s transportation request list \( EC_{CP} \) in ascending order of their distance to the centroid so that those with long distances are pushed at the end of the list.
- Multiply the total number of patients assigned to that specific cluster \( CP_{i} \) by a parameter \( \beta \) to identify a new set of distant members \( R_{CP} \) to be eliminated from the end of the list \( EC_{CP} \). In this study, we took \( \beta = 0.30 \).

4.1.4. Shared destination-based insertion

Match the destination of each of the distant members in \( R_{CP} = \{r_{1}, r_{2}, \ldots, r_{|R_{CP}|}\} \) with its current cluster \( CP_{i} \). If \( r_{j} \in R_{CP} \) shares the same destination with any other member of the cluster, it will stay in the current cluster; otherwise, start matching its destination with the patients assigned to other clusters and insert it into any single cluster whose member share the same destination as to \( r_{j} \).

4.1.5. Updating the centroids

Repeat Steps 3 and 4 for each cluster and update the centroid of each cluster with the updated members. Feed these centroids as new cluster centers to the K-means algorithm in Step 2. This process will continue until EKMIH observes no change in the cluster assignments from one iteration to the next. Upon termination, a cluster corresponds to a set of pick-up points to be served by the same ambulance.

4.2. Execution of EKMIH framework

Scheduling decisions for each cluster will be determined by solving the MILP model presented in Section 3.1. Fig. 1 presents the EKMIH framework developed in this paper. Initially, the value of \( K \) is calculated by dividing the demand by the capacity of the ambulance \( (GC) \). We calculate the total demand by adding up the number of people to be picked at all the pick-up nodes. Thus, value of \( K \) for \( n \), i.e., number of pickup nodes, can be calculated through this expression \( K = \sum_{i=1}^{n} p_{i} / GC \). The resulting value is rounded to the nearest integer not less than the value obtained from the expression. By taking into account this value of \( K \), priority-based initial centroids are computed as discussed in step 1. Subsequently, the K-means algorithm is used to create the clusters, as explained in step 2. The resulting centroids after each convergence of the K means are updated by implementing Steps 3 to 5 where all the clusters are filtered, and distant members of the clusters are inserted in other clusters depending upon their destination nodes. The final clusters obtained from step 5 are individually solved through the MILP model to create scheduling plans. Same ambulance visits all the pick-up nodes included in one cluster. This process continues until the solution value increases against the value of \( K \).

5. Computational experiments

Subsection 5.1 provides detail about the realistic data, parameter values and characteristics of problem instances used in the experiments. Subsections 5.2 to 5.5 present and discuss the results of various types of tests. Subsection 5.2 compares the results against three different solution methods and discusses the quality of obtained solutions. Subsections 5.3 and 5.4 investigate the impact on cost components of changing the values for the maximum cluster size and number of clusters, respectively. Subsection 5.5 discusses the sensitivity analysis to evaluate the trade-off and MILP model behavior under three different scenarios.

We conduct computational experiments to assess the performance of EKMIH on different instance sets. The effects of the variations
in the maximum cluster size and the number of clusters are analyzed in subsequent experiments. Additionally, the impact of weights associated with the performance measures is investigated to determine the trade-off between the performance measures by varying respective weightings in the objective function of the MILP model. The mathematical model formulation and EKMIH are implemented in OPL and solved by CPLEX. The computational tests are performed on a computer with 2.40 GHz Intel Core i5 CPU and 4 GB of RAM.

5.1. Instance generation

The mathematical model developed in this study and the heuristic solution framework are tested on realistic problem instances arising in the context of NEPT services being offered in Hong Kong. We study the operations of the auxiliary medical service (AMS) that offers free non-emergency ambulance transfer service (NEATS) to patients in Hong Kong. The core target customers are those who require medical care and other related services at the outpatient clinics of the Hospital Authority (HA) in Hong Kong or other clinics operated by private hospitals. For smooth functioning of the service, all transportation requests must be received at least one working day in advance of the appointment date and then scheduling status should be confirmed by a phone call. Each ambulance at AMS can carry six people at maximum. AMS has its training centers in all major areas of the city. These training centers are considered depots in our case study. Each ambulance starts and ends its tour at the designated depot. To carry out our experiments, we select one of its regional depots on the Hong Kong Island side as the origin and destination points for the ambulance routes. Fifteen different hospitals and outpatient clinics of the HA in the Central and Western district of Hong Kong are chosen as the locations of health care providers for the patients in this study. Ninety-six locations in the same district considered the potential pickup points or homes of patients who need transportation services. All of these locations are marked on the map shown in Fig. 2. We used Google Maps API to compute the actual distance ($d_{jk}$) and travel time ($t_{jk}$) between each pair of nodes.

The patient availability start times for pick-up are distributed randomly under a uniform distribution over a time period of 7 h. Considering the patients who use a wheelchair, an extra seat can be reserved for such cases. The number of seats required for each patient’s transportation request are randomly assigned between 1 and 2 with equal probabilities (i.e., 0.5). The cost values ($b$, $c$) are selected considering the cost levels in the background of Hong Kong, and these values are 4 Hong Kong Dollars (HKD) per kilometer of distance and 250 HKD per ambulance, respectively. The penalty costs for per seat underutilization ($g$), waiting per minute ($f$), and extra ride per minute ($h$), are kept equal to 1 HKD. For the first group of computational tests, all the performance metrics in the objective function are assigned identical weights ($\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$), although the varying weights are assigned in the sensitivity analysis to determine the trade-off between measures. The capacity of each ambulance ($G_c$) is fixed at six, and the service time is 10 min at each pick-up/delivery node.

To examine the scalability of our solution method, we consider seven problem sizes (named Sets A to G). Based on the dataset previously introduced, nineteen problem instances, in total, are constructed. The length of the problem instances stays identical within a set. In contrast, the length of the instances increases from Set A to Set G. For instance, Set A comprises four problem instances, and each instance within this set has the same size (4 pick-up nodes and 4 drop-off nodes); nevertheless, the geographical positions (i.e., coordinates) of pickup and drop-off locations are different for each instance within this set. Similarly, patient availability start times for pick-up and the number of people to be picked up at each pick-up point are different for each instance within this set. Table 3 presents the characteristics of the problem instances and values of input parameters, such as the number of pick-up nodes ($P$), number

![Fig. 2. Pick-up locations and hospitals/clinics for patients.](image)
of delivery nodes \((D)\), number of hospitals/clinics used as drop-off locations, number of ambulances \((V)\), total number of people to be picked up, and route length \((L)\).

5.2. Results

Table 4 compares the results for the instances against three different types of solution methods: (i) the integrated patient assignment and ambulance routing problem (MILP model) solved through CPLEX; (ii) a simple K-means based heuristic (SKMH) framework, where simple K-means is used for the clustering of patients, and MILP model for scheduling of patients for each cluster; and (iii) the EKMIH framework. Table 4 reports CPU run time (in seconds) and total cost (i.e., objective value) in HKD. Considering the first solution method, CPLEX could solve each instance of Set A optimally, whereas it could only deliver best integer solutions for all the instances of Set B after a run time of 2 h. CPLEX could not deliver any solution for the rest of the sets of the problem instances in a reasonable computational time. However, both SKMH and EKMIH could solve all the problem instances efficiently. The solution time is observed to increase significantly as the problem instances change from Set A to Set B. The value of total cost increases with the increasing size of the problem instances across all the solution methods.

Optimality gap is computed to assess the performance of heuristic solution methods against the MILP model. The CPLEX solutions obtained by employing the first solution method are used to evaluate the heuristic methods. As CPLEX solutions are only available for the first two sets (i.e., Sets A and B) and the results received from the two heuristic methods are the same for the first two sets, optimality gaps are shown in the single Column 9 against two heuristic methods. It is computed as follows:

\[
\left( \frac{\text{Total cost from EKMIH solution} - \text{Total cost from MILP}}{\text{Total cost from MILP}} \right) \times 100\%.
\]

Best integer solution from CPLEX is used to measure the performance for instances of Set B. CPLEX produces these results with less than 45% optimality gap. It can be observed that there is a zero optimality gap for the instances of Set A, whereas the gap remains below 5% for the instances of Set B.

Furthermore, the relative gap (R. gap), as shown in Column 10, is computed to compare two heuristic solution methods. The R. gap is computed as follows:

\[
\left( \frac{\text{Total cost from SKMH solution} - \text{Total cost from EKMIH solution}}{\text{Total cost from EKMIH solution}} \right) \times 100\%.
\]

Out of the eight instances where SKMH and EKMIH produce different solutions, EKMIH achieved better solution quality (i.e., smaller R. gap) than SKMH in six instances. While results remain the same for the rest of the instances, the computational results demonstrate that, in general, EKMIH is able to deliver better solutions than SKMH and the run time for EKMIH is shorter than SKMH.

Table 5 shows the breakup of total cost against all the sets of instances and solution methods. The operating cost, travelling cost (TRC) and ambulance route assignment cost (ARC), increases from the first set to the last set following the increasing size of the instances. The variations in the values of TRC within each set are due to the different geographical locations of transportation requests’ origins and destinations. Underutilization cost (UC) reflects the unused capacity, and its values do not show much variation within the same set as the total number of people to be picked remains the same for each instance within the same set. Interestingly, both waiting cost (WC) and extra ride time cost (ERC) show considerable variation within the same set. Moreover, although per-minute penalty values for waiting at the pickup location and extra ride time on an ambulance are the same, WC observes higher values than ERC. Figs. 3 and 4 compare WC and ERC for both SKMH and EKMIH respectively. It is observed that WC has higher values than ERC in both

Table 3

The characteristics of the problem instances.

| Inst. | # Pickup nodes | # Delivery nodes | # Hospitals | # Ambulances | Total pickups | Route length |
|-------|---------------|-----------------|-------------|--------------|---------------|--------------|
| A1    | 4             | 4               | 2           | 1            | 6             | 40           |
| A2    | 4             | 4               | 2           | 1            | 6             | 40           |
| A3    | 4             | 4               | 2           | 1            | 6             | 40           |
| A4    | 4             | 4               | 2           | 1            | 6             | 40           |
| B1    | 8             | 8               | 3           | 2            | 12            | 40           |
| B2    | 8             | 8               | 3           | 2            | 12            | 40           |
| B3    | 8             | 8               | 3           | 2            | 12            | 40           |
| B4    | 8             | 8               | 3           | 2            | 12            | 40           |
| C1    | 16            | 16              | 6           | 4            | 24            | 100          |
| C2    | 16            | 16              | 6           | 4            | 24            | 100          |
| C3    | 16            | 16              | 6           | 4            | 24            | 100          |
| C4    | 16            | 16              | 6           | 4            | 24            | 100          |
| D1    | 32            | 32              | 8           | 8            | 48            | 125          |
| D2    | 32            | 32              | 8           | 8            | 48            | 125          |
| D3    | 32            | 32              | 8           | 8            | 48            | 125          |
| E1    | 48            | 48              | 10          | 12           | 72            | 125          |
| E2    | 48            | 48              | 10          | 12           | 72            | 125          |
| F     | 64            | 64              | 12          | 16           | 96            | 125          |
| G     | 96            | 96              | 15          | 24           | 144           | 125          |
methods. It indicates that patients spend more time waiting for the ambulance compared to the extra ride time on the ambulance.

5.3. Effect of the maximum cluster size on cost

In this set of computational experiments, we evaluate the effect of variation in the maximum cluster size on the solution quality. Instances B4 and C4 are used to conduct these experiments. The value of \( Q \) is changed between 4 and 8. The value of \( K \) for each instance remains the same as it was observed in the previous experiments for these respective instances. Therefore, the ambulance route assignment cost remains the same as in the original experiments. Figs. 5 and 6 show the results for the variation in the maximum size of the clusters for instances B4 and C4 respectively.

The results depicted in Fig. 5 indicate that travelling cost \( TRC \) and underutilization cost \( UC \) decline with the increasing size of the \( Q \). Hence, the change in the size of the \( Q \) has a positive effect on operating costs and underutilization penalty, whereas a negative impact is witnessed for user inconvenience measures. Waiting cost \( WC \) and extra ride time cost \( ERC \) start increasing when the value of \( Q \) extends beyond 5 but eventually show consistent values at the end of the curve. Thus, a larger cluster size may also increase the waiting time and extra ride time for patients. Similar results can be observed in the case of instance C4. The results presented in Fig. 6 reveal that \( TRC \) and \( UC \) decrease with the increasing size of the \( Q \). While \( WC \) increases after the first experiment, \( ERC \) also starts increasing eventually at the end of the curve.

5.4. Effect of the number of clusters on cost

The number of clusters, \( K \), plays an important role in obtaining better clusters in the EKMIH framework. Therefore, it is necessary to find such a value of \( K \) which best suits the clustering algorithm and characteristics of the considered problem. To analyze the impact of the value of \( K \), we perform computational experiments on Instance C2 by examining different values of \( K \). For these experiments, we fix the value of \( Q \) equal to 7 and change the value of \( K \) between 3 and 5. Fig. 7 depicts the results for different values of \( K \) against the cost components.

The results demonstrate that the penalties associated with the patients’ inconvenience (WC, ERC) incur a minimum cost when \( K = 5 \), while operating cost (OPC) incurs the minimum value when \( K = 3 \). Although OPC observes the minimum value at \( K = 3 \), the total cost does not essentially decrease as the cost for WC and ERC is higher due to a fewer number of clusters (ambulances) and more patients assigned to the ambulances. As the value of \( K \) moves from 3 to 4, OPC surges, whereas the patient’s inconvenience penalties experience a decline in their values. Thus, a balanced value of \( K \) should be sought to achieve the best results for the problem. In Instance C2, the best value of \( K \) is 4 as ERC value is lower and OPC value is not as high as it is when \( K = 5 \).

5.5. Sensitivity analysis

The NEPT planning staff makes decisions bearing in mind real-life situations where they have to follow some limits related to both transportation budget and service quality. Although some of the parameters used in the MILP are patient-specific (e.g., \( p_j \) and \( E_0 \)) while some other parameters (e.g., \( t_{jk} \), \( d_j \) and \( G_c \)) depend on the real-life situation, some crucial parameters must be fine-tuned

---

**Table 4**

| Instance | MILP Solution | SKMH Framework | EKMIH Framework | Gap (%) | R. Gap (%) |
|----------|---------------|----------------|-----------------|---------|------------|
| No.      | Type | CPU (sec) | Total cost | CPU (sec) | Total cost | CPU (sec) | Total cost |         |          |
| 1        | A1   | 17       | 515.40    | 20         | 515.40    | 20         | 515.40    | 0       | 0.00      |
| 2        | A2   | 18       | 519.60    | 21         | 519.60    | 20         | 519.60    | 0       | 0.00      |
| 3        | A3   | 16       | 525.00    | 19         | 525.00    | 19         | 525.00    | 0       | 0.00      |
| 4        | A4   | 15       | 486.60    | 23         | 486.60    | 19         | 486.60    | 0       | 0.00      |
| 5        | B1   | 7200     | 1002.60   | 33         | 1026.60   | 34         | 1026.60   | 2.39    | 0.00      |
| 6        | B2   | 7200     | 1137.50   | 30         | 1192.20   | 33         | 1192.20   | 4.81    | 0.00      |
| 7        | B3   | 7200     | 968.70    | 33         | 996.00    | 29         | 996.00    | 2.82    | 0.00      |
| 8        | B4   | 7200     | 1020.40   | 32         | 1067.40   | 33         | 1067.40   | 4.61    | 0.00      |
| 9        | C1   | -        | -         | 48         | 2233.19   | 46         | 2235.99   | -       | −0.13     |
| 10       | C2   | -        | -         | 44         | 2192.40   | 48         | 2192.40   | -       | 0.00      |
| 11       | C3   | -        | -         | 50         | 2187.40   | 52         | 2101.60   | -       | 4.08      |
| 12       | C4   | -        | -         | 50         | 2329.60   | 47         | 2268.40   | -       | 2.70      |
| 13       | D1   | -        | -         | 68         | 4679.20   | 61         | 4679.20   | -       | 0.00      |
| 14       | D2   | -        | -         | 70         | 4565.40   | 68         | 4568.20   | -       | −0.06     |
| 15       | D3   | -        | -         | 67         | 4669.00   | 76         | 4664.60   | -       | 0.09      |
| 16       | E1   | -        | -         | 135        | 7399.00   | 143        | 7399.00   | -       | 0.00      |
| 17       | E2   | -        | -         | 141        | 7146.00   | 138        | 7111.20   | -       | 0.49      |
| 18       | F    | -        | -         | 330        | 9786.40   | 337        | 9698.80   | -       | 0.90      |
| 19       | G    | -        | -         | 755        | 14965.40  | 788        | 14891.20  | -       | 0.50      |
## Table 5
Cost details against all performance measures.

| Instance | MILP Solution | SKMH Framework | EKMIH Framework |
|----------|----------------|----------------|-----------------|
| No.      | TRC  | ARC  | WC  | UC  | ERC  | TRC  | ARC  | WC  | UC  | ERC  | TRC  | ARC  | UC  | ERC  |
| 1        | A1   | 166.40 | 250 | 40  | 33   | 26   | 166.40 | 250 | 40  | 33   | 26   | 157.60 | 250 | 44  | 32   | 36   |
| 2        | A2   | 157.60 | 250 | 44  | 32   | 36   | 157.60 | 250 | 44  | 32   | 36   | 157.60 | 250 | 44  | 32   | 36   |
| 3        | A3   | 156.00 | 250 | 56  | 32   | 31   | 156.00 | 250 | 56  | 32   | 31   | 156.00 | 250 | 56  | 32   | 31   |
| 4        | A4   | 138.80 | 250 | 43  | 33   | 22   | 138.80 | 250 | 43  | 33   | 22   | 138.80 | 250 | 43  | 33   | 22   |
| 5        | B1   | 285.60 | 500 | 93  | 66   | 58   | 295.60 | 500 | 91  | 60   | 80   | 295.60 | 500 | 91  | 60   | 80   |
| 6        | B2   | 305.50 | 500 | 160 | 68   | 104  | 335.20 | 500 | 195 | 65   | 97   | 335.20 | 500 | 195 | 65   | 97   |
| 7        | B3   | 268.70 | 500 | 76  | 65   | 59   | 290.00 | 500 | 38  | 63   | 105  | 290.00 | 500 | 38  | 63   | 105  |
| 8        | B4   | 291.40 | 500 | 105 | 64   | 60   | 310.40 | 500 | 103 | 64   | 90   | 310.40 | 500 | 103 | 64   | 90   |
| 9        | C1   | -     | -    | -    | -    | -    | 631.20 | 1000 | 205 | 117  | 280  | 634.00 | 1000 | 202 | 118  | 282  |
| 10       | C2   | -     | -    | -    | -    | -    | 664.40 | 1000 | 283 | 129  | 116  | 664.40 | 1000 | 283 | 129  | 116  |
| 11       | C3   | -     | -    | -    | -    | -    | 508.40 | 1000 | 376 | 126  | 177  | 501.60 | 1000 | 299 | 125  | 176  |
| 12       | C4   | -     | -    | -    | -    | -    | 737.60 | 1000 | 277 | 126  | 189  | 682.40 | 1000 | 199 | 122  | 265  |
| 13       | D1   | -     | -    | -    | -    | -    | 1205.20 | 2000 | 857 | 251  | 366  | 1205.20 | 2000 | 857 | 251  | 366  |
| 14       | D2   | -     | -    | -    | -    | -    | 1176.40 | 2000 | 603 | 220  | 566  | 1169.20 | 2000 | 613 | 220  | 566  |
| 15       | D3   | -     | -    | -    | -    | -    | 1202.00 | 2000 | 834 | 244  | 389  | 1135.60 | 2000 | 849 | 236  | 444  |
| 16       | E1   | -     | -    | -    | -    | -    | 1906.00 | 3000 | 1614 | 378  | 501  | 1906.00 | 3000 | 1614 | 378  | 501  |
| 17       | E2   | -     | -    | -    | -    | -    | 1632.00 | 3000 | 1194 | 317  | 1003 | 1673.20 | 3000 | 1157 | 322  | 959  |
| 18       | F    | -     | -    | -    | -    | -    | 2300.40 | 4000 | 1765 | 447  | 1274 | 2398.80 | 4000 | 1892 | 468  | 940  |
| 19       | G    | -     | -    | -    | -    | -    | 3472.40 | 6000 | 3361 | 690  | 1442 | 3469.20 | 6000 | 3264 | 690  | 1468 |
Fig. 3. WC and ERC for the SKMH framework.

Fig. 4. WC and ERC for the EKMIH framework.

Fig. 5. Impact of increase in maximum cluster size on the cost for instance B4.
following the real-life challenges of the specific healthcare organizations. Preference among the performance measures of the objective function is expressible by assigning different values to the respective weights of these measures. We perform a sensitivity analysis to observe the behavior of the MILP model against the variation in the weights of the performance measures. For this analysis, one test instance with seven pickup locations and seven destination points is constructed. The complete problem is solved through CPLEX. This instance follows the same set of rules as defined in Section 5.1.

Table 6
Variation in the weightage of performance measures for all scenarios.

| Sr. # | \(a_1 = 0 \text{ to } 1, a_2 = a_3 = a_4 = (1 - a_4)/3\) | Sr. # | \(a_4 = 0 \text{ to } 1, a_1 = a_2 = a_3 = (1 - a_3)/3\) |
|-------|------------------------------------------------|-------|------------------------------------------------|
|       | \(a_1\)  \(a_2\)  \(a_3\)  \(a_4\)            |       | \(a_1\)  \(a_2\)  \(a_3\)  \(a_4\)            |
| 1     | 0.05  0.32  0.32  0.32                          | 12    | 0.32  0.32  0.05  0.32                          |
| 2     | 0.10  0.30  0.30  0.30                          | 13    | 0.30  0.30  0.10  0.30                          |
| 3     | 0.20  0.27  0.27  0.27                          | 14    | 0.27  0.27  0.20  0.27                          |
| 4     | 0.30  0.23  0.23  0.23                          | 15    | 0.23  0.23  0.30  0.23                          |
| 5     | 0.40  0.20  0.20  0.20                          | 16    | 0.20  0.20  0.40  0.20                          |
| 6     | 0.50  0.17  0.17  0.17                          | 17    | 0.17  0.17  0.50  0.17                          |
| 7     | 0.60  0.13  0.13  0.13                          | 18    | 0.13  0.13  0.60  0.13                          |
| 8     | 0.70  0.10  0.10  0.10                          | 19    | 0.10  0.10  0.70  0.10                          |
| 9     | 0.80  0.07  0.07  0.07                          | 20    | 0.07  0.07  0.80  0.07                          |
| 10    | 0.90  0.03  0.03  0.03                          | 21    | 0.03  0.03  0.90  0.03                          |
| 11    | 1.00  0   0   0                               | 22    | 0   0   1.00  0                               |
| 12    | 0.32  0.32  0.32  0.32                          | 23    | 0.32  0.32  0.32  0.32                          |
| 13    | 0.30  0.30  0.30  0.30                          | 24    | 0.30  0.30  0.30  0.30                          |
| 14    | 0.27  0.27  0.27  0.27                          | 25    | 0.27  0.27  0.27  0.27                          |
| 15    | 0.23  0.23  0.30  0.30                          | 26    | 0.23  0.23  0.30  0.30                          |
| 16    | 0.20  0.20  0.40  0.40                          | 27    | 0.20  0.20  0.20  0.20                          |
| 17    | 0.17  0.17  0.50  0.50                          | 28    | 0.17  0.17  0.20  0.20                          |
| 18    | 0.13  0.13  0.60  0.60                          | 29    | 0.13  0.13  0.13  0.13                          |
| 19    | 0.10  0.10  0.70  0.70                          | 30    | 0.10  0.10  0.10  0.10                          |
| 20    | 0.07  0.07  0.80  0.80                          | 31    | 0.07  0.07  0.07  0.07                          |
| 21    | 0.03  0.03  0.90  0.90                          | 32    | 0.03  0.03  0.03  0.03                          |
| 22    | 0   0   1.00  0                               | 33    | 0   0   0   1.00                              |

Fig. 6. Impact of increase in maximum cluster size on the cost for instance C4.

Fig. 7. Impact of value of K on the cost.

Table 6
Variation in the weightage of performance measures for all scenarios.
In the first set of scenarios (Scenarios 1 to 11), \( \alpha_1 \), the weightage assigned to the operating cost (OPC), is changed between 0 and 1, while the remaining three weights \( \alpha_2, \alpha_3 \) and \( \alpha_4 \) share equal weightage out of 1 – \( \alpha_1 \). Similarly, for the second (Scenarios 12 to 22) and third set (Scenarios 23 to 33) of scenarios, the weightage of patient inconvenience measures, \( \alpha_3 \) and \( \alpha_4 \), is correspondingly varied between 0 and 1, respectively, as for \( \alpha_1 \) in the first set of scenarios. The variation in the weightage for these three sets of scenarios is given in Table 6. Figs. 8–10 show the results for these three sets of scenarios comparing the performance measures.

The experiments in the first set of scenarios (Scenarios 1 to 11) aim to examine the impact of operating cost on underutilization cost (UC) and patient inconvenience costs (patients’ waiting time cost (WC) and extra ride time cost (ERC)). The MILP solution results as shown in Fig. 8 depict that patients’ inconvenience measures, waiting time and extra ride time, have the lowest penalty when operating cost is given the least weightage. The lowest value for waiting time and extra ride time penalty is achieved when \( \alpha_1 = 0.05 \). Their values remain constant for some consecutive observations and rise significantly as we give higher weightage to the operating cost in the last few tests. The values for patient’s inconvenience measures particularly waiting time experience a significant high escalation for the last three values of \( \alpha_1 (\alpha_1 = 0.8, \alpha_1 = 0.9, \alpha_1 = 1.0) \). Underutilization costs remain unaffected and maintain the same value from the first to the last test. From this set of experiments, a trade-off exists between the operating cost and the patient inconvenience.

In the second set of scenarios (Scenarios 12 to 22), the tests seek to assess the effect of the first user inconvenience measure, waiting time, on the rest of the measures. Fig. 9 shows that operating costs observe small values against the lower weightage of \( \alpha_3 \), whereas extra ride time cost does not show any sensitivity to the lower weightage given to waiting time. The underutilization penalty does not show much variation as expected due to the fixed capacity of the ambulances. Moreover, the underutilization penalty depends on the number of visited nodes, and these remain the same in every test. Hence, the lower preference for waiting time keeps the operating costs lower. Reciprocally, operating costs receive a significant rise as the weightage for \( \alpha_3 \) increases in the last four tests. The waiting penalty observes the minimum value equal to 0 when it is assigned the maximum weightage \( \alpha_3 = 1 \). Interestingly, the second inconvenience measure, i.e., extra ride time penalty, experiences a high rise when the waiting time penalty has the lowest value on account of the maximum weightage assigned to it. Hence, the trade-off exists not only between operating cost and patient inconvenience, but also between the two inconvenience measures.

To observe the impact of the second inconvenience measure (i.e., extra ride time) on other performance measures, the value of \( \alpha_4 \) is varied in the third set of scenarios (Scenarios 23–33). Fig. 10 displays the results for this set of scenarios. It is evident from the results that the penalty on extra ride time does not have any impact on other performance measures, particularly when it is assigned the lower weightage. However, it affects the waiting time in the last test while other measures remain unaffected. The low sensitivity of extra ride time cost in this set of scenarios can be attributed to the bigger gaps between patient availability start times for pickup. However, as observed in the first set of scenarios, the extra ride time shows high sensitivity when higher weightage is assigned to the operating cost. This behavior of extra ride time penalty is understandable as patients spend more time in the ambulance or observe more waiting time due to high preference given to operating costs.

The results of the sensitivity analysis indicate the behavior of the MILP model when different preferences are given to the performance measures. The findings clearly answer the research questions raised in the Introduction Section. The results of the experiments in the first set of scenarios (Scenarios 1–11) evidently show that there is a trade-off between user inconvenience measures and operating cost. Patients’ inconvenience measures have higher values when the operating cost is prioritized, and vice versa. When the weightage for operating cost (value of \( \alpha_1 \)) is varied between 0.8 and 1.0, the penalty values for waiting time and extra ride time show a significant escalation. Waiting time, in particular, increases by 300% against the last three weightage values of operating cost. Interestingly, the results of the second set of scenarios (Scenarios 12 to 22) not only approve the results of the first set of scenarios but also show that there is a trade-off between the two inconvenience measures in specific situations. The operating cost incurs a substantial increase of approximately 94% when the waiting time weightage (value of \( \alpha_2 \)) surges in the last four experiments of the second set of scenarios, whereas the extra ride time experiences a growth of 80% in its penalty values for the same last two experiments of the second set of scenarios. The results presented in subsections 5.3 and 5.4 also reveal the trade-off by showing how the user inconvenience measures respond when we increase the number of clusters (high operating cost) or increase the cluster size (low operating cost). Hence this study provides valuable insights to the NEPT planners to execute their plans considering the trade-off between these performance measures. The decision-makers can reduce the patient inconvenience by making proper preference rules among operating costs, the waiting time penalty and the extra ride time penalty. Service quality levels and operating budgets are critical for sustaining the operations. Hence, healthcare managers can adjust their NEPT service quality targets considering the trade-off between the performance indicators.

Making the decision-making process more inclusive, health care managers can also seek feedback from patient groups on their views on trade-offs. Perhaps they might provide some insights on which is more upsetting: (1) waiting time or extra ride time?, and (2) quick but expensive service or inexpensive service with some discomfort? Providing patients with more information and authority in the decision-making process can lead to well-informed decisions and increased satisfaction.

6. Conclusion

NEPT services are rapidly growing on account of the increasing elderly population and demographic changes in today’s society. However, elderly patients still face unique barriers and challenges in accessing efficient and cost-effective services. These NEPT services are not only beneficial to the elderly, but they are also effective in mitigating the risk of contagious disease outbreaks (e.g., Covid-19 and SARS) by reducing the chance of patients spreading the diseases in public transportation networks. This article explores the NEPT services being provided by the Hospital Authority in Hong Kong. The problem formulation is developed considering the DARP featuring a weighted objective function that contains four different performance measures. The problem takes into account the
Fig. 8. Results for performance measures against the first set of scenarios (Scenarios 1 to 11).

Fig. 9. Results for performance measures against the second set of scenarios (Scenarios 12 to 22).

Fig. 10. Results for performance measures against the third set of scenarios (Scenarios 23 to 33).
limitations relating to the earliest patient availability time, ambulance route length and capacity restriction. We develop an iterative heuristic framework, EKMIH, which groups the patients into clusters and subsequently creates schedules through CPLEX for the patients assigned to each cluster. The mathematical model and EKMIH framework are tested on a real-life NEPT case study of Hong Kong. The computational results on 19 problem instances demonstrate the effectiveness and efficiency of the heuristic framework and successful implementation of the MILP model to perform the scheduling decisions. For all instances where CPLEX delivered the result, our heuristic framework gives less than a 5% optimality gap. Moreover, we also perform additional experiments to analyze the impact of cluster size and the number of clusters on the solution quality and cost components. Additionally, a sensitivity analysis is conducted to study the behavior of the MILP model concerning different performance measures. It is noted that trade-off exists between operating costs and patient inconvenience measures. When operating costs are weighted at 0.05 in the objective function, the penalty value for user inconvenience measures is the lowest. Interestingly, trade-off also appears between the inconvenience measures (waiting time and extra ride time) when one of them is assigned a higher weightage.

In future work, these performance metrics can be further explored by integrating dynamic patient arrivals and coordinating between emergency and non-emergency ambulance fleets. The NEPT problem studied in this article has very important transport and health implications. The historical data of patient attributes and the outcome of scheduling and routing decisions can be used to evolve future policies. Hence, another important future research direction is to integrate the idea of context histories and context prediction in our proposed NEPT problem. The NEPT problem involves a lot of data about patient mobility and service standards, and these patient-specific data (appointment date, clinic visited, pickup location, time windows, waiting time and ride time) are usually available in the records of service providers. The term user context usually refers to data on people, groups, and physical or computational things, such as their current situation, identity, and space-time location (Dupont et al., 2020). Availability of context history assists in exploring the past behavior of users (patients). Thus, the service providers can act proactively based on contextual data of patients, notifying them in advance about the potential delays, and waiting time or assisting them in performing their healthcare checkup tasks more efficiently. Using the machine learning approaches, Bala and Chana (2015) designed intelligent task failure prediction models for supporting proactive fault tolerance by anticipating task failures in scientific workflow applications. Rosa et al. (2015) developed a multi-Temporal Context-aware System to help workers build their competencies by utilizing their current and past contexts. Researchers have recently explored several methods for analyzing the context data, including context prediction methodologies (da Rosa et al., 2016) and similarity analysis (Filippetto et al., 2021; Wiedmann et al., 2016). Such data analysis strategies can help to predict the future behavior of users. Consequently, the decision-makers can take proactive steps to formulate the policies and manage the relevant resources. Interestingly, the concepts of context histories and context prediction have found applications in many different fields, for instance, workers competence management (Rosa et al., 2015), failure prediction in cloud computing services (Bala and Chana, 2015), risk prediction (Filippetto et al., 2021) and intelligent services in transport systems (Zanella Gomes et al., 2019). Also, the context prediction adaptive model, ORACON (da Rosa et al., 2016), and CHSPAM model (Dupont et al., 2020) for discovery and monitoring of sequential patterns can be applied in many different domains, including healthcare, because of their flexibility. In conclusion, the context histories and context prediction-based methodologies can be applied to the NEPT problem to predict the anticipated values for the performance measures (waiting time, extra ride time) and operational costs. Furthermore, healthcare managers can use such data analysis and prediction tools to develop better policies for patient mobility and resource management.

Statement of contribution

Formulation of research questions: Jamal Abdul Nasir, Yong-Hong Kuo, Reynold Cheng.
Overall design of the methodologies: Jamal Abdul Nasir, Yong-Hong Kuo, Reynold Cheng.
Development of computational algorithms: Jamal Abdul Nasir.
Computational experiments: Jamal Abdul Nasir.
Result analysis: Jamal Abdul Nasir.
Manuscript writing: Jamal Abdul Nasir, Yong-Hong Kuo, Reynold Cheng.
Supervision: Yong-Hong Kuo.

Financial disclosure

Jamal Abdul Nasir and Yong-Hong Kuo were supported by ECS grant 14202115 from the Hong Kong Research Grants Council, the 2019 Guangdong Special Support Talent Program—Innovation and Entrepreneurship Leading Team (China) (2019B02S593), and HKU Seed Fund for Basic Research grant 201910159164. Reynold Cheng was supported by the COVID-19 Action Seed Funding (Second Round), HKU (No. 176FRCKCHE, 177DRCKCHE) and the Guangdong–Hong Kong–Macau Joint Laboratory Program 2020 (Project No: 2020B1212030009).

Declaration of competing interest

Nil.
Parragh, Sophie N., 2011. Introducing heterogeneous users and vehicles into models and algorithms for the dial-a-ride problem. Transport. Res. C Emerg. Technol. 19 (5), 912–930.

Parragh, Sophie N., Doerner, Karl F., Hartl, Richard F., Gandibleux, Xavier. 2009. A heuristic two-phase solution approach for the multi-objective dial-a-ride problem. Networks 54 (4), 227–242.

Parragh, Sophie N., Doerner, Karl F., Hartl, Richard F., 2010. Variable neighborhood search for the dial-a-ride problem. Comput. Oper. Res. 37 (6), 1129–1138.

Peiris, J.S.M., Lai, S.T., Poon, L.L.M., Guan, Y., Yam, L.Y.C., Lim, W., Nicholls, J., Yee, W.K.S., Yan, W.W., Cheung, M.T., et al., 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 361 (9366), 1319–1325.

Qu, Yuan, Bard, Jonathan F., 2013. The heterogeneous pickup and delivery problem with configurable vehicle capacity. Transport. Res. C Emerg. Technol. 32 (1–20).

Rasmussen, Matias Sevel, Justesen, Tor, Dohn, Anders, Larsen, Jesper. 2012. The home care crew scheduling problem: preference-based visit clustering and temporal dependencies. Eur. J. Oper. Res. 219 (3), 598–610.

Rest, Klaus-Dieter, Hirsch, Patrick. 2016. Daily scheduling of home health care services using time-dependent public transport. Flex. Serv. Manuf. J. 28, 495–525.

Ross, Tim, 2013. Advancing ontario’s accessibility: a study of linguistic, discursive, and conceptual barriers. Can. J. Urban Res. 22, 126–144.

Schildt, M., Doerner, K.F., Hartl, R.F. 2011. Metaheuristics for the dynamic stochastic dial-a-ride problem with expected return transports. Comput. Oper. Res. 38 (12), 1719–1730.

Shao, Yufen, Bard, Jonathan F., Jarrah, Ahmad I., 2012. The therapist routing and scheduling problem. IIE Trans. 44 (10), 868–893.

Smith, Matthew Lee, Prohaska, Thomas R., MacLeod, Kara E., Ory, Marcia G., Eisenstein, Amy R., Ragland, David R., Irmiter, Cheryl, Towne, Samuel D., Satariano, William A., 2017. Non-emergency medical transportation needs of middle-aged and older adults: a rural-urban comparison in Delaware, USA. Int. J. Environ. Res. Publ. Health 14 (2).

Souza, André da Silva, Chagas, Jonatas, Penna, Puca, Souza, Marcone, 2020. A hybrid heuristic algorithm for the dial-a-ride problem, pp. 53–66.

Titchkosky, Tanya, 2001. Disability: a rose by any other name;’people-first’ language in canadian society’. Can. Rev. Sociol./Revue canadienne de sociologie 38, 125–140.

Van Den Berg, P.L., Van Essen, J.T., 2019. Scheduling non-urgent patient transportation while maximizing emergency coverage. Transport. Sci. 1–18.

Wallace, Richard, Hughes-Cromwick, Paul, Mull, Hillary, Khannabis, Snehamay, 2005. Access to health care and nonemergency medical transportation: two missing links. Transport. Res. Rec. 1924 (1), 76–84, 2005.

Wiedmann, Tiago, Barbosa, Jorge Luís Vitória, Rigo, Sandro José, Barbosa, Débora Nice Ferrari, 2016. Recsim: a model for learning objects recommendation using similarity of sessions. J. Univers. Comput. Sci. 22, 1175–1200.

Yalçındağ, Semih, Matta, Andrea, Şahin, Evren, George Shanthikumar, J., 2014. A two-stage approach for solving assignment and routing problems in home health care services. In: Proceedings of the International Conference on Health Care Systems Engineering. Springer International Publishing, pp. 47–59.

Zanella Gomes, Joneval, Vitória Barbosa, Jorge Luís, Resin Geyer, Claudio Fernando, Santos dos Anjos, Julio Cesar, Vicente Canto, José, Pesin, Gustavo, 2019. Ubiquitous intelligent services for vehicular users: a systematic mapping. Interact. Comput. 31, 465–479.

Zhang, Zhenzhen, Liu, Mengyang, Lim, Andrew, 2015. A memetic algorithm for the patient transportation problem. Omega 54, 60–71.

Zhang, Zizhen, Qin, Hu, Wang, Kai, He, Huang, Liu, Tian, 2017. Manpower allocation and vehicle routing problem in non-emergency ambulance transfer service. Transport. Res. E Log. 106, 45–59.