Journal of Industrial and Management Optimization

https://www.aimspress.com/journal/jimo

JIMO, 22(6): 2843–2865.
DOI:10.3934/jimo.2026104
Received: 26 March 2026
Revised: 05 May 2026
Accepted: 14 May 2026
Published: 22 May 2026

Research article

Inter-hospital ambulance routing in Sri Lanka

Sudheeraka Wickramarachchi[1,*], **Kazuki Hasegawa**[2] **and Wei Wu**[1,2]

[1] Graduate School of Science and Technology, Shizuoka University, Japan

[2] Graduate School of Integrated Science and Technology, Shizuoka University, Japan

* **Correspondence:** Email: sudheeraka.22@shizuoka.ac.jp.

Abstract: In this study, we examine the emergency medical transportation system in the Sri Lankan public health service and develop an efficient routing framework for inter-hospital patient transfers, where patients are moved between hospitals according to clinical requirements and hospital capability levels. In this transfer process, medical experts emphasize that the minimum risk time of each patient should not be exceeded. To address this problem, we propose a mixed-integer programming (MIP) model and test it on random instances with up to 25 patients. Computational results showed that the MIP approach fails to obtain a feasible solution within a 300-second time limit. To solve real-world cases with up to 132 patients, we propose three heuristic approaches. The first is a two-phase method that combines machine learning with a simplified MIP model that reduces the feasible region (ML-MIP). The second heuristic, TP-VRP, is also a two-phase approach in which patient–hospital assignments are determined in the first phase, and the remaining problem is solved as a vehicle routing problem with time windows (VRP-TW) in the second phase. The third heuristic, RCPSP-VRP, combines a resource-constrained project scheduling problem for hospital assignments with the VRP-TW solution approach used in the second method. On 15 tested real-world instances, TP-VRP and RCPSP-VRP obtained feasible solutions for all cases, whereas MIP and ML-MIP failed to obtain feasible solutions for 8 out of 15 instances within 300 seconds. Compared to MIP, RCPSP-VRP also reduced average runtime by approximately 98.3% on the seven instances where both methods yielded feasible solutions, and achieved equal or better objective values, improving two instances.

Keywords: inter-hospital ambulance routing problem; minimum risk time; mixed integer programming; machine learning; vehicle routing problem with time windows
Mathematics Subject Classification: 90-10, 90B90, 90C11

1. Introduction

Emergency medical transportation systems (EMTSs) encompass a range of services, including ambulance dispatching, routing, and tracking, aimed at providing immediate medical assistance to individuals in emergency situations. The existing literature on EMTS focuses on optimizing response times, resource allocation, and overall system efficiency [1]. Several studies have explored various aspects, such as dispatch protocols [1], ambulance routing optimization [2], and dynamic resource allocation strategies [3]. One critical component of EMTS is the rapid response to emergencies, which involves picking up patients and transporting them to hospitals. However, the inter-hospital EMTS, which entails transferring patients between hospitals, has been relatively under-explored in healthcare optimization. In this paper, we propose solution methods for optimization problems arising in inter-hospital emergency medical transportation in Sri Lanka.

The ambulance routing problem (ARP) involves determining efficient routes for ambulances to reach emergency locations, considering factors such as traffic congestion, geographic constraints, and patient priority. The ARP is a specialized type of vehicle routing problem (VRP) tailored to the unique needs of emergency services. Techniques developed for solving the VRP are often adapted to the ARP, incorporating real-time data such as incident locations and ambulance availability to make dynamic routing decisions. Accordingly, the ARP inherits concepts and methodologies from the broader field of vehicle routing.

Researchers have used different optimization techniques to address the ARP, including mathematical models, heuristic algorithms, and simulation-based approaches [1, 4]. Real-time data, such as traffic information and ambulance availability, has also been integrated to dynamically update routing plans. In some previous studies, the ARP has been generalized as a VRP with pickup and delivery or an open vehicle routing problem [5]. Fundamentally, the performance of an EMTS depends on how efficiently the ARP is solved. Talarico et al. [4] considered the ARP in a disaster-response scenario with a large number of injured patients requiring medical assistance simultaneously. In this scenario, ambulances are used to transport both medical personnel and patients, providing on-site aid to mildly injured individuals while transporting severely injured patients to hospitals.

Other researchers [6] have investigated the green ARP, which adapts to the requirements of the home health care (HHC) centers. In this context, ambulances serve patients with varying requirements within an HHC network. Based on these settings, they proposed an approach to minimize the total traveling cost and greenhouse gases emitted by ambulances.

Recent work has also examined dynamic ambulance routing under richer operational settings. For example, Zidi et al. [7] combined k-means++ clustering with a time-variant multi-objective strength Pareto evolutionary algorithm 2 (SPEA2) for a dynamic ambulance routing problem, highlighting the importance of fast re-optimization under changing demand. This line of work is highly relevant to our setting, but it addresses emergency dispatching with dynamic incident arrivals rather than planned inter-hospital transfers with explicit bed compatibility and mandatory-transfer subsets.

The inter-hospital setting considered in this paper differs from standard emergency ARP settings in four structural aspects: (i) *hospital-level compatibility*, where each patient can only be assigned to hospitals with sufficient care level; (ii) *bed-level receiving capacity*, where receiving capacity is modeled as individual destination beds; (iii) *mandatory-transfer subset*, where a subset of patients must be served due to clinical urgency or carry-over from previous days; and (iv) *centralized scheduled dispatching*,

where all ambulances start from a central depot and scheduled-transfer requests are optimized jointly each day. These characteristics create a tightly coupled selection–assignment–scheduling–routing problem that classical ARP formulations, which focus only on dispatching or routing, do not capture.

We refer to this optimization problem as the inter-hospital ambulance routing problem (IH-ARP). Because IH-ARP simultaneously integrates patient selection, bed assignment, ambulance scheduling, and routing under a clinically motivated min–max objective, it requires a tailored solution framework. The methodological contribution of this study therefore lies in decomposing the problem into well-understood subproblems that can be solved efficiently, while adapting the resulting heuristics to the specific characteristics of the Sri Lankan scheduled-transfer setting.

Our contributions are summarized as follows:

- We formulate IH-ARP as a MIP model that integrates patient selection, patient-bed assignment, ambulance scheduling, and routing while minimizing the worst patient delay (maximum tardiness) relative to MRT.
- We propose three decomposition-based heuristics (ML-MIP, TP-VRP, and RCPSP-VRP) for larger instances where the full MIP becomes computationally difficult.
- We provide computational evidence on both random and real-world instances from Colombo district, and we analyze the strengths and limitations of each method.

Many ambulance routing studies have adopted generic metaheuristic approaches. Compared to such methods, the present decomposition strategy offers three practical advantages. First, it keeps hospital-level feasibility explicit at the assignment stage rather than repairing infeasible routes afterward. Second, it preserves clinical interpretability because the controlling quantity remains tardiness relative to MRT. Third, it enables modular extensions, such as replacing the routing module with a robust VRP-TW solver or embedding rolling-horizon updates when operational disruptions occur. However, other metaheuristics can be attractive for dynamic or highly stochastic settings, although they often require extensive parameter calibration and may provide less transparent clinical guarantees.

The remainder of this paper is organized as follows: Section 2 reviews the most closely related literature. Section 3 presents the real-world background and problem setting. Section 4 introduces the notation and the MIP model. Section 5 describes the proposed heuristic approaches. Section 6 reports the experimental setup and computational results on both random and real-world instances. Finally, Section 7 concludes the paper and discusses directions for future research.

2. Related work

The literature most closely related to IH-ARP lies at the intersection of ambulance routing, vehicle routing with time windows, healthcare transfer planning, and optimization under uncertainty. Standard ARP studies typically prioritize rapid incident response or ambulance redeployment [1, 2], whereas our setting concerns planned inter-hospital transfers in which requests are known before dispatch and hospital receiving capacity is an explicit component of the decision problem. Disaster-response ambulance routing [4] and dynamic home-healthcare routing [6] both motivate timeliness-aware models, but neither captures the combination of bed-level compatibility, mandatory-transfer requests, and centralized same-day dispatching considered here.

From a VRP perspective, IH-ARP is also distinct from classical VRP-TW and pickup-and-delivery models. Transferring a patient from an origin hospital to an admissible receiving bed creates a coupled

assignment–routing structure: before route construction, each patient must be matched to a feasible destination bed, and this match changes both the subsequent route cost and the attainable tardiness. Existing VRP-TW methods [8,9] provide strong routing technology for the second stage of our heuristics, but they do not directly model patient-to-bed assignment or the mandatory-transfer subset.

Uncertainty-aware optimization is also relevant. Robust and stochastic ambulance planning has been studied under uncertain demand locations, traffic conditions, and service availability [10, 11]. Likewise, stochastic and robust VRP variants address travel-time or demand uncertainty through recourse actions or uncertainty sets. Our model is deterministic because the scheduled-transfer problem is solved on a short daily planning horizon using prevalidated bed availability and a fixed ambulance pool. This assumption is most reasonable when traffic conditions are stable enough to allow reliable travel-time estimation and when bed confirmations are made before dispatch. In practice, however, a bed may become unavailable or an ambulance may break down after the plan has been computed, in which case a deterministic plan can degrade quickly. For this reason, the proposed framework should be interpreted as a baseline day-ahead scheduler that can be extended with robust assignment buffers, rolling-horizon re-optimization, or two-stage recourse decisions.

More specifically, a two-stage stochastic extension could assign patients and reserve tentative routes in the first stage, and then revise the realized routing plan after observing travel-time or bed-availability scenarios. A robust optimization variant could instead protect the TP-VRP or RCPSP-VRP assignment modules against interval uncertainty in travel times and receiving capacity by optimizing against the worst case within calibrated uncertainty sets. These extensions are natural next steps because the decomposition structure already separates assignment and routing decisions, making them amenable to scenario-based or robust replacements without changing the overall problem definition.

3. Problem description

Public health is one of the major services provided free of charge to the people of Sri Lanka. Currently, the medical transportation system within the public health service is undergoing improvement. These efforts aim to achieve several objectives, including ensuring high-quality healthcare that meets patient requirements, minimizing potential risks, and responding rapidly to emergency situations.

3.1. Healthcare background in Sri Lanka

In Sri Lanka, each regional healthcare community is organized around its nearest hospital. However, hospital capabilities vary by location, for example, between urban and rural areas. Based on their available facilities, hospitals in Sri Lanka can be broadly categorized into six levels.

- National hospitals: Level 1
- Teaching hospitals: Level 2
- Provincial/district general hospitals: Level 3
- Base hospitals: Level 4
- Divisional hospitals: Level 5
- Primary medical care units: Level 6.

As the level number increases, the availability of space, specialist personnel, intensive care, and surgical facilities decreases. Patients are typically admitted to the nearest hospital or to a hospital at the preferred

level of care. In most cases, patients seek treatment at their nearest accessible hospital. However, based on the recommendations of medical experts, patients may be transferred to higher-level hospitals with better facilities, depending on their current condition. Transfers from higher-level hospitals to lower-level hospitals are uncommon.

In this study, we focus on the Colombo district, which has the highest population density in Sri Lanka. There are 45 hospitals in this region, covering all levels listed above.

3.2. Transfer process

Patient transfers can be categorized into two types: immediately incurred transfers (IITs) and scheduled transfers (STs). An IIT refers to a patient transfer caused by an unpredictable event such as an accident. Each hospital in Sri Lanka is equipped with one or more ambulances for emergency use, enabling it to handle IIT cases effectively. In the ST setting, patients requiring transfer are identified in advance. Information such as candidate hospitals, travel times between hospitals, and other medical recommendations for each patient is also determined in advance, typically on the day before routing. Unlike IIT cases, ambulances for ST cases in Sri Lanka are dispatched from a central depot with a fixed fleet size.

Centralized dispatching in ST scenarios allows more efficient ambulance allocation. Consequently, solving the corresponding optimization problem has become an important part of the Sri Lankan EMTS. The target patients for ST are those who cannot receive adequate treatment at their current hospitals. ST demand may not be fully satisfied because of the capacity constraints of receiving hospitals or the limited availability of ambulances. From these requests, a subset is selected and scheduled at predetermined start times each day. As a result, the optimization problem is solved on a daily basis.

In both ST and IIT scenarios, medical experts recommend that the minimum risk time (MRT) for each patient should not be exceeded. MRT refers to the latest allowable time by which a patient should arrive at the receiving hospital, beyond which the patient's condition may deteriorate. Accordingly, MRT is a critical factor in this case study. Therefore, the ARP in Sri Lanka involves a time-window constraint for each patient transfer.

The operational distinction between IIT and ST is central to this study. In the Sri Lankan EMTS, IIT cases are event-driven, geographically uncertain, and dispatched immediately by hospital-based emergency fleets. By contrast, ST cases are planned requests collected before daily dispatching, and they are handled jointly by a dedicated fleet that starts from a central depot. Because these two transfer types use separate ambulance systems in Sri Lanka, their associated optimization problems are largely independent. This separation allows us to focus on the ST setting without explicitly modeling random emergency arrivals. At the same time, centralized dispatching creates strong interactions among patient-specific constraints: Each selected patient competes for the same ambulance pool while also requiring a clinically admissible receiving bed and an arrival time that does not exceed the patient's MRT.

In this study, we therefore focus on the ST scenario and develop an optimization-based solution framework for the inter-hospital ambulance routing system.

3.3. Objective and mandatory transfers

The objective of IH-ARP is to minimize the maximum tardiness. This min–max criterion is motivated by discussions with practitioners, who emphasized that, in the scheduled-transfer process, avoiding

extremely late transfers is the most critical operational target: Reducing one patient's very large delay is operationally more important than slightly improving several transfers that are already safe. Alternative criteria, such as total tardiness or weighted tardiness, are also relevant and could be examined in future sensitivity analyses, but they may obscure extreme outcomes for the most time-critical patients.

Among the patient-transfer requests, we distinguish a mandatory-transfer subset that must be served. This subset may represent patients prioritized because of clinical urgency, referral rules, or carry-over obligations from previous days. In addition, actual deployment decisions remain subject to physician approval and patient consent. Although these considerations are not modeled as decision variables, they motivate the mandatory-service constraints and reinforce the interpretation of the computed plan as a decision-support tool rather than an autonomous dispatch rule.

4. Mathematical programming

We first formulate a MIP model for the patient transfer system that minimizes the maximum tardiness relative to MRT. A fixed number of ambulances is available at the depot. Depending on the transfer demand, all ambulances may be occupied at some times, whereas at others, some ambulances may remain idle.

In IH-ARP, we are given a set of patients $V_s = \{1, 2, \ldots, m\}$ located at lower-level hospitals and requiring transfer to higher-level hospitals, a set of receiving beds $V_d = \{m + 1, m + 2, \ldots, m + n\}$, and a set of ambulances $K = \{1, 2, \ldots, k_{\max}\}$. Figure 1 illustrates the placement of patients and receiving beds across hospitals. Receiving capacity is modeled at the bed level, so each selected patient is transferred from the origin hospital to an individual bed in a higher-level hospital. The ambulance depot is denoted by 0. Each patient $i \in V_s$ is associated with an MRT d_i and requires a bed at level l_i or better, where a smaller level indicates higher capability. Similarly, each bed $j \in V_d$ is associated with its hospital level l_j. Multiple patients and multiple beds may belong to the same hospital; see, for example, the hospital on the left in Figure 1.

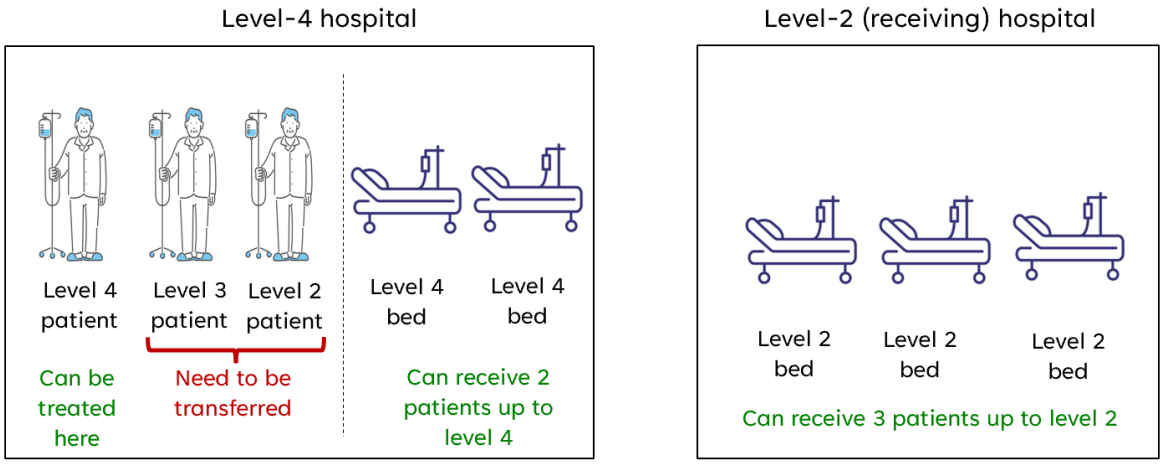

Figure 1. An example of patients and receiving beds.

If patient i is selected, the patient must be transferred to a bed j whose hospital level satisfies $l_j \le l_i$. The arrival time at the receiving hospital should not exceed the patient's MRT d_i; otherwise, positive

tardiness is incurred. A subset $F \subset V_s$ represents mandatory-transfer patients who must be transferred because of urgent medical needs or carry-over from previous days.

We denote by c_{ij} the travel time from the origin hospital of patient $i \in V_s$ to bed $j \in V_d$ in a receiving hospital. Similarly, c_{ji} denotes the return travel time from bed $j \in V_d$ to the origin hospital of patient $i \in V_s$. The travel times from the depot to the origin hospital of patient $i \in V_s$ and from bed $j \in V_d$ to the depot are denoted by c_{0i} and c_{j0}, respectively.

All ambulances are assumed to have homogeneous travel times. According to the situation in Sri Lanka, the number of patients requiring transfer typically exceeds the number of available beds at each level l', that is, $|\{i \in V_s \mid l_i = l'\}| > |\{j \in V_d \mid l_j = l'\}|$.

Input data, including travel times, bed availabilities, service times, and ambulance availability, are treated as deterministic. This approximation is reasonable for short day-ahead planning horizons, in which receiving hospitals confirm available beds before dispatch, and travel times are based on stable historical or map-based estimates. Nevertheless, it may understate operational risk when congestion, emergency admissions, or ambulance breakdowns occur. In practice, a simple contingency mechanism is to re-optimize the second-stage routing problem in a rolling horizon whenever a disruption invalidates the current plan, while preserving transfers that have already begun.

The objective of IH-ARP is to construct an ambulance routing plan that minimizes the maximum tardiness over all transferred patients. For clarity, the tardiness τ_i of patient i relative to MRT d_i is defined as

$$\tau_i = \max\{0, a_i - d_i\},$$

where a_i is the arrival time of patient i at the assigned receiving bed. The maximum tardiness is then defined by

$$T_{\max} = \max_{i \in V_s} \tau_i.$$

Thus, the model limits the worst violation of the clinically recommended latest arrival time.

From the input data, we construct a directed graph $G = (V, E)$, where

- $V = \{0\} \cup V_s \cup V_d$,
- $E = E_{sd} \cup E_{ds} \cup \{(0, j) \mid j \in V_s\} \cup \{(i, 0) \mid i \in V_d\} \cup \{(0, 0)\}$, where
  - $E_{sd} = \{(i, j) \mid i \in V_s, j \in V_d, l_j \le l_i\}$
  - $E_{ds} = \{(i, j) \mid i \in V_d, j \in V_s\}$.

In the model, the arc $(0, 0)$ represents an ambulance that remains idle at the depot and is not dispatched for patient transfers. We use c_{sum} to denote the total travel time over all arcs, that is, $c_{\text{sum}} = \sum_{(i,j)\in E} c_{ij}$.

We now introduce the decision variables used in the model. The binary variable x_{ijk} indicates whether ambulance k traverses arc (i, j). We use q_i to denote the departure time associated with patient i, q_j to denote the arrival time at bed j, and $T_{\max}$ to denote the maximum tardiness.

Table 1 summarizes the main notations used throughout the paper, including notations introduced in later sections.

With these input and variables, IH-ARP is formulated as follows:

$$\min \quad T_{\max} \tag{4.1}$$

Table 1. Notation list.

	input data
V_s	set of patients requiring transfer
V_d	set of receiving beds
K	set of ambulances stationed at the central depot
F	mandatory-transfer subset of patients
l_i, l_j	care level required by patient i / offered by bed j
d_i	minimum risk time (MRT) of patient i
c_{ij}	travel time on arc (i, j)
	decision variables
x_{ijk}	1 if ambulance k traverses arc (i, j); 0 otherwise
q_i	dispatch or service-start time associated with node i
y_{ij}	1 if patient i is assigned to bed j; 0 otherwise
z_{it}	1 if patient i occupies an ambulance at time t; 0 otherwise
τ_i	tardiness of patient i relative to MRT
$T_{\max}$	maximum tardiness over all transferred patients

$$\text{s.t.} \quad \sum_{i\in V} x_{jik} = \sum_{i\in V} x_{ijk} \qquad \forall j \in V, \forall k \in K \tag{4.2}$$

$$\sum_{k\in K}\sum_{j\in V_d} x_{ijk} \le 1 \qquad \forall i \in V_s \setminus F \tag{4.3}$$

$$\sum_{k\in K}\sum_{j\in V_d} x_{ijk} = 1 \qquad \forall i \in F \tag{4.4}$$

$$\sum_{k\in K}\sum_{i\in V_s} x_{ijk} = 1 \qquad \forall j \in V_d \tag{4.5}$$

$$\sum_{i\in V_s\cup\{0\}} x_{0ik} = 1 \qquad \forall k \in K \tag{4.6}$$

$$q_0 = 0 \tag{4.7}$$

$$q_j \ge q_i + c_{ij} - (1 - x_{ijk})c_{\text{sum}} \qquad \forall (i, j) \in E_{ds} \cup E_{sd}, \forall k \in K \tag{4.8}$$

$$T_{\max} \ge q_i + \sum_{k\in K}\sum_{j\in V_d} c_{ij}x_{ijk} - d_i - c_{\text{sum}}\left(1 - \sum_{k\in K}\sum_{j\in V_d} x_{ijk}\right) \qquad \forall i \in V_s \tag{4.9}$$

$$q_i \ge 0 \qquad \forall i \in V \tag{4.10}$$

$$T_{\max} \ge 0 \tag{4.11}$$

$$x_{ijk} \in \{0, 1\} \qquad \forall (i, j) \in E, \forall k \in K. \tag{4.12}$$

Constraints (4.2) enforce flow conservation. Constraints (4.3) and (4.4) ensure that patients in F must be transferred, whereas all other patients can be assigned to at most one bed. Constraints (4.5) enforce that each receiving bed is matched with exactly one transferred patient. This equality is intentional because the planning process predefines the set of active receiving beds for the day, and these active beds are required to be fully utilized. Constraint (4.6) states that each ambulance departs from the depot, and (4.7) sets the start time. Constraints (4.8) define arrival times at each location. Constraints (4.9)

ensure that $T_{\max}$ captures delays beyond MRT d_i. Constraints (4.10)–(4.12) define the domains of the decision variables.

5. Heuristic approaches

The model proposed in Section 4 is deterministic and should therefore be interpreted as a baseline planning tool rather than as a fully robust real-time control policy. A natural stochastic extension would treat uncertain travel times or bed cancellations in a second stage with recourse, whereas a robust extension would protect the assignment and routing decisions against bounded uncertainty in c_{ij} or active-bed availability. However, even in its deterministic form, model (4.1)–(4.12) becomes difficult to solve with a commercial MIP solver on medium-to-large instances, as shown by the experimental results in Section 6.

Accordingly, this section presents three decomposition-based heuristic approaches—ML-MIP, TP-VRP, and RCPSP-VRP—for obtaining high-quality solutions to instances with up to 132 patients. Beyond their computational role, these frameworks also provide a natural basis for future extensions to on-site uncertainty: TP-VRP can incorporate robust assignment costs or scenario-dependent feasibility checks, and RCPSP-VRP can account for uncertainty through protected processing times or ambulance-capacity buffers.

In the first approach, ML-MIP, we fix the set of patients to be transferred and solve the remaining problem using a simplified MIP model. The second and third approaches, TP-VRP and RCPSP-VRP, share the same overall structure: In the first phase, patients are assigned to suitable hospitals, and in the second phase, ambulance routes are constructed to minimize the maximum tardiness of the system.

These three heuristics play different roles. ML-MIP serves as a learning-assisted reduction of the original decision space; TP-VRP is the simplest decomposition baseline with an explicit assignment parameter; and RCPSP-VRP is our most competitive approach because it anticipates ambulance scarcity already in the assignment stage. We therefore regard the latter as the strongest practical method, while the first two methods provide interpretable baselines and help explain which modeling ingredients are most valuable.

5.1. Approach using machine learning

The first heuristic approach, ML-MIP, utilizes machine learning to identify patients for transfer. In the first phase, we apply binary classification to distinguish between patients who are assigned and those who are not. The receiving hospitals of the assigned patients are not determined at this stage. We test several classification algorithms, including logistic regression (LR) [12], gradient boosting (GB) [13], random forest (RF) [14], support vector machines (SVM) [15], and artificial neural networks (ANNs) [16].

We select the classifier with the highest accuracy to determine the patients to be transferred. To construct the training data, we solve 50 small-scale instances using the MIP model (4.1)–(4.12) and use the resulting patient-selection decisions as labels. Let γ_i denote the predicted probability for each test sample $i \in V_s \setminus F$. Using these probabilities, we define a subset of patients V_s' by taking all patients in F together with the $|V_d| - |F|$ patients in $V_s \setminus F$ that have the largest values of γ_i. Thus, after the first phase, the selected set contains at most as many patients as there are available beds.

Let V' and E' denote the node and arc sets induced by V_s':

- $V' = \{0\} \cup V'_{\mathrm{s}} \cup V_{\mathrm{d}}$,
- $E' = E'_{\mathrm{sd}} \cup E'_{\mathrm{ds}} \cup \{(0, j) \mid j \in V'_{\mathrm{s}}\} \cup \{(i, 0) \mid i \in V_{\mathrm{d}}\} \cup \{(0, 0)\}$, where
  - $E'_{\mathrm{sd}} = \{(i, j) \mid i \in V'_{\mathrm{s}}, j \in V_{\mathrm{d}}, l_j \le l_i\}$
  - $E'_{\mathrm{ds}} = \{(i, j) \mid i \in V_{\mathrm{d}}, j \in V'_{\mathrm{s}}\}$.

In the second phase, we solve the following MIP based on the graph (V', E'):

$$\min \quad T_{\max} \tag{5.1}$$

$$\text{s.t.} \quad \sum_{i \in V'} x_{jik} = \sum_{i \in V'} x_{ijk} \qquad \forall j \in V', \forall k \in K \tag{5.2}$$

$$\sum_{k \in K} \sum_{j \in V_{\mathrm{d}}} x_{ijk} = 1 \qquad \forall i \in V'_{\mathrm{s}} \tag{5.3}$$

$$\sum_{k \in K} \sum_{i \in V'_{\mathrm{s}}} x_{ijk} = 1 \qquad \forall j \in V_{\mathrm{d}} \tag{5.4}$$

$$\sum_{i \in V'_{\mathrm{s}} \cup \{0\}} x_{0ik} = 1 \qquad \forall k \in K \tag{5.5}$$

$$q_0 = 0 \tag{5.6}$$

$$q_j \ge q_i + c_{ij} - (1 - x_{ijk}) c_{\mathrm{sum}} \qquad \forall (i, j) \in E'_{\mathrm{ds}} \cup E'_{\mathrm{sd}}, \forall k \in K \tag{5.7}$$

$$T_{\max} \ge q_i + \sum_{k \in K} \sum_{j \in V_{\mathrm{d}}} c_{ij} x_{ijk} - d_i \qquad \forall i \in V'_{\mathrm{s}} \tag{5.8}$$

$$q_i \ge 0 \qquad \forall i \in V' \tag{5.9}$$

$$T_{\max} \ge 0 \tag{5.10}$$

$$x_{ijk} \in \{0, 1\} \qquad \forall (i, j) \in E', \forall k \in K. \tag{5.11}$$

Model (5.1)–(5.11) is a simplified version of the original model (4.1)–(4.12) obtained by fixing the selected patients. Because all patients in V'_{s} are required to be transferred, one of constraints (5.3) and (5.4) is redundant. We nevertheless retain both constraints for ease of comparison with the original model and for readability. By treating the selected patients as fixed input, the feasible region is reduced, which accelerates computation while preserving the routing and assignment decisions for the preselected patients.

Algorithm 1 summarizes ML-MIP. Its computational complexity is dominated by the second-stage MIP model, while the prediction stage is negligible once the classifier is trained. Hence, ML-MIP is attractive only when fixing the selection variables removes enough combinatorial difficulty to compensate for a possible classification error.

5.2. Approach based on transportation problem

In the second approach, TP-VRP, patients are assigned to hospitals in the first phase using a modified assignment cost c'_{ij} that combines the travel time c_{ij} between patient i and receiving bed j with the patient's MRT d_i as follows:

$$c'_{ij} = c_{ij} - \alpha d_i \qquad (i, j) \in E_{\mathrm{sd}}, \tag{5.12}$$

Algorithm 1 ML-MIP heuristic.

1: Train candidate binary classifiers on solved small instances.
2: Select the classifier with the best validation accuracy.
3: Predict transfer probabilities γ_i for $i \in V_s \setminus F$.
4: Build V_s' from F and the highest-probability patients until $|V_s'| = |V_d|$.
5: Solve the reduced MIP on (V', E').
6: **Output:** the best feasible reduced-model solution.

where α is a parameter determined through preliminary empirical testing. If α is set to 0, the modified assignment cost reduces to the original travel cost. Conversely, if α is assigned a large value, patients with smaller MRT values are favored more strongly in the assignment stage.

We use the minus sign in (5.12) so that smaller MRT values effectively receive higher priority in the assignment stage. The tested values $\alpha \in \{0, 0.5, 1, 1.5, 2, \infty\}$ were chosen to span the spectrum from pure distance minimization ($\alpha = 0$) to urgency-dominated assignment ($\alpha \to \infty$). In practice, $\alpha = 0$ performed best on average (see the results in Section 6), which suggests that explicit routing feasibility in the second phase already captures much of the urgency effect.

Using a binary variable y_{ij}, where $y_{ij} = 1$ indicates that patient i is assigned to receiving bed j, the first phase can be formulated as follows:

$$\min \quad \sum_{(i,j)\in E_{sd}} c'_{ij} y_{ij} \tag{5.13}$$

$$\text{s.t.} \quad \sum_{j\in V_d} y_{ij} \le 1 \qquad \forall i \in V_s \setminus F \tag{5.14}$$

$$\sum_{j\in V_d} y_{ij} = 1 \qquad \forall i \in F \tag{5.15}$$

$$\sum_{i\in V_s} y_{ij} = 1 \qquad \forall j \in V_d \tag{5.16}$$

$$y_{ij} \in \{0, 1\} \qquad \forall (i, j) \in E_{sd}. \tag{5.17}$$

Constraints (5.14) and (5.15) play the same roles as constraints (4.3) and (4.4), respectively: Patients outside F can be assigned to at most one bed, whereas patients in F must be assigned to exactly one bed. Constraints (5.16) ensure that hospital capacity is fully utilized, and constraints (5.17) define the domain of the decision variable.

Problem (5.13)–(5.17) has the structure of a transportation problem (TP), in which supply (patients to be transferred) is matched to demand (hospital capacity) while minimizing transportation cost. It can therefore be solved in polynomial time by algorithms such as the network simplex method or the transportation algorithm [17]. To ensure that total supply equals total demand, we introduce a dummy destination node $j^\dagger$ representing the non-assignment of optional patients when physical receiving capacity is insufficient. By assigning zero cost to arcs $(i, j^\dagger)$ and adding the balance constraint $\sum_{i\in V_s} y_{ij^\dagger} = |V_s| - |V_d|$, we keep the transportation model balanced while preserving the interpretation that only patients assigned to real beds proceed to the second-stage routing problem. This TP structure, together with the dummy node, makes the first phase computationally efficient.

In the second phase, we treat the patient-bed assignments produced in the first phase as *task nodes*. Each task node v_{ij} represents the transfer of patient i to bed j. We then construct a complete directed graph on these task nodes together with a special depot node v_{00}. The cost on each arc $(v_{ij}, v_{i'j'})$ is defined as $c_{ji'} + c_{i'j'}$. This graph is used as input to a vehicle routing problem with time windows (VRP-TW), where the time window of node v_{ij} is $[0, d_i + T]$ for a fixed tardiness threshold T.

VRP-TWs are well studied, and many effective algorithms and solvers are available [8]. However, our objective, minimizing the maximum tardiness, is less standard. To exploit powerful VRP-TW solvers such as PyVRP [9], we perform a binary search over the tardiness threshold T. The initial bounds are set to $T^{\mathrm{L}} = 0$ and $T^{\mathrm{U}} = \max_{i \in V_{\mathrm{s}}} \left\{ \max_{j:(i,j) \in E_{\mathrm{sd}}} c_{ij} - d_i \right\}_+ + t_{\max}$, where $(x)_+ = \max\{0, x\}$. At each iteration, we solve a VRP-TW feasibility problem with time windows $[0, d_i + T]$ at $T = (T^{\mathrm{L}} + T^{\mathrm{U}})/2$. If the problem is feasible, we update $T^{\mathrm{U}} \leftarrow T$; otherwise, we set $T^{\mathrm{L}} \leftarrow T + 1$ for integer-minute data. This process continues until $T^{\mathrm{L}} = T^{\mathrm{U}}$, which yields the minimum feasible tardiness threshold.

Algorithm 2 summarizes TP-VRP. The first phase is polynomial in the size of the assignment graph, whereas the second phase repeatedly solves a VRP-TW feasibility problem inside a binary search over tardiness thresholds. The overall runtime is therefore driven by the number of VRP-TW feasibility checks and the difficulty of each routing call.

Algorithm 2 TP-VRP heuristic.

Choose α and solve the transportation model with dummy destination $j^{\dagger}$.
Keep only patient-bed pairs assigned to real beds.
Build task nodes v_{ij} and the induced routing graph.
Initialize tardiness bounds T^{L} and T^{U}.
while $T^{\mathrm{L}} < T^{\mathrm{U}}$ **do**
    Set $T \leftarrow \lfloor (T^{\mathrm{L}} + T^{\mathrm{U}})/2 \rfloor$.
    Solve the VRP-TW feasibility problem with windows $[0, d_i + T]$.
    if feasible **then**
        $T^{\mathrm{U}} \leftarrow T$.
    else
        $T^{\mathrm{L}} \leftarrow T + 1$.
    end if
end while
Output: the routes associated with the minimum feasible tardiness threshold.

5.3. Approach incorporating resource-constrained project scheduling problem

The TP-VRP approach decomposes the original model into an assignment phase and a routing phase. However, the assignment phase completely ignores the limit on the number of ambulances, and the parameter α must be tuned. To address these two issues, we propose a third approach.

In this approach, we replace the TP in the first phase with a resource-constrained project scheduling problem (RCPSP) to obtain patient-hospital assignments. A related idea appears in [18], where RCPSP with routing extends the traditional RCPSP by incorporating the movement of resources between activities using a limited fleet of vehicles. In our approach, we adapt the RCPSP framework to assign patients to suitable beds while minimizing the maximum tardiness in the system.

Let $\Xi = \{0, 1, \ldots, t_{\max}\}$ denote the time horizon, and let $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)$ denote the vector of processing times, where p_i is the processing time of patient (job) i; further details are given below. We use the same binary variable y_{ij} as in model (5.13)–(5.17) to represent the assignment of patient i to bed j, and the same continuous variable q_i as in model (4.1)–(4.12) to denote the departure time of patient i. In addition, we introduce a binary variable z_{it} indicating whether patient i uses an ambulance at time t. The model RCPSP($\boldsymbol{p}$), which depends on $\boldsymbol{p}$, is formulated as follows:

$$\min \quad c_{\text{sum}} T_{\max} + \sum_{(i,j)\in E_{\text{sd}}} c_{ij} y_{ij} \tag{5.18}$$

$$\text{s.t.} \quad T_{\max} \geq q_i + p_i - d_i - t_{\max}\left(1 - \sum_{j\in V_{\text{d}}} y_{ij}\right) \qquad \forall i \in V_{\text{s}} \tag{5.19}$$

$$\sum_{i\in V_{\text{s}}} z_{it} \leq k_{\max} \qquad \forall t \in \Xi \tag{5.20}$$

$$q_i + 1 - t_{\max}(1 - z_{it}) \leq t \leq q_i + p_i - 1 + t_{\max}(1 - z_{it}) \qquad \forall i \in V_{\text{s}}, \forall t \in \Xi \tag{5.21}$$

$$q_i \geq 0 \qquad \forall i \in V_{\text{s}} \tag{5.22}$$

(5.14)–(5.17).

The objective function (5.18) is formulated as a single-objective function. The second term, representing the total assignment cost, acts solely as a scalarized support term (with a small relative weight inherently provided by c_{sum}) to break ties among assignments with the same maximum tardiness, thereby guiding the solver toward more efficient routing structures.

Constraints (5.20) ensure the maximum number of ambulances used at any time in the time horizon Ξ does not exceed the available limit. Constraints (5.21) link q_i and z_{it} through a standard big-M linearization: when $z_{it} = 1$, time t must lie in the active interval $[q_i + 1, q_i + p_i - 1]$; when $z_{it} = 0$, the inequalities are relaxed by $t_{\max}$.

In this study, we set $p_i = 2 \min_{j:\,(i,j)\in E_{\text{sd}}} c_{ij}$ in $\boldsymbol{p}$ to approximate the round-trip travel time to the nearest available bed. After solving RCPSP($\boldsymbol{p}$), we obtain an optimal assignment $\boldsymbol{y}^*$. Based on these patient-bed assignments, we then determine the routes by solving VRPTW($\boldsymbol{y}^*$) using the same second-phase procedure as in the TP-VRP approach.

Although this approach already provides satisfactory solutions, it can be improved further. We therefore extend it to an iterative framework.

Let Y be the set of assignment vectors generated in previous iterations, initially empty. In each iteration, we solve RCPSP($\boldsymbol{p}, Y$) by adding the following constraint to RCPSP($\boldsymbol{p}$):

$$\sum_{(i,j)\in E_{\text{sd}}:\, y'_{ij}=0} y_{ij} \geq 1 \qquad \forall \boldsymbol{y}' \in Y.$$

This constraint ensures that the assignment obtained in the current iteration differs from all assignments found previously. Algorithm 3 summarizes the resulting iterative framework.

The strength of RCPSP-VRP comes from representing ambulance contention already in the assignment phase through the resource constraints (5.20)–(5.21). This helps explain why the method is especially effective on the largest instances, where assignment decisions that ignore ambulance scarcity can drive the second-stage routing problem toward poor or infeasible regions.

Algorithm 3 The iterated approach combining RCPSP and VRP-TW.

$\boldsymbol{p} \leftarrow \boldsymbol{p}^{\text{init}}$, where $p_i^{\text{init}} = 2 \min_{j:\,(i,j)\in E_{\text{sd}}} c_{ij}$.
$Y \leftarrow \emptyset$.
Set incumbent $v_{\text{incumbent}}$ as $v_{\text{incumbent}} \leftarrow +\infty$.
while the time limit is not reached **do**
    Solve RCPSP($\boldsymbol{p}, Y$), and obtain an optimal solution $\boldsymbol{y}^*$.
    $Y \leftarrow Y \cup \{\boldsymbol{y}^*\}$.
    Solve VRP-TW($\boldsymbol{y}^*$) using binary search based on the assignments $\boldsymbol{y}^*$, and obtain an optimal set of routes $\boldsymbol{\sigma}^*$ as well as its maximum tardiness v^*.
    For each patient i appearing in $\boldsymbol{\sigma}^*$, update p_i in $\boldsymbol{p}$ as the sum of the travel time from the previous location to i and the transfer time for i.
    if $v_{\text{incumbent}} > v^*$ **then**
        $v_{\text{incumbent}} \leftarrow v^*$, and $\boldsymbol{\sigma}_{\text{incumbent}} \leftarrow \boldsymbol{\sigma}^*$
    end if
end while
Output: $v_{\text{incumbent}}$ and $\boldsymbol{\sigma}_{\text{incumbent}}$.

6. Computational experiments

This section describes the test instances and presents the computational results for the proposed model and heuristic approaches.

6.1. Instance description

To evaluate the performance of the MIP model and the heuristic algorithms, we generated two types of datasets: randomly generated instances and instances derived from real-world data.

Real-world (*real*) instances were constructed from operational data to reflect the complexities and constraints of the patient transfer system in the Colombo district. This district comprises one level-1 hospital, five level-2 hospitals, one level-3 hospital, two level-4 hospitals, nine level-5 hospitals, and 27 level-6 hospitals. Random (*rand*) instances are synthetic datasets generated to evaluate the performance of the MIP model and heuristic algorithms in controlled but still challenging settings. For these instances, we randomly selected 15 to 25 patient-transfer requests from large-scale real-world instances and retained the corresponding level-compatibility rules, travel-time matrix entries, and ambulance counts. These random instances can be harder than some real instances because random sampling may produce concentrated urgent requests (small MRT values) together with tight bed compatibility, thereby increasing combinatorial difficulty.

The random-instance generation process can be summarized as follows: (i) choose a source real-world day, (ii) sample 15–25 requests without replacement, (iii) retain the original admissible receiving beds and ambulance count associated with that sampled subset, and (iv) verify feasibility of the mandatory-transfer set against the active receiving beds before solving the instance. These instances are intentionally small because they were also used to generate exact labels for ML training. We acknowledge that this design is better suited for controlled validation than for large-scale synthetic stress testing, and we identify broader random-instance generation as an important extension.

Detailed information on the random and real-world instances is given in Table 2. Each instance is denoted by "*wwww-xx-yy-z*," where type = *wwww* specifies the instance type, $m = xx$ indicates the number of patients, $\sum b_j = yy$ is the total receiving capacity (beds), and $k_{\max} = z$ represents the number of ambulances. For all instances, $t_{\max}$ is set to 480 minutes (8 working hours).

Across all tested instances, the number of patients ranges from 15 to 132, total active receiving beds from 14 to 129, and the number of ambulances from 3 to 24. The planning horizon is one working day (480 minutes). In the real-world data, ST requests are collected before dispatch and are paired with expert-provided MRT values that define the latest desirable arrival times. IIT requests are not part of the computational dataset because they are handled by a separate emergency fleet.

Table 2. Instance information.

name	type	m	$\sum b_j$	$k_{\max}$
rand-15-14-3	rand	15	14	3
rand-17-16-3	rand	17	16	3
rand-20-18-4	rand	20	18	4
real-24-17-4	real	24	17	4
real-24-19-5	real	24	19	5
rand-25-22-5	rand	25	22	5
rand-25-24-6	rand	25	24	6
real-25-17-4	real	25	17	4
real-25-22-5	real	25	22	5
real-26-24-6	real	26	24	6
real-45-42-9	real	45	42	9
real-70-66-14	real	70	66	14
real-76-74-16	real	76	74	16
real-80-77-16	real	80	77	16
real-96-88-19	real	96	88	19
real-96-90-22	real	96	90	22
real-96-94-22	real	96	94	22
real-110-104-22	real	110	104	22
real-126-121-23	real	126	121	23
real-132-129-24	real	132	129	24

To train the machine learning models described in Section 5.1, we constructed an additional dataset from 50 small-scale instances, each with 15–25 patients and generated in the same way as the *rand* instances. The corresponding optimal solutions were obtained from the MIP model. These instances yielded 1236 training records, each representing an individual patient. The dataset included the following features:

- the MRT of each patient,
- the distance from the patient to the depot,
- the distances from the patient to beds in each receiving hospital (if a receiving hospital cannot meet the required level for the patient, the corresponding distance is set to infinity),
- the capacity of each hospital,

- the ratio of the number of available ambulances to the total number of patients.

6.2. Implementation details

All algorithms and mathematical models were implemented in Python 3.10.7 and executed on a PC with an 11th Gen Intel(R) Core(TM) i7-1185G7 processor (3.00 GHz) and 16 GB of memory. All mathematical programming models were solved using `Gurobi Optimizer` (version 13.0.1). For the binary classification models, we used the open-source Python library `scikit-learn` (version 1.3.2). For ML-MIP, we report results for RF because it achieved the highest validation accuracy among LR, GB, RF, SVM, and ANN. To solve the VRP-TW subproblems arising in the binary search, we used `PyVRP` (version 0.9.1) with a stopping criterion that terminates the search after 10 consecutive iterations without improvement of the incumbent solution. All computational experiments were run with a time limit of 300 seconds.

6.3. Experimental results

We begin with the computational results for the MIP model described in Section 4. Table 3 reports the running time (column "time") in seconds and the corresponding objective value (column "obj") obtained within the time limit for the random instances. The notation "t.l." indicates that the instance was not solved to optimality within the time limit.

Table 3. Results obtained by MIP model on random instances.

instance	time	obj
rand-15-14-3	73.8	0
rand-17-16-3	t.l.	20
rand-20-18-4	t.l.	54
rand-25-22-5	t.l.	3
rand-25-24-6	t.l.	42

The MIP model obtained a proven optimal solution for only one instance, namely the case with 15 patients and 3 ambulances. For the remaining random instances, all of which had more than 15 patients and beds in total, the MIP model failed to prove optimality within the time limit, which motivates the use of heuristic approaches. This table is included only to illustrate the computational capacity and limitations of the original MIP formulation on random instances. At this stage, our goal is not to claim strong performance on the random set but to show how quickly the exact model becomes challenging as instance size increases, thereby motivating the need for heuristic methods in the subsequent experiments. Because of constraints (4.5), (5.4), and (5.16), the number of transferred patients is fixed at $|V_d|$ in all tested solutions; therefore, the number of non-transferred patients is $m - |V_d|$ for each instance.

Parameter tuning for the second heuristic approach (TP-VRP), described in Section 5.2, was performed over six values of α on 15 real-world instances. The results are summarized in Table 4.

Table 4. Results of the TP-VRP approach based on different values of α.

	$\alpha = 0$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2$		$\alpha = +\infty$	
instance	time	obj	time	obj	time	obj	time	obj	time	obj	time	obj
real-25-17-4	0.6	0	0.5	0	0.6	0	0.7	12	0.6	14	0.5	36
real-25-22-5	0.6	0	0.5	11	0.6	0	0.7	0	0.6	0	0.6	0
real-24-17-4	0.5	0	0.6	0	0.6	0	0.5	0	0.5	0	0.6	15
real-24-19-5	0.5	0	0.6	0	0.6	0	0.6	0	0.5	0	0.6	0
real-26-24-6	0.6	0	0.5	0	0.7	0	0.6	0	0.6	0	0.6	5
real-45-42-9	0.8	0	0.8	0	0.8	0	0.9	0	0.7	0	0.8	0
real-70-66-14	1.1	0	1.3	0	1.3	0	1.2	0	1.3	0	1.1	0
real-76-74-16	1.6	0	1.4	0	1.7	0	1.3	0	1.8	0	1.5	1
real-80-77-16	1.3	0	2.0	0	1.4	0	1.7	0	1.5	0	1.2	0
real-96-88-19	2.2	0	2.4	0	2.3	13	2.3	6	1.6	12	1.6	14
real-96-90-22	1.7	0	2.3	0	1.8	0	1.8	0	2.0	0	2.1	0
real-96-94-22	2.2	0	2.6	0	2.3	0	2.1	0	2.6	0	1.7	0
real-110-104-22	2.6	0	2.7	4	2.3	4	2.5	22	2.3	9	2.5	21
real-126-121-23	3.4	50	3.1	38	2.5	39	2.7	43	2.1	45	3.2	54
real-132-129-24	3.5	50	3.3	51	3.0	52	3.2	52	3.4	55	3.2	58

The notation "$\alpha = +\infty$" indicates that assignment priorities are determined solely by MRT, that is, pairs are ranked by increasing d_i rather than by travel time. All other notations are consistent with Table 3. Table 4 shows that TP-VRP generally performs better for smaller values of α than for larger ones. For readability, Table 5 also reports the best objective value for each instance and the absolute gap of each α setting from that best value. For instance, an optimal solution for real-25-22-5 was obtained when $\alpha \in \{0, 1, 1.5, 2\}$, which indicates that the best setting of α depends on the instance. Because TP-VRP terminates in less than 3 seconds for any fixed value of α, far below the 300-second time limit, we treat TP-VRP as a six-setting multi-start procedure over $\alpha \in \{0, 0.5, 1, 1.5, 2, +\infty\}$ and report the best objective value together with the total running time in Table 6.

Finally, we compare the results on 15 real-world instances obtained by the proposed mixed integer programming model (MIP) in Section 4 with those obtained by the three heuristic approaches introduced in Section 5: ML-MIP (Section 5.1), TP-VRP (Section 5.2), and RCPSP-VRP (Section 5.3). The notation "—" indicates that no feasible solution was found within the time limit.

To improve comparability across methods, Table 7 reports the best objective value per instance and the absolute optimality gap of each method from that best value.

Table 6 shows that MIP and ML-MIP perform similarly in both running time and objective value. Both approaches obtained optimal solutions for five instances with up to 70 patients. However, for all instances with more than 70 patients, neither MIP nor ML-MIP found a feasible solution within the 300-second time limit. This suggests that reducing the feasible region by fixing only the binary patient-selection variables is not sufficient to improve computational performance substantially.

Table 5. Best objective value and absolute objective gaps for TP-VRP parameter settings.

		absolute gap					
instance	best	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = \infty$
real-25-17-4	0	0	0	0	12	14	36
real-25-22-5	0	0	11	0	0	0	0
real-24-17-4	0	0	0	0	0	0	15
real-24-19-5	0	0	0	0	0	0	0
real-26-24-6	0	0	0	0	0	0	5
real-45-42-9	0	0	0	0	0	0	0
real-70-66-14	0	0	0	0	0	0	0
real-76-74-16	0	0	0	0	0	0	1
real-80-77-16	0	0	0	0	0	0	0
real-96-88-19	0	0	0	13	6	12	14
real-96-90-22	0	0	0	0	0	0	0
real-96-94-22	0	0	0	0	0	0	0
real-110-104-22	0	0	4	4	22	9	21
real-126-121-23	38	12	0	1	5	7	16
real-132-129-24	50	0	1	2	2	5	8
avg.	5.87	0.80	1.07	1.33	3.13	3.13	7.73

Table 6. Comparison of four approaches.

	MIP		ML-MIP		TP-VRP		RCPSP-VRP	
instance	time	obj	time	obj	time	obj	time	obj
real-25-17-4	26.1	0	25.4	0	3.5	0	1.7	0
real-25-22-5	t.l.	33	t.l.	43	3.6	0	1.8	0
real-24-17-4	24.3	0	56.2	0	3.3	0	1.2	0
real-24-19-5	67.8	0	43.2	0	3.4	0	1.4	0
real-26-24-6	t.l.	42	t.l.	50	3.6	0	1.5	0
real-45-42-9	37.6	0	13.1	0	4.8	0	2.6	0
real-70-66-14	127.7	0	61.6	0	7.3	0	5.1	0
real-76-74-16	t.l.	—	t.l.	—	9.3	0	5.1	0
real-80-77-16	t.l.	—	t.l.	—	9.1	0	5.0	0
real-96-88-19	t.l.	—	t.l.	—	12.4	0	5.7	0
real-96-90-22	t.l.	—	t.l.	—	11.7	0	5.3	0
real-96-94-22	t.l.	—	t.l.	—	13.5	0	5.4	0
real-110-104-22	t.l.	—	t.l.	—	14.9	0	7.6	0
real-126-121-23	t.l.	—	t.l.	—	17.0	38	15.5	38
real-132-129-24	t.l.	—	t.l.	—	19.6	50	17.2	46

Table 7. Best objective values and method-wise absolute optimality gaps on real-world instances.

		absolute gap			
instance	best	MIP	ML-MIP	TP-VRP	RCPSP-VRP
real-25-17-4	0	0	0	0	0
real-25-22-5	0	33	43	0	0
real-24-17-4	0	0	0	0	0
real-24-19-5	0	0	0	0	0
real-26-24-6	0	42	50	0	0
real-45-42-9	0	0	0	0	0
real-70-66-14	0	0	0	0	0
real-76-74-16	0	—	—	0	0
real-80-77-16	0	—	—	0	0
real-96-88-19	0	—	—	0	0
real-96-90-22	0	—	—	0	0
real-96-94-22	0	—	—	0	0
real-110-104-22	0	—	—	0	0
real-126-121-23	38	—	—	0	0
real-132-129-24	46	—	—	4	0
avg.	5.60	10.71	13.29	0.27	0.00

By contrast, TP-VRP and RCPSP-VRP obtained matching or better feasible solutions in substantially shorter time for all instances with at most 70 patients. For the six instances with 76–110 patients, both TP-VRP and RCPSP-VRP obtained optimal solutions, whereas MIP and ML-MIP failed to find feasible solutions within the time limit. For the two largest instances, with more than 120 patients, RCPSP-VRP outperformed TP-VRP by achieving better objective values.

A concise summary of method-level performance on the 15 real-world instances is as follows: MIP and ML-MIP each found feasible solutions for 7 of the 15 instances (46.7%) within the 300-second time limit, whereas TP-VRP and RCPSP-VRP did so for all 15 instances (100%). The average absolute optimality gap from the best known objective value is 10.71 for MIP, 13.29 for ML-MIP, 0.27 for TP-VRP, and 0.00 for RCPSP-VRP. On the seven instances where MIP and RCPSP-VRP both found feasible solutions, the average runtime decreased from 126.2 to 2.19 seconds. These figures reinforce that the main benefit of decomposition is not only feasibility recovery but also a drastic reduction in search effort.

For 13 of the 15 instances, the best objective value was 0, and both TP-VRP and RCPSP-VRP reached this value quickly. Positive tardiness remains only in the two largest instances. Thus, although a 300-second time limit is sufficient for most tested cases, longer time limits may still be useful on the largest ones.

The prevalence of zero maximum tardiness deserves interpretation. It does not imply that the instances are trivial; rather, it indicates that many daily scheduled-transfer cases can be completed within all MRT limits once compatibility and routing are handled well. In such cases, the problem behaves like a highly constrained feasibility problem, while the largest instances preserve nonzero tardiness

and therefore discriminate more clearly between methods. A natural extension is to study secondary objectives—for example, total travel time or total tardiness—conditional on achieving $T_{\max} = 0$.

In addition to the real-world instances, random instances can be used as synthetic stress cases for heuristic validation. Although these instances are smaller, they can still be structurally tight; this is consistent with the observed MIP difficulty and the combinatorial coupling between compatibility and MRT constraints.

Because the current random instances were primarily generated for exact-label creation and controlled validation, they do not yet provide the kind of broad scalability study that a larger synthetic benchmark set would enable. We therefore view the real-world instances as the main evidence for large-scale performance and identify expanded synthetic testing as an important next step.

The difference in algorithm performance can be explained by analyzing the structure of the problem. The MIP and ML-MIP models must explore a large and tightly coupled decision space, where routing, timing, and feasibility constraints interact combinatorially. Even with a reduced number of binary decisions, the ML-MIP model struggles to ensure feasibility as the number of patients increases. Moreover, inaccuracies in the ML-driven patient selection may also negatively influence solver performance. In contrast, the TP-VRP and RCPSP-VRP approaches decompose the problem more effectively by constructing feasible routing and scheduling structures before optimization. This decomposition significantly reduces the search space and enables faster convergence. In the TP-VRP method, the algorithm parameter also balances the contributions of both the patient's MRT and the travel time in the patient-bed assignment, helping to avoid poor search directions. Additionally, the RCPSP-VRP better captures ambulance-resource contention in the assignment phase, which explains its superior performance for the largest two instances.

From an ablation perspective, the progression ML-MIP $\rightarrow$ TP-VRP $\rightarrow$ RCPSP-VRP is informative. Fixing patient selection alone offers limited benefit, so the assignment/routing decomposition is the main source of improvement. Incorporating ambulance-capacity information in the assignment phase yields a further gain on the hardest cases, explaining the advantage of RCPSP-VRP over TP-VRP. This interpretation clarifies which design choices are most consequential: decomposition matters more than learning-based selection, and resource-aware assignment matters most on large instances.

The methods also differ in their sensitivity to data perturbations. ML-MIP is the most brittle because an early misclassification can remove a clinically important patient from the reduced model. TP-VRP is easier to re-run quickly after a bed cancellation, but it may still assign too many concurrent transfers because ambulance contention is handled only later. RCPSP-VRP is more resilient to ambulance scarcity, although all three methods would benefit from a rolling-horizon contingency rule in practice: when a bed becomes unavailable or an ambulance breaks down, already started transfers are fixed, and the remaining unserved tasks are re-optimized from the current time.

Overall scalability is governed primarily by the number of patients, but the interaction with bed capacity and ambulance count is also important. More patients increase the assignment and routing search spaces, tighter bed capacity intensifies competition among compatible destinations, and fewer ambulances increase route coupling over time. This is precisely why RCPSP-VRP becomes relatively stronger on the largest instances: it internalizes the scarce-ambulance effect before route construction.

A practical limitation of the present study is that we do not compare our approach against a wider family of external metaheuristics. Our goal here is to compare three decomposition strategies that are naturally aligned with the proposed MIP formulation. Expanding the benchmark set to include general-

purpose metaheuristics or uncertainty-aware methods is an important direction for future empirical validation.

7. Conclusions

In this study, we investigated an inter-hospital ambulance routing problem (IH-ARP) motivated by challenges in the Sri Lankan health service system. To address it, we formulated a mixed integer programming (MIP) model that minimizes the maximum tardiness. However, the exact model struggled to produce feasible solutions within a realistic time limit, which motivated the development of three heuristic approaches.

The first approach, ML-MIP, uses binary classification to identify a subset of patients to be transferred and then solves a simplified mathematical model. The second approach, TP-VRP, decomposes the problem into two phases: patient-to-bed assignment via a transportation problem and route construction via a vehicle routing problem with time windows (VRP-TW) solved through binary search on tardiness. The third approach, RCPSP-VRP, replaces the transportation problem with a resource-constrained project scheduling problem (RCPSP) to account for ambulance availability and then improves the solution within an iterative framework.

Our computational experiments on real-world data highlight the strengths and limitations of these approaches. MIP and ML-MIP perform similarly across instances but fail to solve the large-scale cases within the time limit. By contrast, TP-VRP and RCPSP-VRP match the optimal solutions found by MIP on the instances solved to optimality, while requiring much less time. Among them, RCPSP-VRP performs best on the largest instances, especially when the number of patients exceeds 100.

The main novelty of the work lies in the integrated IH-ARP formulation and in the problem-specific decomposition of assignment, scheduling, and routing under an MRT-based worst-case objective. We further clarified the deterministic assumptions, the distinction between ST and IIT operations, the role of the parameter α, and the practical robustness implications of the proposed heuristics.

Several directions for future research remain. First, the proposed methods could be improved further, and one promising direction is the development of an automated tuning procedure for α in TP-VRP to improve adaptability and performance across diverse instances. Second, it would be valuable to design algorithms that produce lower bounds for IH-ARP, especially for large-scale instances where the current MIP lower bounds are consistently 0. Such bounds would provide stronger benchmarks for evaluating heuristic solution quality. Finally, although this study focuses on scheduled inter-hospital transfers, developing algorithms tailored to immediate emergency transfers is another important direction.

Acknowledgments

We sincerely thank Dr. (Ms) Thilina Wanigasekera, Medical Superintendent of Base Hospital Puttalam, Sri Lanka, and Dr. Sampath Samaraweera, Medical Officer (Organization Development), Ministry of Health, Sri Lanka, for their support and for sharing information on the background of EMTSs in Sri Lanka.

This work was supported by JSPS KAKENHI [Grant No. 25K00180].

Data availability

The data that support the findings of this study are not publicly available due to confidentiality and institutional restrictions. Data may be made available from the corresponding author upon reasonable request and with permission from the relevant authorities.

Use of Generative-AI tools declaration

The authors declare that this paper reports the original research work by the research group and is not generated by Generative-Al tools.

Author contributions

Sudheeraka Wickramarachchi: Formal analysis, writing-original draft, methodology, data curation, and validation; **Kazuki Hasegawa**: Conceptualization and writing-review; **Wei Wu**: Supervision, methodology, software, and funding acquisition.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. S. Lee, The role of centrality in ambulance dispatching, *Decis. Support Syst.*, **54** (2012), 282–291. https://doi.org/10.1016/j.dss.2012.05.036
2. S. N. Shetab-Boushehri, P. Rajabi, R. Mahmoudi, Modeling location–allocation of emergency medical service stations and ambulance routing problems considering the variability of events and recurrent traffic congestion: A real case study, *Healthc. Anal.*, **2** (2022), 100048. https://doi.org/10.1016/j.health.2022.100048
3. D. Neira-Rodado, J. Wilmer Escobar-Velasquez, S. McClean, Ambulances deployment problems: Categorization, evolution and dynamic problems review, *ISPRS Int. J. Geo-Inf.*, **11** (2022), 109. https://doi.org/10.3390/ijgi11020109
4. L. Talarico, F. Meisel, K. Sörensen, Ambulance routing for disaster response with patient groups, *Comput. Oper. Res.*, **56** (2015), 120–133. https://doi.org/10.1016/j.cor.2014.11.006
5. T. Tlili, M. Harzi, S. Krichen, Swarm-based approach for solving the ambulance routing problem, *Procedia Comput. Sci.*, **112** (2017), 350–357. https://doi.org/10.1016/j.procs.2017.08.012
6. F. Ziya-Gorabi, A. Ghodratnama, R. Tavakkoli-Moghaddam, M. S. Asadi-Lari, A new fuzzy tri-objective model for a home health care problem with green ambulance routing and congestion under uncertainty, *Expert Syst. Appl.*, **201** (2022), 117093. https://doi.org/10.1016/j.eswa.2022.117093
7. I. Zidi, A novel approach for dynamic ambulance routing: Integrating k-means++ clustering with time-variant multi-objective SPEA2, *J. Adm. Econ. Sci.*, **18** (2025), 619–642. https://doi.org/10.25259/JAES_18_2_619

8. D. Bredström, M. Rönnqvist, Combined vehicle routing and scheduling with temporal precedence and synchronization constraints, *Eur. J. Oper. Res.*, **191** (2008), 19–31. https://doi.org/10.1016/j.ejor.2007.07.033
9. N. A. Wouda, L. Lan, W. Kool, Pyvrp: A high-performance vrp solver package, *Informs J. Comput.*, **36** (2024), 943–955. https://doi.org/10.1287/ijoc.2023.0055
10. A. Bozorgi-Amiri, S. Tavakoli, H. Mirzaeipour, M. Rabbani, Integrated locating of helicopter stations and helipads for wounded transfer under demand location uncertainty, *Am. J. Emerg. Med.*, **35** (2017), 410–417. https://doi.org/10.1016/j.ajem.2016.11.024
11. T. Carnes, S. G. Henderson, D. B. Shmoys, M. Ahghari, R. D. MacDonald, Mathematical programming guides air-ambulance routing at ornge, *Interfaces*, **43** (2013), 232–239. https://doi.org/10.1287/inte.2013.0683
12. S. Buya, P. Tongkumchum, B. E. Owusu, Modelling of land-use change in thailand using binary logistic regression and multinomial logistic regression, *Arab. J. Geosci.*, **13** (2020), 1–12. https://doi.org/10.1007/s12517-020-05451-2
13. J. H. Friedman, Stochastic gradient boosting, *Comput. Stat. Data Anal.*, **38** (2002), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2
14. J. L. Speiser, M. E. Miller, J. Tooze, E. Ip, A comparison of random forest variable selection methods for classification prediction modeling, *Expert Syst. Appl.*, **134** (2019), 93–101. https://doi.org/10.1016/j.eswa.2019.05.028
15. Z. Wang, X. Xue, *Multi-class support vector machine*, 23–48, Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-02300-7_2
16. P. C. Pendharkar, A comparison of gradient ascent, gradient descent and genetic-algorithm-based artificial neural networks for the binary classification problem, *Expert Syst.*, **24** (2007), 65–86. https://doi.org/10.1111/j.1468-0394.2007.00421.x
17. G. N. Frederickson, A note on the complexity of a simple transportation problem, *SIAM J. Comput.*, **22** (1993), 57–61. https://doi.org/10.1137/0222005
18. P. Lacomme, A. Moukrim, A. Quilliot, M. Vinot, Integration of routing into a resource-constrained project scheduling problem, *EURO J. Comput. Optim.*, **7** (2019), 421–464. https://doi.org/10.1007/s13675-018-0104-z

AIMS Press