Minimizing delays of patient transports with incomplete information: A modeling approach based on the Vehicle Routing Problem

Dennis Adelhütte[1], Kristin Braun[*,1,2], Frauke Liers[1], and Sebastian Tschuppik[1]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Department of Data Science, Cauerstraße 11, 91058 Erlangen, Germany

[2]Fraunhofer Institute for Integrated Circuits IIS, Nordostpark 84, 90411 Nuremberg, Germany

8th May, 2024

Abstract

We investigate a challenging task in ambulatory care, the minimizing of delays of patient transports. In practice, a limited number of vehicles is available for non-rescue transports. Furthermore, the dispatcher rarely has access to complete information when establishing a transport plan for dispatching the vehicles. If additional transport is requested on demand then schedules need to be updated, which can lead to long delays. We model the scheduling of patient transports as a vehicle routing problem with general time windows and solve it as a mixed-integer linear problem that is modified whenever additional transport information becomes available. We propose a modeling approach that is designed to determine fair and stable plans. Furthermore, we show that the model can easily be modified when transports need to satisfy additional requirements, e.g., during pandemics, exemplarily the Covid-19 pandemic. To show the applicability and efficiency of our modeling approach, we conduct a numerical study using historical data from the region of Middle Franconia. The results reveal and show that, by applying mathematical optimization - or, to be more precise by solving mixed-integer linear problem formulations - one can significantly decrease delays and have considerable potential for optimized patient transports.

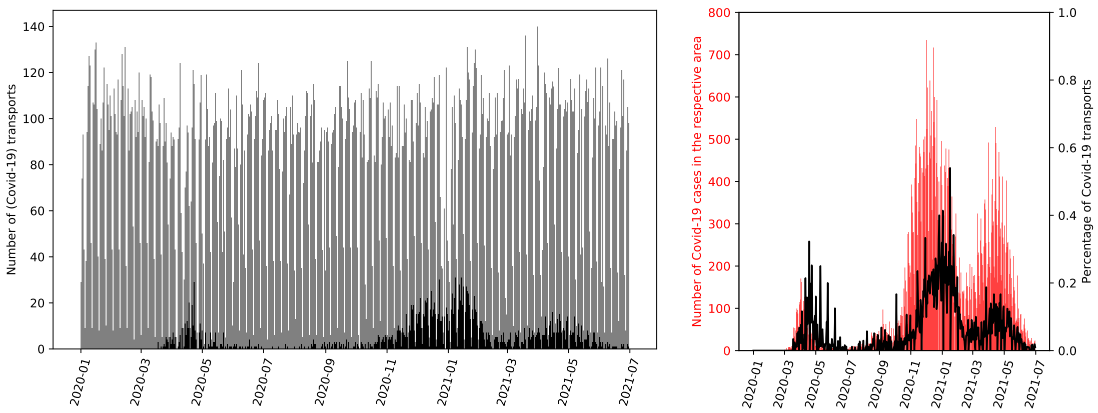
Keywords: OR in Health Services, Vehicle Routing, Heuristics, Mixed-Integer Linear Optimization, Patient Transport

*Corresponding author, kristin.braun@fau.de

1 Introduction

In a healthcare system, patient transport needs to be well planned to ensure a functioning system. Because such a system consists of many components, e.g., health promotion, primary care, specialized services, and hospitals, transporting patients without delays is by no means trivial, especially when the respective transports are not emergencies and can be postponed. This could be the case when a patient needs to be transported home from a hospital or when patients receive certain treatments at a specific destination. While in principle those transports could be carried out by the vehicles that carry out rescue transports, a different fleet is typically used for other patient transports in Germany. In contrast to rescue transports, patient transports allow delays, though they are not desirable. Scheduling them is a challenge due to different reasons. While the number of vehicles in the transport fleet is limited, the drivers' work shifts need to be respected. Moreover, a vehicle can transport only one patient at once. Finally, not all transport requests are known at the time when the plans are established: Many transports are requested during the day when a transport schedule is already in operation. The last point in particular can lead to long delays for patients waiting for their transport, as it is often impossible to handle all transports at once.

A natural scheduling approach is a greedy approach: whenever a transport is requested, the vehicle that can reach the patient most quickly is dispatched. This approach is performed in Middle Franconia, according to the local dispatcher, the Integrierte Leitstelle Nürnberg (ILS). For the resulting transport plans, optimization potential is usually disregarded and sometimes transports even have to be rescheduled to the following day. As a motivation, we provide some statistics about the transport data in the following.



(a) Absolute number of transports, distinguished on the basis whether the patient is known to be infected (black bars) or not (gray bars).

(b) Percentage of known Covid-19 transports and total number of Covid-19 cases in the catchment area of the ILS.

Figure 1: Statistics on the numbers of transports and the percentage of known Covid-19 transports for the available data.

Statistics about the patient transports. We consider the number of transports from January 2020 to July 2021. In Figure 1a, we plot the number of patient transports. The number of weekly transports on weekends, especially on Sundays, is decreasing, while the total number of transports shows a similar order of magnitude. Over the turn of the year, due to Christmas and other holidays, fewer transports are requested. In 2020, there is another small decrease starting in April. There is no such decline before 2020, so this is likely due to the onset of the Covid-19 pandemic. Further, we plot the percentage of transports that involved an patient either infected, or, at least, suspected to be infected, with Covid-19. These are represented by the black portion of the bars in Figure 1a. For better visibility, the percentage of Covid-19 transports is also given in Figure 1b. In the peak phase, up to 60% of all transports have been classified as infected cases. This is – as there are a lot of suspected cases – not the actual number of infections. However, the trend given in Figure 1b

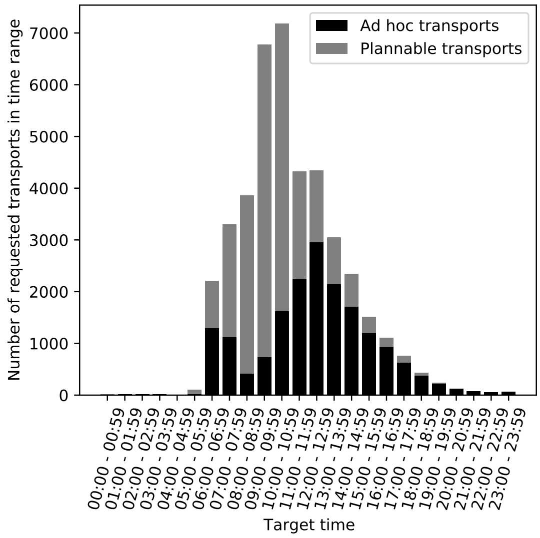


Figure 2: Number of plannable and ad hoc transports, differentiated by the target time.

is quite similar to the number of actual Covid-19 cases in Middle Franconia, cf. [46]. This trend is shown in the red part of the plot.

Furthermore, we consider the number of transports during the course of a day in Figure 2: Transports are distinguished depending on the time when they are requested and the number of requested transports per hour is plotted for the same time window as before. The majority of the transports are requested at 6 am or later. The peak is in the late morning, thereafter the number slowly decreases. After about 7 pm, the number of transports is again relatively low compared to the previous hours.

The number of plannable transports[1] is shown in the gray bars, while the black bars represent the number of ad hoc transports.[2] In the early morning hours, i.e., between 6 and 8 am a lot of transports are ad hoc ones. After that, until noon, most transports are plannable. Then, the percentage of ad hoc transports increases over the course of the day. After 3 pm, more than 80% of all transports are ad hoc transports.

Our contribution. In this paper, we model the problem of finding fair schedules for patient transports. In our case, 'fair' means that the maximum delay over all patient transports is minimized as the first priority, before, secondly, the total delay over all transports is minimized. Besides the aim of distributing the minimal delay among the patients roughly uniformly we need to respect the drivers' shifts whenever possible.

We present two approaches to handle transports that are requested during the course of the day: On the one hand, if the dispatcher has knowledge of the requested transport in advance, but does not yet know the time it is supposed to happen (for example a patient needs to be taken home after a treatment), so-called dummy transports are introduced to block vehicle capacities. These transports can be expected due to, for example, requests that typically come in at specific times of the day or as follow-up transports. On the other hand, whenever an ad hoc transport becomes known to the dispatcher, we reoptimize the part of the schedule that has not yet begun and establish a new schedule that incorporates the ad hoc transport. The model is based on the vehicle routing problem with general time windows (VRPGTW) that was introduced in [28]. We solve the optimization problems using state-of-the-art solvers for mixed-integer linear problems (MIPs) that we enhance with heuristics in order to improve their running time. We demonstrate that one can modify our proposed model whenever necessary by introducing a general and adaptable way to add new inequalities, equations and penalty variables. As an application, we discuss the

[1] Transports that are known when establishing a schedule.
[2] Transports that become known during the day.

issues that need to be taken into account during the Covid-19 pandemic to minimize the risk of infection and how to model them. In our numerical study, we show that our optimized procedure for scheduling transports is effective and efficient in practice. Regarding infectious illnesses, we show that transporting patients that are (at least suspected to be) infected using separate vehicles is desirable.

In several key points, our approach differs from previous work. Firstly, rather than treating fairness as a separate model parameter, we incorporate it directly into our objective function. Secondly, our methodology ensures robustness without the use of random variables. Thirdly, rather than focusing on optimizing required time or returned distance, we prioritize minimizing delays for patients even if this causes detours of vehicles. Our research is based on a dynamic and deterministic vehicle routing problem (VRP) and employs a wait-first strategy. Finally, because rescue and non-rescue transports in Germany use separate vehicle fleets, we also use a separate fleet dedicated solely to patient transports, distinct from the fleet designated for rescue vehicles. We discuss these points in more detail in our literature review in Section 2.

Structure of the paper. In Section 2, we present related work concerning vehicle routing problems and other problems arising in the healthcare sector. Furthermore, we show our chosen way of modeling the VRPGTW. Thereupon, in Section 3, the patient transport problem is described on an abstract level. We first define different transports and discuss the amount of information available at the time of planning. We also introduce techniques for modifying models on an abstract level. In Section 4, we discuss how the patient transport problem can be modeled as a MIP. Subsequently, the model is modified to incorporate the required updates when a previously unknown transport is requested. We then describe how to incorporate Covid-19-related requirements, including disinfection time and the goal of separating known Covid-19 transports from other requests as far as possible.

In Section 5, we elaborate on algorithmic methods to improve the running time for our models' solving process and evaluate the methods by using the available historical data. In particular, we compare our methods to an implementation simulating the current scheduling practice at the ILS. Finally, in Section 6, we discuss our results and provide some ideas for future research.

2 Literature review

Since the VRP generalizes the traveling salesperson problem (TSP), it is naturally NP-hard. Before we present our chosen approach to model the VRP in Subsection 2.1, we present an overview of research about VRPs in the literature. Further, we present and discuss some literature concerning transport problems in healthcare that have been discussed in the context of mathematical optimization. Finally, we classify our problem depending on these findings.

Vehicle routing problems. The VRP presented in [15] is a generalization of the well-known TSP: Given an (un-)directed graph, and given a fixed nonnegative integer m and a fixed node, the question is whether the given graph can be partitioned into at most m Hamiltonian cycles that share only the given fixed node. The interpretation given in [15] is that the given number m is the number of available vehicles to carry out (petrol) deliveries such that each customer is served exactly once by exactly one vehicle and that each vehicle starts and ends its tour at a depot (which corresponds to the given fixed node). For a general overview of vehicle routing problems, we refer to [56]. In this work, an overview of different methods for modeling and solving variations of vehicle routing problems and its modifications are presented and evaluated, including branch-and-bound-algorithms, branch-and-cut-algorithms, set-covering-based algorithms and heuristic methods, among others. It is also possible to model VRPs as MIPs. This is also our solution approach since we use state-of-the-art software in our numerical study. In Subsection 2.1 we justify this choice in more detail.

The works of [13, 14] present an overview of different problem classes of vehicle routing problems. Two generalizations are the vehicle routing problems with time windows and the vehicle routing problem with pick-up and delivery. The latter one usually contains time windows, so one mostly omits the pick-up and delivery part. In our work, we use the VRPGTW which was introduced in

[28]. Contrary to the vehicle routing problem with time windows, the bounds on target times can be soft, i.e., their violation is penalized, or hard, i.e., the corresponding constraints need to be satisfied. [55] discusses many current VRP extensions, such as service quality, equity, and working hours, among others. Another generalization discussed in [28] is the dial-a-ride problem. This problem aims to model transports of passengers. Thus, human factors like their satisfaction must also be included. For an overview we simply refer to [26].
VRPs are either static or dynamic and either deterministic or stochastic. In a static VRP, all transports are known beforehand, while in a dynamic one they can change over time. Further, new transports can be requested over time. The distinction between deterministic and stochastic VRP depends on whether some data is uncertain.

Dynamic VRPs. Dynamic VRPs require more sophisticated solution techniques, while static VRPs are, in comparison, usually easy to solve by applying state-of-the art MIP approaches. [5, 43, 4] present overviews of dynamic VRPs. The first consists of a collection of problems while the latter focuses on the distinction between periodic and continuous solving methods. A periodic solving method re-optimizes the problem after a certain (fixed) time period or as soon as new data is available while a continuous solving method is performed throughout the whole day. Another overview from [44] further introduces a new classification scheme for dynamic VRPs that consists of eleven criteria, for example the type of the objective function (i.e., whether one minimizes costs, distances, travel times, etc.), the fleet size, the type of time constraints and the solution method. Another solution method for dynamic VRPs is given in [18]. There, the authors introduce dummies to precautionarily schedule ad hoc transports to areas where new requests are likely to occur. Further, there are different possibilities, how currently waiting vehicles can behave to handle incoming transport requests better. The work of [38] describes different waiting strategies, namely the drive-first strategy where a vehicle leaves its current location as soon as possible. In contrast to the drive-first strategy, for the wait-first strategy, the vehicle waits as long as possible. Further, two mixtures of both are introduced. Due to the stochastic probability distribution for new customers' location, heuristics focusing on the waiting location are considered.

Stochastic VRPs. In addition to dynamic VRPs, uncertainties add another layer of complexity. Stochastic VRPs can consist of different types of uncertainties. The work of [50] summarizes where those may occur: The main source of uncertainty is demand, which refers to the number of requests, as well as where, if, and when they will occur. Furthermore, the environment may be uncertain, such as fluctuating travel times due to traffic. Finally, one may have uncertain resources, which means that the availability of, for example, vehicles and drivers is not guaranteed. Further overviews over stochastic VRPs can, for example, be found in [6, 40]. The work of [39] focuses on solution methods for these problems. [8] is one of the first works discussing dynamic and stochastic VRPs. There, the authors propose heuristics to solve these more difficult problems. Also, [20] distinguishes stochastic, dynamic and dynamic/stochastic VRPs. They further mention that it is helpful to gather patient data to determine a probability distribution. Similarly, the more current survey [45] investigate these three types of VRPs.
Stochastic VRPs are frequently solved with Markov Decision Processes (MDPs). A general MDP has four components: states, actions, transitions, and rewards. For VRPs, each state contains the vehicle(s)' current location, arrival time at the current node, and the status of all customers. An action assigns times to customers, and transitions are used for changing from one state to the next after an action is chosen. The reward is determined by how the problem is defined and what specific goals are set, e.g. minimizing travel times, maximizing the number of visited customers or balancing the workload between drivers.
Unlike general MDPs, route-based MDPs explicitly model route plans as sequences of possible future actions, thereby redefining the feasible action space. This enables real-time decisions that directly adjust and optimize future routes. This also implies that states in route-based MDPs now include routes. [52] shows that MDP and route-based MDP formulations are equivalent when using a slightly restricted formulation of the Bellman equation.
[31] and [30] apply MDPs to solve stochastic VRPs with uncertain travel times. The first uses heuristics and real-time data, while the latter deals with nonstationary stochastic travel times and attempts to determine a probability distribution. Furthermore, [31] optimizes driver attendance

times and vehicle coverage, which are fixed in our case. In both cases, the customers and their demands are known in advance. Further, [64, 57] and [48] solve stochastic versions of a dial-a-ride problem with uncertain travel times.
In [49], the authors assume that they have complete information about all customers, with the exception of stochastic demands. As a result, due to limited capacity, vehicles may be required to return to the depot during replanning. MDPs and heuristics are used to solve the problem for a single vehicle with known probability distributions.
There is also some work to be done to address the uncertainty concerning whether multiple requests will occur at all. In [53], possible customer locations are known beforehand. Additionally, the number of requests is available. [60] discusses vacant taxi routing, which involves deciding where taxis that are not currently transporting passengers should drive or wait. In this case, customer requests may appear from any position. This is handled by clustering positions into zones, resulting in a discrete action space. Both works use MDPs to solve stochastic VRPs. [51] also solves a dynamic and stochastic VRP with unknown requests. The author pays special attention on the question where vehicles should wait.
Other solution methods feature combinations of MDPs with rolling horizon approaches [36] and genetic ant colony algorithms [25].
Many applications can be modeled as order picking and delivery problems. In [63, 62], the authors consider delivery for business-to-costumer and online-to-offline supermarkets, respectively. A more general work can be found in [11] where a VRP for minimizing the driven road distance is modeled and evaluated on benchmark instances. In contrast to their approach, we minimize delays and not a road distance. In [22], the main question is to decide whether ad hoc transports should be accepted or rejected what is not possible in healthcare. During the whole time period, the problem is optimized continuously. [32] introduce a new model, namely a VRP with time-dependent travel times. Varying travel times, such as those caused by traffic jams, are taken into account in this specific model. The travel time is determined by the origin and destination, as well as the requested departure time. By incorporating this information, the authors intend to enhance the reliability of planned routes. In [7], the authors apply the solution of snapshots to handle another dynamic VRP, namely scheduling ad hoc transports for online taxi routing. In our work, we consider a dynamic and deterministic VRP.

Transport problems in healthcare. There is also a lot of work that model passenger transporting problems that occur in healthcare as vehicle routing problems. The work of [27] is a taxonomy of healthcare decisions and thus, a good overview. In [3] and [29], the patient transport problem within a hospital is considered. In the former one, they assume that the hospital is one building, while in the latter one, the buildings of the hospital are spread over the whole city — in this specific example, in Tours (France). A static version of the VRP for the patient transport problem is discussed in [37], an example for a work about rescue transports is [23]. Important work in the field of patient transport in the European area is described in [54], where the focus is on the situation in the Netherlands. There, rescue vehicles can also be used for patient transports. This, however, reduces the number of vehicles available for rescue missions. Although in principle this is possible in Germany as well, this option is usually avoided by the dispatcher due to the separated fleets, and we do not consider this option in our work. In [16], the authors consider the situation in Austria where they describe the stationing of rescue vehicles or the periodic delivery of blood reserves. Another work that uses data from the US is [61]. They consider the patient transport problem, but model it stochastically. In their work, service duration and travel time can be stochastic. They also model random cancellations and apply a k-means clustering-based algorithm to solve the dynamic part of the problem. In [35], the authors consider a real-world example from Copenhagen, Denmark. There, elderly and disabled people need to be transported, where each transport has either a pick-up or a delivery time window. Further, the dispatcher needs to be made aware of different equipment that might be needed, for example when a patient requires a wheelchair. The objective of the problem represents waiting times, driving times, the number of vehicles or a combination of those. Another real-world example is examined in [42]. In addition to the necessary equipment for wheelchair transports, they consider the possibility of accompanying persons wishing to be transported, or the need for a second staff member. Furthermore, time windows for the staff are considered for the start and end of the shift, as well as for a break during

the shift. A variable neighborhood search and column generation is used to solve the problem. [19] develop a scheduling algorithm for routing tasks in Austrian hospitals, mostly for the delivery of goods. A different application in healthcare is presented in [47]. They consider the issue of (mostly rural) areas lacking hospitals. To keep the travel distance of residents as short as possible, doctors visit these areas on a regular basis, depending on the area's size. Recently, the authors have considered the need for Covid-19 testing centers there. Their data is from Turkey, and they model the problem as a periodic location routing problem.
Concerning additional requirements due to the Covid-19 pandemic, [41] describe an application of the vehicle routing problem in practice, which is solved heuristically. The authors used tabu search to heuristically plan the distribution of face masks in Spain. [2] tested the feasibility of schedules for transporting dialysis patients under worst case assumptions for the spreading of the virus using Monte Carlo simulations. Contactless delivery of food to settlements during the pandemic was considered in [10] and solved by applying a genetic algorithm. [21] consider vaccine distribution across different regions. They pose the question when vaccines should be available and which amount is required. The problem has multiple objectives and is solved using an optimal control approach combined with dynamic programming.

2.1 The vehicle routing problem with general time windows

To model our transport problem as a MIP, we choose a VRPGTW model. This formulation is defined on a directed graph $G := (V, A)$ with vertex set $V := \{0, \dots, n\}$, $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, arc set $A \subseteq V \times V$ and a set of homogeneous vehicles $K := \{1, \dots, m\}$, $m \in \mathbb{N}$. For all $i \in V$, the set of outgoing arcs is denoted by $\delta^+(i)$ and the set of incoming arcs is denoted by $\delta^-(i)$. Each node $i \in \{1, \dots, n\} =: N$ corresponds to one customer, while node 0 is the depot where all vehicles start and end their trip. The duration $\tau_{i,j} \geqslant 0$ denotes the time that is required to serve node i and to travel to node j afterwards.
For all $(i, j) \in A$ and $k \in K$, the binary variable $x_{i,j,k} \in \{0, 1\}$ indicates whether vehicle k travels on the arc (i, j). For all $i \in V$, $y_i \in \mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geqslant 0\}$ denotes the time when the transport for customer i starts at its origin. Finally, for each $i \in V$, $p_i \colon \mathbb{R}_+ \to \mathbb{R}_+$ is a piecewise linear penalty function which yields a penalty dependent on the delay necessary to serve customer i. The goal is serving each customer exactly once, with as little penalty – w.r.t. a given goal – as possible. This can be modeled as a mixed-integer nonlinear optimization problem (MINLP) that is easy to linearize, see Model (1). Our objective is to minimize the penalty terms. Constraints (1b) – (1e) ensure that there is a tour where each vehicle starts and ends at the depot and each customer is served exactly once, see [56]. Together with Constraint (1f), valid tours are defined. It ensures that a vehicle which serves customer j directly after customer i finishes the service of customer i and travels from customer i to j before the vehicle starts serving customer j:

$$\min \quad \sum_{i \in N} p_i(y_i) \tag{1a}$$

$$\text{s.t.} \quad \sum_{k \in K} \sum_{j \in \delta^+(i)} x_{i,j,k} = 1 \qquad \forall i \in N, \tag{1b}$$

$$\sum_{j \in \delta^+(0)} x_{0,j,k} = 1 \qquad \forall k \in K, \tag{1c}$$

$$\sum_{j \in \delta^+(i)} x_{i,j,k} - \sum_{j \in \delta^-(i)} x_{i,j,k} = 0 \qquad \forall k \in K, i \in N, \tag{1d}$$

$$\sum_{i \in \delta^-(0)} x_{i,0,k} = 1 \qquad \forall k \in K, \tag{1e}$$

$$x_{i,j,k}(y_i + \tau_{i,j} - y_j) \leqslant 0 \qquad \forall k \in K, (i, j) \in A, \tag{1f}$$

$$x_{i,j,k} \in \{0, 1\} \qquad \forall k \in K, (i, j) \in A, \tag{1g}$$

$$y_i \geqslant 0 \qquad \forall i \in N. \tag{1h}$$

Linearizing piecewise linear functions and Constraint (1f) are standard techniques. In particular, Constraint (1f) is equivalent to

$$y_i + \tau_{i,j} - y_j \leqslant M_{i,j}(1 - x_{i,j,k})$$

for a sufficiently large real number $M_{i,j} > 0$. Thus, the resulting model is a MIP that is NP-hard in theory.

Remark 2.1. *In [28], the vehicles also have a capacity for transporting goods, and other costs are also incurred. However, for our purposes, the given parameters are sufficient since our vehicles can only transport exactly one patient a time and we are only interested in minimizing delays of transports.*

Classification of our problem and solution approach. Our model faces uncertainty about incoming requests, i.e., we are unaware of when, or how many requests will come in and where their pick-up and drop-off locations are. Further, we assume that the environment and resources are known. To handle this uncertainty, we avoided stochastic methods because of a lack of reliable distribution information, with the exception of backward transports. Still, in this case, we use an approach with no random variables but incorporate dummy transports into our model. While traditional Markov Decision Processes as explained previously require knowledge of potential customer locations, in our scenario, customers can be located anywhere. Although clustering methods like those in [60] are conceivable, the issue of unknown distribution persists, complicating the adoption of standard probabilistic models. Interestingly, experiments where we assumed to have a full knowledge about all transports failed to yield significant improvements in solution quality. Thus, we use a periodic dynamic VRP formulation that re-optimizes our schedule as soon as a new request becomes known. Further, we use the wait-first strategy in our model.
Rather than modeling human satisfaction directly like in dial-a-ride problems, we built our objective function to ensure a fair distribution of delay: We want to achieve fair schedules by minimizing the maximum delay across all transportation modes per day and respect drivers' shift times due to employment law reasons.
Modeling our dynamic VRP as a MIP provides two significant benefits that align well with the complexities and uncertainties of our operational setup. First, we can use off-the-shelf optimization software to find global optimal solutions. Second, using a MIP formulation allows us to swiftly modify our model as our operations change, without having to reconstruct our software or algorithms. This allows us to easily adapt to new challenges and try out new ideas, as demonstrated in Subsection 4.1. Furthermore, the MIP formulation allows us to easily include various requirements and goals, such as ensuring that our schedules are fair and compliant with rules and regulations. So, by using this formulation, we not only improve our scheduling, but we also ensure that we can handle unexpected changes while still providing good transportation services in a fair and legal manner.

3 Mathematical description of the problem

In this chapter, we present the concepts and definitions required to formalize the patient transportation problem from Subsection 2.1. We begin with definitions of transports and schedules and then extend them by incorporating incomplete information.
The following elements are required to define a patient transport: A person has to be transported from one place to another. Each patient transport has an origin and a destination, and a vehicle may only transport one patient at a time. It also has a target time and a specified duration (the time required for travel and other tasks). In practice, some of this information can be incomplete.

Complete information. First, we formalize the definition of a transport where the information is complete.

Definition 3.1 (Plannable transport)**.** *A* plannable transport T *is a tuple* (O_T, D_T, d_T, t_T) *that consists of*

i) its origin O_T*, i.e., the place where a patient is picked up,* *ii)* *its* destination D_T*, i.e., the place where a patient is dropped off,*

iii) *its* duration d_T*, i.e., the total duration of the transport from reaching* O_T *until leaving* D_T*, and,*

iv) *its* target time t_T*, i.e., the time when* T *is supposed to be started at* O_T*.*

The set of plannable transports is denoted as $\mathcal{T}_{\text{plan}}$*. The dispatcher has complete information about the set* $\mathcal{T}_{\text{plan}}$ *when scheduling the vehicles.*

A daily schedule is a collection of all vehicle tours throughout the day. Each vehicle begins its daily tour at its depot and returns there once all assigned transports have been completed. This behavior is independent of the current schedule. A vehicle waits at D_T after finishing its transport. To formalize whether or not a schedule can be carried out, we define the following:

Definition 3.2 (Feasible schedule)**.** *Let* $\mathcal{T}_{\text{plan}}$ *be a set of plannable patient transports,* K *the set of available vehicles, and let, for all* $T \in \mathcal{T}_{\text{plan}}$*,* $y_T \in \mathbb{R}_+$*.*[3] *The set* $\mathcal{P}$ *with* $(T, k, y_T) \in \mathcal{P}$ *if and only if* $k \in K$ *is scheduled to carry out* $T \in \mathcal{T}_{\text{plan}}$ *at* y_T *is called the* schedule*. The schedule is called* feasible *if*

i) $y_T \geqslant t_T$*, i.e., a transport is supposed to be carried out earliest at its specified target time,*

ii) *for two subsequent transports* $T_i, T_j \in \mathcal{T}_{\text{plan}}$ *carried out by the same vehicle* $k \in K$*, inequality* $y_{T_j} \geqslant y_{T_i} + d_{T_i}$ *holds, and,*

iii) *each vehicle* $k \in K$ *starts and ends its trip within predefined shift times.*

The set of feasible schedules is henceforth denoted as $\mathcal{X}$*.*

Property (*iii*) specifies shift times that need to be followed. These shift times are taken from reality, as vehicle drivers work in shifts. As a result, a vehicle's tour should begin and end during its shift. This typically turns out to be a bottleneck as we will see in our numerical study.
To be able to distinguish between schedules that respect and do not respect the shift times, we introduce weakly feasible schedules:

Definition 3.3 (Weakly feasible schedule)**.** *If a schedule* $\mathcal{P}$ *satisfies Properties (i) and (ii) of Definition 3.2 then* $\mathcal{P}$ *is called* weakly feasible*. The set of weakly feasible schedules is denoted by* $\mathcal{X}_{\text{weak}}$*.*

The delay of vehicles is not taken into account in these definitions of schedules, as the target time of a transport is not considered for feasibility. However, our goal is establishing schedules that minimizes patients' delays while simultaneously respecting the shift times as much as possible. Thus, having the two competing goals of maintaining shift times while minimizing delays in a fair manner raises the question of what makes up a 'good' feasible schedule. To address multiple objectives, we propose the following:

Definition 3.4 (Measure of quality)**.** *Let* $\mathcal{X}_{\text{weak}}$ *be a set of weakly feasible schedules and assume that we are given* $l \in \mathbb{N}$ *real-valued functions* $g_1(x), \ldots, g_l(x)$*, defined on* $\mathcal{X}_{\text{weak}}$*. Let* $\gamma_1, \ldots, \gamma_l \geqslant 0$ *be a set of weights. We call the function* $f \colon \mathcal{X}_{\text{weak}} \to \mathbb{R}$ *with* $f(x) := \sum_{i=1}^{l} \gamma_i g_i(x)$ *the measure of quality with respect to the weights* $\gamma_1, \ldots, \gamma_l \geqslant 0$.

Naturally, this definition also works for feasible schedules $\mathcal{X}$ instead of weakly feasible schedules $\mathcal{X}_{\text{weak}}$. Possible criteria include the maximum delay of all transports and the sum of the delays of all transports or the violation of shifts.

Remark 3.5. *The method of weighting objectives to obtain exactly one objective value is referred to as scalarization and is applied in multi-objective optimization. Since we use single-objective optimization methods, we simply refer to [17].*

[3] Naturally, here, and in the following, all defined times are in $\mathbb{R}_+$.

We now formulate the optimization problem on an abstract level: For a given set of transports and given quality criteria with weights, the task is to solve

$$\min_{x \in \mathcal{X}} f(x) \quad \text{or, if necessary,} \quad \min_{x \in \mathcal{X}_{\text{weak}}} h(x). \tag{2}$$

If we optimize over $\mathcal{X}_{\text{weak}}$ then we assume that the violation of vehicles' shift times is one of the measures of quality of the schedule in the sense of Definition 3.4, i.e., $h(x) := f(x) + \sum_{k \in K} \gamma_k g_k(x)$ for some $\gamma_k \geqslant 0$ and real-valued function g_k defined on $\mathcal{X}_{\text{weak}}$, for all $k \in K$.

Incomplete information. Incomplete information can manifest in two ways: Firstly, a transport has been requested but its target time is not known. This can occur, for example, if a patient needs to be returned home following a treatment but the dispatcher is unsure when it will be completed. Secondly, a transport is requested that was not known when the schedule for $\mathcal{T}_{\text{plan}}$ was established. This includes patients who have recently been discharged from the hospital. In both cases, the transport requests needs to be incorporated into the already planned schedule.
We formalize the different levels of information:

Definition 3.6 (Semiplannable transport)**.** *A* semiplannable transport T *is a tuple* (O_T, D_T, d_T) *that consists of*

i) its origin O_T*, i.e., the place where a patient is picked up,*

ii) its destination D_T*, i.e., the place where a patient is dropped off, and,*

iii) its duration d_T*, i.e., the total duration of the transport from reaching* O_T *until leaving* D_T*.*

For a semiplannable transport, no target time is known. The set of semiplannable transports is denoted as $\mathcal{T}_{\text{semi}}$*.*

Definition 3.7 (Ad hoc transport)**.** *Let* $\mathcal{P}$ *be a (weakly) feasible schedule for* $\mathcal{T}_{\text{plan}}$*. An* ad hoc *transport* T *is a transport that has to be planned after vehicles have already started carrying out* $\mathcal{P}$*. The set of ad hoc transports is denoted as* $\mathcal{T}_{\text{adhoc}}$*.*

The distinction between plannable, semi-plannable and ad hoc transports and the different level of information leads to a natural division of the scheduling process into two parts.
In the planning phase, the transports of $\mathcal{T}_{\text{plan}}$ are scheduled. After the planning phase, when the schedule is carried out, the operational phase begins. While the transports of $\mathcal{T}_{\text{plan}}$ are carried out, transports of $\mathcal{T}_{\text{adhoc}}$ need to be incorporated and the schedule usually has to be updated. For transports $T \in \mathcal{T}_{\text{semi}}$, no target time is yet defined. When one treats them as ad hoc transports, the dispatcher ignores them until their target time becomes known.
Alternatively, the estimated target time t_T^{est} could guide scheduling, treating the transport as a plannable transport in $\mathcal{T}_{\text{plan}}$. If the actual target time differs, the dispatcher adjusts accordingly. If earlier, it is treated ad hoc; if later, the vehicle waits. However, since estimated times may vary significantly, a maximum waiting time is introduced. If a vehicle exceeds this waiting time, it is reassigned, treating the transport as ad hoc and deleting the dummy transport. This is formalized as follows:

Definition 3.8 (Dummy transport)**.** *A transport* T *is called a* dummy transport*, if*

i) $T \in \mathcal{T}_{\text{semi}}$*, i.e., it is semiplannable,*

ii) its target time is estimated with $t_T^{\text{est}} \geqslant 0$*, and,*

iii) it has a waiting time $w_T \geqslant 0$ *after which* T *will be treated as an ad hoc transport.*

The set of dummy transports is denoted as $\mathcal{T}_{\text{dummy}}$*.*

To estimate the target time, historical data or medical expertise can be used. An example for a scenario where dummy transports are applicable in practice in our real-world example are return trips from dialysis. We come back to this in Subsubsection 4.2.1.
This concludes the discussion of the patient transport problem. In the following section, we model the specific optimization problems and describe how they can be solved, also when incorporating semiplannable and ad hoc transports.

4 Mathematical optimization

The patient transport problem described in the previous section is now modeled as a VRPGTW. This type of modeling has proved to be the most appropriate for our application's practical requirements and standards as discussed in Subsection 2.1. We begin by modeling the plannable patient transport problem in Subsection 4.1. Then, we extend this formulation for semiplannable and ad hoc transports in Subsubsection 4.2.1 before we elaborate on the modeling of requirements of the Covid-19 pandemic in Subsubsection 4.2.2.

4.1 Modeling the patient transport problem as a VRPGTW

To model our problem, we use the VRPGTW formulation from Subsection 2.1. Assume we have n plannable transports (Definition 3.1) from $\mathcal{T}_{\text{plan}}$ that have to be scheduled with m vehicles. We begin by explaining how the plannable model is created.
The parameters of the VRPGTW are defined as follows:

- $N := \{1, \dots, n\}$ is the set of plannable transports. Each node i corresponds to transport T_i. The depot is denoted as node 0 and has a copy $n + 1$. The vehicles start at 0 and end their trip at $n + 1$. This avoids modeling issues and does not have other implications.
- $K := \{1, \dots, m\}$ is the set of available vehicles.
- $A_N := \{(i, j) \in N \times N : i \neq j\}$ is the set of arcs 'between two transports'. An arc $(i, j) \in A_N$ is used if and only if a vehicle carries out T_j directly after transport T_i.
- $A_K := \{(0, j) : j \in N\} \cup \{(i, n + 1) : i \in N\} \cup \{(0, n + 1)\}$ is the set of arcs from the depot to all $j \in N$, from each $i \in N$ to its depot and the arc $(0, n + 1)$ that is used by a vehicle if it does not transport any patients.
- The digraph is $G := (V, A)$ with $V := N \cup \{0, n + 1\}$ and $A := A_N \cup A_K$.
- For each $i \in N$, t_i is the target time of transport T_i.
- For $k \in K$, $[a_k, b_k] \subseteq \mathbb{R}_+$ denotes the shift of the driver(s) of vehicle k.
- For $i, j \in N$, $\tau_{i,j} \geqslant 0$ denotes the time a vehicle needs to reach O_j after starting transport T_i, i.e., the sum of d_i and the travel time $\text{dist}_{i,j} \geqslant 0$ from D_i to O_j.
- For $j \in N$, $\tau^k_{0,j} > 0$ denotes the travel time for vehicle $k \in K$ to reach O_j from its depot and $\tau^k_{j,n+1} > 0$ denotes the travel time for vehicle $k \in K$ to reach its depot from D_j. We set $\tau^k_{0,n+1} := 0$ for all $k \in K$. We need the superscript k here, since there can be multiple depots.

The variables of our model are:

- $x_{i,j,k} \in \{0, 1\}$ is the binary variable that indicates whether vehicle k travels from node i to j, i.e., whether vehicle k carries out transport T_j directly after it carries out transport T_i.
- For all $i \in N$, $y_i := y_{i,k} \in \mathbb{R}_+$ denotes the time when vehicle k arrives at node i, i.e., when transport T_i starts. We only need one variable for each transport T_i as at most one vehicle serves it.
- $y_{0,k} \in \mathbb{R}_+$ denotes the time when a vehicle starts its trip and $y_{n+1,k} \in \mathbb{R}$ denotes the time when it ends its trip. Here we need a variable for each vehicle.

As a choice of an objective function that evaluates the quality of the schedule adequately, we use piecewise linear measures of quality that we will linearize whenever necessary. To simplify notation, we thus introduce the following concept for penalty weights:

Definition 4.1 (Penalty weights). *Let Λ be a set of variables. A* penalty weight *is a parameter $\gamma \in \mathbb{R}$ that is used to penalize all variables $\lambda \in \Lambda$. The* penalty set *Γ contains the tuples (γ, Λ). Using this representation, an objective function has the form*

$$\sum_{(\gamma,\Lambda)\in\Gamma} \sum_{\lambda\in\Lambda} \gamma \cdot \lambda.$$

Since our goal is to create a fair schedule, we attempt to minimize the maximum delay throughout the day. To this end, we introduce a penalty parameter $\gamma_{\max} > 0$. Further, we minimize the individual delays (possibly weighted by γ_i), yielding the (piecewise linear) measure of quality

$$\gamma_{\max} \cdot \max\{0, y_1 - t_1, \dots, y_n - t_n\} + \sum_{i \in N} \gamma_i \max\{0, y_i - t_i\} \tag{3}$$

as we defined it in Definition 3.4. Objective function (3) can be easily linearized, as one can see in the Objective function (4a) and Constraints (4h), (4i) of the following MIP where our (linearized) instance is formulated:

$$\begin{aligned}
\min \quad & \gamma_{\max} \cdot \Delta^{\max} + \sum_{i \in N} \gamma_i \Delta_i && && \text{(4a)}\\
\text{s.t.} \quad & \text{Constraints (1b)} - \text{(1e)}, && && \text{(4b)}\\
& y_i + \tau_{i,j} - y_j \leqslant M_{i,j}(1 - x_{i,j,k}) && \forall k \in K, (i,j) \in A_N, && \text{(4c)}\\
& y_{i,k} + \tau_{i,j}^k - y_{j,k} \leqslant M_{i,j}(1 - x_{i,j,k}) && \forall k \in K, (i,j) \in A_K, && \text{(4d)}\\
& a_k \leqslant y_{0,k} && \forall k \in K, && \text{(4e)}\\
& y_{n+1,k} \leqslant b_k && \forall k \in K, && \text{(4f)}\\
& t_i \leqslant y_i && \forall i \in N, && \text{(4g)}\\
& y_i - t_i \leqslant \Delta_i && \forall i \in N, && \text{(4h)}\\
& 0 \leqslant \Delta_i \leqslant \Delta^{\max} && \forall i \in N, && \text{(4i)}\\
& x_{i,j,k} \in \{0,1\} && \forall k \in K, (i,j) \in A, && \text{(4j)}\\
& y_{i,k} \geqslant 0 && \forall i \in N, k \in K. && \text{(4k)}
\end{aligned}$$

We can represent the current penalties for the delays using the penalty weights of Definition 4.1 by

$$\Gamma := \{(1, \{\Delta_i \mid i \in N\})\} \cup \{(\gamma_{\max}, \{\Delta^{\max}\})\},$$

so we do not state the delays explicitly in the following.
The remaining constraints of Model (4) that are not directly taken over from the general formulation model the following: Constraints (4e) and (4f) are hard bounds that ensure shift times are met and feasible schedules as introduced in Definition 3.2 are obtained. Constraint (4g) prevents a vehicle from starting a transport before its target time.
With the solution of Model (4), we can consequently establish a feasible schedule: If $x_{i,j,k} = 1$ then vehicle k carries out transport T_j after carrying out transport T_i. Thus, from the optimal solution (x^*, y^*), it is possible to reconstruct the path of vehicle k from 0 to $n+1$ to the form $(v_{k_1}, \dots, v_{k_s})$, where $s \in \mathbb{N}$ denotes the number of transports for the respective vehicle. Along with the optimal arrival times y^*, an optimal schedule is obtained.

4.2 Extending the plannable formulation

So far, we have modeled the problem of finding optimal schedules for transports for which all information is known, i.e., plannable transports. This model is now extended to handle transports with incomplete information, including ad hoc transports. In addition, using the Covid-19 pandemic as an example, we propose and model how to treat the situation of endemic diseases, or, to be more precise, when a patient is at least suspected to be infected with a highly infectious disease. We begin by introducing labels for transports:

Definition 4.2 (Transport labels)**.** *A* (transport) label *is a function* $\phi\colon \mathcal{T}_{\text{plan}} \cup \mathcal{T}_{\text{semi}} \cup \mathcal{T}_{\text{dummy}} \cup \mathcal{T}_{\text{adhoc}} \to \{0,1\}$ *that indicates whether a transport* $T \in \mathcal{T}_{\text{plan}} \cup \mathcal{T}_{\text{semi}} \cup \mathcal{T}_{\text{dummy}} \cup \mathcal{T}_{\text{adhoc}}$ *fulfills some property. Multiple labels can be collected in a* label set Φ.

Label information may include whether a transport involves infectious diseases, needs to fulfill certain priorities, or the question whether certain equipment is needed. Another example can be the information about the type (plannable, semiplannable, ad hoc, dummy) of a transport. Recall that all types of transports are defined in Definitions 3.1, 3.6, 3.7 and 3.8. A similar set can also be introduced for vehicles, e.g., for indicating whether certain equipment is available. In this case, only such vehicles can handle transports that need this equipment. Using the definition of labels from Definition 4.2, we can introduce additional constraints for the patient transport problem on an abstract level. These constraints can be hard or soft, the latter using penalty weights and additional variables.

Now we are presenting different modeling possibilities that allow us to extend Model (4) in various ways. Let Φ be a set of labels. One possibile extension is to limit the number of transports per vehicle with a label $\phi \in \Phi$ to a fixed number $a \in \mathbb{N}$. This can be modeled by

$$\sum_{(i,j)\in A} x_{i,j,k} \cdot \phi(T_i) \leqslant a \ \forall k \in K. \tag{5}$$

If a specific label is not of importance then we can just omit it. With $a = 1$, each vehicle can handle at most one transport of a specific type. Let $\phi_1, \phi_2 \in \Phi$ and assume that it is not allowed to handle a transport j with $\phi_2(T_j) = 1$ immediately after a transport T_i with $\phi_1(T_i) = 1$. This is represented by adding constraints

$$\phi_1(T_i) \cdot \phi_2(T_j) \cdot x_{i,j,k} = 0 \ \forall k \in K, (i,j) \in A. \tag{6}$$

to Model (4). Naturally, it is also possible to set $\phi_1 = \phi_2$ and, thus, prohibit the handling of similar transports after another.

4.2.1 Incorporation of transports with incomplete information

We now look at how to modify the plannable formulation, first to handle semiplannable transports, and then to handle ad hoc transports. Therefore, we apply the strategy for the operational phase that have been proposed in Section 3. In fact, similar to the planning phase, this means creating or extending the VRP formulation and solving the resulting MIP.

Whenever possible, we estimate the target times of semiplannable transports and treat dummy transports like plannable transports. In other words, in Model (4), N is modified such that the nodes corresponds to the transports in $\mathcal{T}_{\text{plan}} \cup \mathcal{T}_{\text{dummy}}$ and the extended VRP is solved. To take the data-driven nature of dummy transports into account, we establish the following rules: Firstly, for the maximum delay as a measure of quality, delays of dummy transports are not taken into account, i.e., only the delays of the initial plannable transports are relevant. Secondly, the individual delay of dummy transports is a quality of measure and is weighted by some parameter γ_i for $T_i \in \mathcal{T}_{\text{dummy}}$. Thus, after incorporating the dummy transports, Objective (4a) of Model (4) is

$$\gamma_{\max} \cdot \Delta^{\max} + \sum_{\substack{i\in N,\\ T_i\notin \mathcal{T}_{\text{dummy}}}} \gamma_i \Delta_i + \sum_{\substack{i\in N,\\ T_i\in \mathcal{T}_{\text{dummy}}}} \gamma_i \max\{0, y_i - t_i^{\text{est}}\}. \tag{7}$$

With the scheduling of $\mathcal{T}_{\text{plan}}$, the planning phase is completed.

In addition to semiplannable transports, the dispatchers usually need to incorporate ad hoc transports. They are called whenever such a transport has to be scheduled at a certain time σ on the fly. With a schedule already in place, some vehicles are already on their tour. Therefore, we need to reoptimize our schedule.

We now introduce some notation. Let $\sigma \in \mathbb{R}_+$ be a given point of time. The set $V^\sigma \subseteq V$ denotes the nodes that correspond to transports known before an ad hoc transport is requested at σ. Using the optimal solution (x^*, y^*) of Model (4) at σ, the set

$$\mathcal{P}^\sigma := \left\{(i,k,y_i^*) \in V^\sigma \times K \times \mathbb{R}_+ \mid \exists j \in V^\sigma : \ x^*_{i,j,k} = 1\right\}$$

denotes the schedule that is carried out at σ. If no ad hoc transports have been requested before σ, we have $\mathcal{P}^\sigma = \mathcal{P}^0$, i.e., the solution of the plannable model. Transports already in operation are not changed. They are elements of the set

$$\mathcal{P}^\sigma_{\text{fixed}} := \left\{(i,k,y_i^*) \in \mathcal{P}^\sigma \mid \forall j \in V^\sigma : x^*_{i,j,k} = 1 \Rightarrow \sigma \geqslant y^*_{j,k} - \text{dist}_{i,j}\right\}$$

that contains all transports where the vehicle is either on the way from D_{T_i} to O_{T_j} or has started or finished transport T_j. If an ad hoc transport is requested for some time point σ then transports of $\mathcal{P}^{\sigma}_{\text{fixed}}$ are not changed while the transports in $\mathcal{P}^{\sigma} \setminus \mathcal{P}^{\sigma}_{\text{fixed}}$ are removed of the schedule. This ensures that vehicles not carrying out transports at σ are available again since the schedule after σ was deleted. Thus, every transport scheduled after σ, including the ad hoc transport, can be rescheduled by solving Model (4) with the additional restriction that all transports in $\mathcal{P}^{\sigma}_{\text{fixed}}$ are unchanged. Therefore, we first update G by including the new ad hoc transport in V^{σ} and $A^{\sigma} := \{(i,j) \in A : i,j \in V^{\sigma}\}$, respectively. Thereupon, to ensure that fixed transports are not changed, we introduce two additional constraints to Model (4), namely

$$x_{i,j,k} = x^*_{i,j,k} \;\forall (i,j,k,y^*_i) \in \mathcal{P}^{\sigma}_{\text{fixed}} \tag{8}$$

and

$$y_j = y^*_j \;\forall (i,j,k,y^*_i) \in \mathcal{P}^{\sigma}_{\text{fixed}}. \tag{9}$$

If we have dummy transports, we need to be cautious about their waiting times since, whenever a transport $T \in \mathcal{T}_{\text{dummy}}$ exceeds its waiting time, it is deleted and treated as an ad hoc transport (once requested). Then, in order to use the vehicle that should have handled T, we reoptimize our schedule.

4.2.2 Adjusting the model during the Covid-19 pandemic

To decrease the risk of infections for patient transports during the Covid-19 pandemic, it is desirable that transports are distinguished into two types, depending on whether a patient is (suspected to be) infected with Covid-19 or not, resulting in an additional property of transports. From now on, we refer to patients with a suspected infection with Covid-19 as 'infected' as well. In the following, we will discuss some ideas how such Covid-19 transports can be handled and proceed with a mathematical formulation how minimizing the risk of infections can be incorporated in our Model (4) using Definition 4.2.

There are different possibilities to decrease the risk of infections. In practice, when dealing with dangerous infectious diseases like Covid-19, the staff needs to wear protective clothing and the vehicle is disinfected after every transport. In addition to these protective measures, it can be helpful to minimize the number of changes from a Covid-19 transport to a non-Covid-19 transport. This will reduce the number of contacts between patients and staff and thus the risk of infection even further, even if vehicles are disinfected after each transport.

We present two different approaches: reducing the number of vehicles that are allowed to carry infected patients and limiting the number of infected patients per vehicle. In addition to the goal of reducing the number of changes, these approaches have been developed due to the fact that the supply of protective clothing is limited and therefore should be distributed as efficiently as possible. Nevertheless, in both cases, we still aim to minimize the delays for patients. This is incorporated by using different penalty parameters for, e.g., the delay and the number of changes in the objective function.

The additional time for disinfection and changing of clothes needs to be taken into account when establishing schedules. This is easily done by increasing the duration d_i of each transport by a constant. Thus, it is not necessary to introduce additional constraints to Model (4).

To indicate, for which transports additional Covid-19 requirements are necessary, we introduce a label $c\colon \mathcal{T} \to \{0,1\}$ where $\mathcal{T} \subseteq \mathcal{T}_{\text{plan}} \cup \mathcal{T}_{\text{semi}} \cup \mathcal{T}_{\text{dummy}} \cup \mathcal{T}_{\text{adhoc}}$. A value of 1 corresponds to a patient's infection. This label is used to create additional constraints. For sake of notation, we write c_i instead of $c(T_i)$. Furthermore, we write $\bar{c}$, meaning a transport is not a Covid-19 transport, with $\bar{c}_i = 1 - c_i$ for all $T_i \in \mathcal{T}$. To model travels from and to the depot adequatly, we further define $c_0 := c_{n+1} := 0$ and, consequently, $\bar{c}_0 := \bar{c}_{n+1} := 1$.

Minimizing the number of changes. To formalize minimizing the number of changes, we use the labels c and $\bar{c}$ with Constraint (6):

$$c_i \cdot \bar{c}_j \cdot x_{i,j,k} = 0 \;\forall k \in K,\; (i,j) \in A_N. \tag{10}$$

Equation (10) models that no vehicle is allowed to carry a non-Covid-19 transport directly after a Covid-19 transport. We do not wish to prohibit all changes because otherwise, the solution quality w.r.t. the delay would decrease or we would not able to create feasible schedules at all. Therefore, we use a soft constraint. Instead of minimizing the number of changes, we introduce the additional variable $\lambda_{i,j,k}^{\text{change}} \in \{0,1\}$ for $(i,j) \in A_N$, $k \in K$ and modify (10) accordingly:

$$c_i \cdot \bar{c}_j \cdot x_{i,j,k} = \lambda_{i,j,k}^{\text{change}} \; \forall k \in K, \; (i,j) \in A_N. \tag{11}$$

Now the variables $\lambda_{i,j,k}^{\text{change}}$ are penalized using γ^{change} and we modify Γ by appending the tuple $(\gamma^{\text{change}}, \{\lambda_{i,j,k}^{\text{change}} \mid (i,j) \in A_N, k \in K\})$.

Approach 1: Minimizing the number of Covid-19 vehicles. An option to reduce contact between uninfected and infected persons is dividing the vehicle fleet into different pools, i.e., one pool of vehicles that only handle Covid-19 transports and one pool for non-Covid-19 transports. It is possible to additionally use so-called floater vehicles that are allowed to handle both types of transports to maintain some degree of flexibility. Here, the protective clothing can be distributed among the Covid-19 vehicles and the floater vehicles so that no vehicle is carrying it needlessly. The corresponding constraint is

$$c_i \cdot x_{i,j,k} \leqslant \lambda_k^{\text{vehicle}}, \; \forall k \in K, \; (i,j) \in A. \tag{12}$$

Here, for $k \in K$, $\lambda_k^{\text{vehicle}}$ are new binary variables. They are penalized using $\gamma_k^{\text{vehicle}}$, i.e., we update $\Gamma \leftarrow \Gamma \cup (\gamma^{\text{vehicle}}, \{\lambda_k^{\text{vehicle}} \mid k \in K\})$.

Approach 2: Limiting the number of Covid-19 transports for each vehicle. Another approach to incorporate Covid-19 requirements is to distribute protective clothing equally among all vehicles. In this case, every vehicle is able to serve a limited number of Covid-19 transports before it needs to return to its depot to obtain new sets of clothing. An advantage of this approach is that each vehicle is able to carry out Covid-19 transports with less delay than when separating the fleets completely. Nevertheless, this modeling approach might increase the number of switches between infected and non-infected patients in vehicles.
To implement the limitation of protective clothing, we use a tuple of a penalty weight and integer variables $\lambda^{\text{clothing}} \in \mathbb{N}_0^{|K|}$, namely $(\gamma^{\text{clothing}}, \Lambda^{\text{clothing}})$ where $\Lambda^{\text{clothing}} := \{\lambda_k^{\text{clothing}} \mid k \in K\}$. The objective function is increased by the penalty value every time a vehicle has to return to the depot to obtain new clothing. Thus, for all $k \in K$, we introduce the constraint

$$\sum_{(i,j) \in A} c_i \cdot x_{i,j,k} \leqslant \alpha(1 + \lambda_k^{\text{clothing}}) \tag{13}$$

where $\alpha \in \mathbb{N}$ is the amount of sets of clothes per vehicle. Inequality (13) is Constraint (5), with the difference that not every exceedance of α is penalized. Instead, every time α is reached again, we increment the variable $\lambda_k^{\text{clothing}}$ by one. In our model, it is not possible that vehicles return to the depot during their shift, and, thus, we prohibit this by choosing the penalty for this scenario quite high. This means that a vehicle only exceeds its limit when it is not avoidable, i.e., if there are more than $\alpha|K|$ Covid-19 transports. However, this never happens in our numerical study.
This concludes the modeling of the patient transport problem. In the next section, we present and discuss our numerical experiment and show the efficiency of our plannable approach and its extensions.

5 Implementation and numerical results

In this section, we provide details on the implementation as well as the insights that can be obtained from the optimized schedules. The models presented in the earlier section are solved via state-of-the-art available global MIP solvers like Gurobi [24] or SCIP [9] which are applying solution approaches for MIPs, cf. [33], [58] or [59]. The rationale behind the usage of available solvers is to enable possible transfer of the developed approaches to the practitioners so that they

can also maintain the program in the future. Moreover, using a MIP formulation, we are able to incorporate the extensions for pandemic requirements.
All models and algorithms were implemented in Python 3.7.7. To solve the MIPs, we used Gurobi 9.0.2 on the NHR@FAU clusters with Intel Xeon E3-1240 v5 or Intel Xeon E3-1240 v6 CPUs, respectively. Each of these has four cores with 3.5 GHz each and a RAM of 32 GB.
We start with the description of a simulated reality to evaluate our schedules in Subsection 5.1. Thereupon we present some heuristic approaches in Subsection 5.2 and, finally, we evaluate the performance of our approach in Subsection 5.3. There, we also cover the incorporation of semiplannable transports and the extensions to cover issues and problems for patient transports during the Covid-19 pandemic.
For all numerical results, we use historical data from 2019 to mid-2021, provided by the ILS that covers regions in Middle Franconia. In practice, this area is divided into different counties, and each one is scheduled individually, with a transport T assigned based on its origin location O_T. With our optimization strategy, we proceed in a similar manner. Counties are different in size and population density. We have rural counties with a low population density in comparison to their size, as well as urban counties with a high population density, what causes a higher number of transports and, thus, more difficult instances.
We optimize each day separately because they do not influence each other as during the night almost no transports are requested. Unless otherwise mentioned, we specify a time limit of 60 minutes. We will elaborate on the time limit in Subsection 5.3.
We must overcome some issues with the available historical data: On the one hand, some transports are stated incorrectly, such as missing timestamps or locations. Missing data have been handled during preprocessing by either estimating travel times using a distance matrix or deleting the corresponding transports if too much data were missing. For the estimation of travel times, we refer to [34]. On the other hand, there are cases where the dispatcher needs to make decisions on the fly that could have not been planned beforehand. For example, vehicles, can be lent between different counties. Technically, each vehicle is assigned to a specific county but in exceptional circumstances and if necessary, this can be relaxed. Further, the distinction between patient and rescue transports can be neglected when absolutely necessary. Moreover, if a transport's delay is excessive, it can be postponed for another day. We do not consider any decisions out of the set of rules or usual decision makings in our model since we are not in a position to make them. In practice, the set of feasible schedules can possibly be improved by incorporating expert decisions.

5.1 Implementation of a simulated reality

For the reason mentioned above, we cannot compare our schedules directly to the ones in the historical data. So, in order to evaluate the optimized schedules, we require a baseline solution.
To ensure the most accurate comparison, we have implemented a simulation of the ILS's decision making that operates similarly to the practice. They apply a greedy approach: Every time a transport is necessary, the vehicle that would arrive the fastest is assigned to it. However, vehicles currently involved in a transport cannot be used, and shift times need to be respected whenever possible.
For the baseline implementation, we thus sort all plannable transports by their target time given in the historical data. Then the vehicles are assigned in that order: For each vehicle, we calculate the earliest time it could reach the requested transport's origin by adding the travel time to the time it is expected to become available, which is either the start of its shift or the expected end of the previous transport. Furthermore, we check whether the vehicle could reach its depot without violating its shift times using the estimated duration, i.e., expected travel times, of the transport. The vehicle that can carry out the transport request with the smallest delay is assigned to it. If there is more than one vehicle with minimal delay then the one with the shortest travel time is chosen. In the case that no vehicle can handle the transport without violating its shift times, we choose the vehicle with the smallest shift violation time.

5.2 Heuristic methods for the MIP solver

We present some algorithmic approaches for speeding up the MIP solution process for the arising models. This is necessary because, while conducting our numerical study (see Subsection 5.3), we have realized that without any improvements, our solution approach is incapable of solving many instances to optimality. For the evaluation of the heuristic methods, we use a subset of real historical data, namely data corresponding to 59 days (January and February 2020) from two counties, for a total of 118 instances. The counties differ in size; the smaller one has about 100,000 residents, the larger one about 500,000.

Using a primal heuristic. The first heuristic we have implemented is a primal heuristic that aims to find good feasible solutions early in the MIP solution process. At every kth node of the branch-and-bound tree within the MIP solver, we tentatively round a part of the optimal solution of the LP relaxation. In every feasible solution, the number of variables $x_{i,j,k}$ that are set to 1 is given by $|K| + |V|$. For a primal heuristic, all values for $x_{i,j,k}$ in the solution of the LP relaxation are sorted and a number of them are fixed in order to faster obtain a feasible MIP solution. Here, we have chosen the number of vehicles $|K|$ and set the $|K|$ greatest variables to 1. If a feasible solution is found by using this start solution then it is used for the following solving process. Otherwise, it is discarded. It is important not to fix too many variables as the solution may then become infeasible since we could, e.g., accidentally schedule two vehicles to the same transport. However, we also note that fixing too few variables will not speed up the solving process and that the respective values are based on empirical analysis.

Using a greedy heuristic to obtain a first feasible solution. The simulation of the current scheduling approach given in Subsection 5.1 creates an (at least weakly) feasible schedule as defined in Definitions 3.2 and 3.3. It can be computed very quickly. Thus, it can be used as a starting solution for the variables $x_{i,j,k}$. The remaining variables depend on the values for $x_{i,j,k}$ and, thus, are computed by the applied solver. This greedy heuristic can be applied once at the beginning of each (re-)scheduling process.

Determining a branching priority. Within each branching step, a variable is chosen from all binary and integer variables, depending on some measure that aims to create preferably small branch-and-bound trees, see for example [1]. We consider the determination of optimized plans under Covid-19 restrictions as described in Subsubsection 4.2.2.
When the values for the binary variables $x_{i,j,k}$ are decided, the unique values for the remaining binary and integer variables can easily be computed. The values for the remaining binary variables, i.e., Covid-19 vehicle and change labels, are directly induced by Constraints (11), (12) and (13). The values for y_i, $i \in N$, model the time when transport T_i starts. Starting at the depot, the values are computed successively. This is possible because each value only depends on its unique predecessor, so we can simply follow the paths of the vehicles. Thus, we obtain a branching priority and branch on the former variables first that we provide as input.

Handling of difficult instances. Especially on days where many transports need to be scheduled, the solving process of the resulting MIPs can take a long time. We have days with around 140 transports in our data, as presented in Section 1. While we can limit the total running time, we can also directly limit the time required for each MIP. To accomplish this, the solving process is interrupted when the time limit is reached and the current best solution is used for establishing a schedule at this time. To ensure that a (weakly) feasible solution is found before the time limit, we provide a start solution, namely the previously described solution of our greedy heuristic. The idea behind this is that short-term decisions can be dealt with directly while long-term decisions, i.e., decisions that are made several hours beforehand, can be postponed. Further possibilities for improving the VRPGTW's solving process, e.g. determining bounds of the optimal solution, are discussed in [12].

Comparison of the heuristic methods. We now evaluate the performance of the heuristics. In Table 1, both heuristic methods to enhance the MIP solver are compared to each other and

Table 1: Comparison of different speed up methods for smaller problems in the upper table and larger problems in the lower table. The unit s stands for seconds.

	Primal heuristic	Greedy heuristic	No heuristic
Average time for solved problems	$69.3s$	$58.0s$	$62.8s$
Solved problems in 60 minutes	55	55	55
Solved problems in 10 minutes	54	55	55
Solved problems in 5 minutes	53	55	55
Solved problems in 3 minutes	53	55	53
Average time for solved problems	$520.3s$	$437.8s$	$642.6s$
Solved problems in 60 minutes	27	33	31
Solved problems in 30 minutes	23	32	26
Solved problems in 15 minutes	21	27	22
Solved problems in 10 minutes	21	25	22

to the results when no such heuristic is used. We assess an easier set as well as a more difficult set with a larger number of transports. The easier test set requests around 20 transports per day, while the more difficult test set requests up to 65 transports per day. The instances are smaller on weekends. The number of available vehicles ranges between 5 and 20, depending on the weekday and the size of the county under consideration.
The number of binary and continuous variables are estimated by $\mathcal{O}(|K| \cdot |V|^2)$ and $\mathcal{O}(|K| \cdot |V|)$, respectively. The number of constraints is $\mathcal{O}(|K| \cdot |V|^2)$, i.e., also quadratic in the number of patients. In general, the number of transports is greater than the number of vehicles. On the easier problem set, as shown in the upper part of Table 1, all approaches are able to solve most of the problems to optimality. The application of a greedy heuristic is slightly faster than the primal heuristic but there are no significant differences.
In contrast, on the more difficult test set, the usage of the greedy heuristic yields a high improvement. Compared to the other approaches, instances are solved significantly faster. If we use the baseline solution then a weakly feasible solution is given at the root node, where many branch-and-bound nodes that yield a worse solution can be pruned. In fact, in most cases, the baseline solution is even feasible and not only weakly feasible, which further improves the running time of our algorithm.
The solution time is improved as a result of the branching priority from Subsection 5.2. This enables us to increase the number of problems solved within 60 seconds by around 50%. Note that if we do not consider Covid-19 requirements, this priority will have no impact. Thus, we always prioritize the variables $x_{i,j,k}$ for branching.

5.3 Improvements applying the VRPGTW model

For the following results, we applied the greedy heuristic together with a branching priority. The proposed results are chosen exemplary and are characteristically similar for other counties. In the following, we compare the historical data's solution and the existing course of action with our optimization approach. Firstly, we consider the results of our general model including ad hoc transports in Subsubsection 5.3.1. Secondly, in Subsubsection 5.3.2, we consider insights that can be drawn from these results. Finally, in Subsubsection 5.3.3 we present the exemplary results for our extensions for handling semiplannable transports and the Covid-19 pandemic.

5.3.1 Plannable and ad hoc transports without further requirements

We evaluate the optimized schedules against the baseline solution of Subsection 5.1. We use the data from 2019 as there have been no Covid-19 transports that distort the scheduling process. Here, semiplannable transports (Definition 3.6) are treated as ad hoc transports (Definition 3.7). We investigate semiplannable transports in Subsubsection 5.3.3.
In each plot, we compare the baseline solution to the optimized schedules regarding shift time violations and delays for different regions and all instances that could be solved within 60 minutes. Instances that are not solved to optimality at this time are very likely also not solvable to optimality within a larger time horizon since the MIPs' respective gaps are too large. All results are sorted according to our objectives, which are reducing shift time violations, maximum delay, and total delay. We recall that it is critical to adhere to the shift times of the drivers, so this is our primary goal.

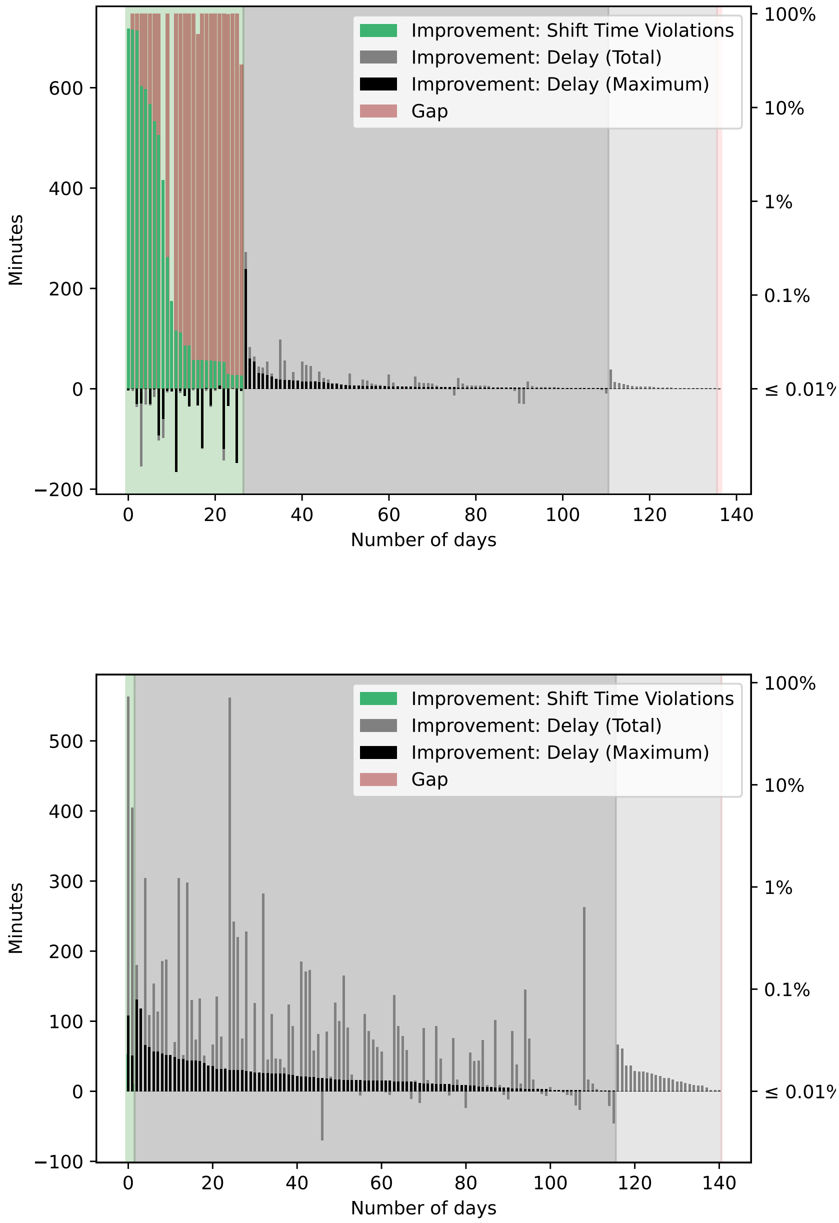


Figure 3: Improvement that can be gained using our optimization approach, when compared to the simulated reality. Two different regions are considered, the above one is a smaller urban region while the bottom one is larger but more rural. The areas from left to right correspond to: Decreased shift time violations (green), decreased maximum delay (dark gray), decreased total delay (light gray), increase in any metric (red). Further, we show the maximum MIP gap for all MIPs occurring in one instance.

In Figure 3, the results are compared for two different regions, Erlangen at the top and the county Erlangen-Höchstadt at the bottom. Each plot displays the differences in shift time violations and delays computed by the optimization model compared to the baseline. A positive value indicates that an improvement could be made, whereas a negative value indicates that some measure has increased with respect to the simulation of the current scheduling approach. The leftmost green section displays all instances in which shift times were improved. As a result, some transportation delays may increase. The second area shows all improvements in maximum delay. Again, the less prioritized measure, total delay, could worsen. The third area, shown in light gray, contains all results in which there is no difference in shift time violation or maximum delay between the optimal and baseline solutions. Thus, our approach only improved the overall delay. All instances in which no measurement changes are neglected.
Finally, the rightmost red part displays all instances in which we obtained a slightly worse solution. This could be due to a numerical error or uncertainties in the realization of transports what we will elaborate later on.
In the upper plot, an improvement with respect to shift times was achieved in more than 20 days. As respecting shift times is highly desirable, a somewhat longer maximum delay is acceptable. Apart from that, the optimized maximum delay is up to 60 minutes shorter than that of the baseline solution. In particular, there is one day for which the improvement is much larger than on the other days. On this day, many transports were requested for the late afternoon.
The lower plot shows results for problems in a larger but more rural county. In such an county there are less available vehicles but also a smaller number of requested transports. The optimized schedule often decreases the maximum delay by up to 60 minutes. Furthermore, on two days, there is a reduced number of shift time violations as well as a smaller maximum delay. Since we aim to only violate them if necessary, this is a great result. The overall delay is reduced by about 10 minutes on a daily average.
In addition to the objective improvements, we plot the maximum gap per day, which is the largest MIP gap observed across all VRPs that must be solved in a single instance. The gap is represented on a logarithmic scale. In our smaller instance set (bottom plot), all gaps are smaller than 0.01%. For the more difficult instance set, we only exceed this value in cases where we can improve the shift time violations. This could be due to the fact that we have another optimization goal to meet, and the complexity of these instances has, thus, increased significantly.
There are also instances where optimized plans perform worse than the baseline scheduling: On one day, the optimized schedule yields a maximum delay that is around ten minutes longer than that of the baseline schedule. This is caused by some ad hoc transports. The schedule of the optimized solution up to this point might be at least as good as the baseline schedule but the vehicles are organized differently. Then, although the baseline schedule is often not the optimal solution of Model (4), it could lead to a better final schedule when an ad hoc transport is requested. This is not completely preventable as there is no information about such transports earlier in the day. However, our approach concerning semiplannable transports from Subsubsection 4.2.1 aims to prevent such settings by including dummy transports for expected transports, and so this situation occurs only rarely.

Using a smaller time limit for each MIP. Now, instead of solving each MIP of an instance to optimality and aborting after a total time limit, we specify a time limit after which the solving process of a single MIP is aborted and the best, w.r.t. the objective value of Model (4), is used. For practical reasons, namely scheduling on the HPC cluster, there is still a total time limit of 24 hours, but this was never reached in our experiments.
Using this approach, it is possible to solve instances in larger counties. In these counties, we have to schedule up to 130 transports per day, but more vehicles are available. In the following, we evaluate the resulting schedules. Figure 4 shows a comparison to the simulated reality described in Subsection 5.1. The representation is similar to that in Figure 3. Here, we enforced a time limit of 60 minutes per MIP.
In many instances there is a much greater delay in our optimized schedule than in the baseline solution, which results from the fact that we are able to decrease the number and duration of shift time violations. Sometimes, it is even possible to find a feasible solution where the baseline solution is just weakly feasible. Thus, it proves to be possible to respect the drivers' shift times

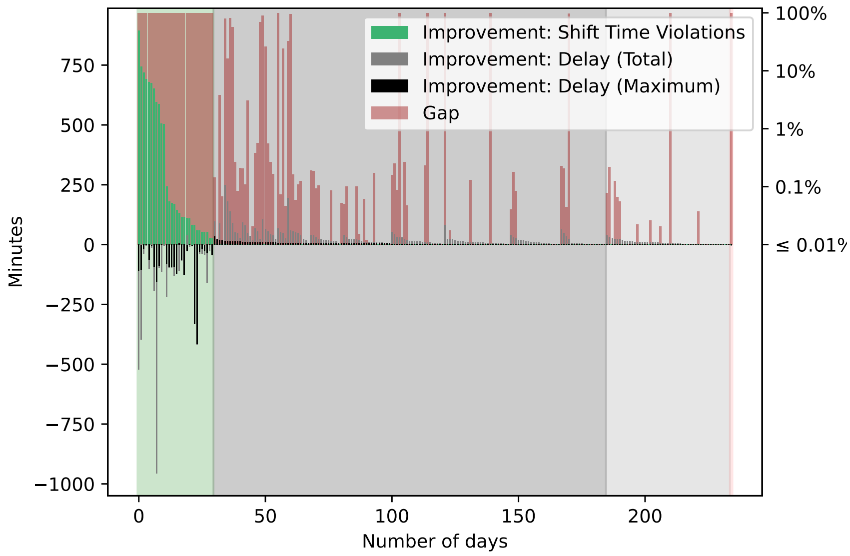
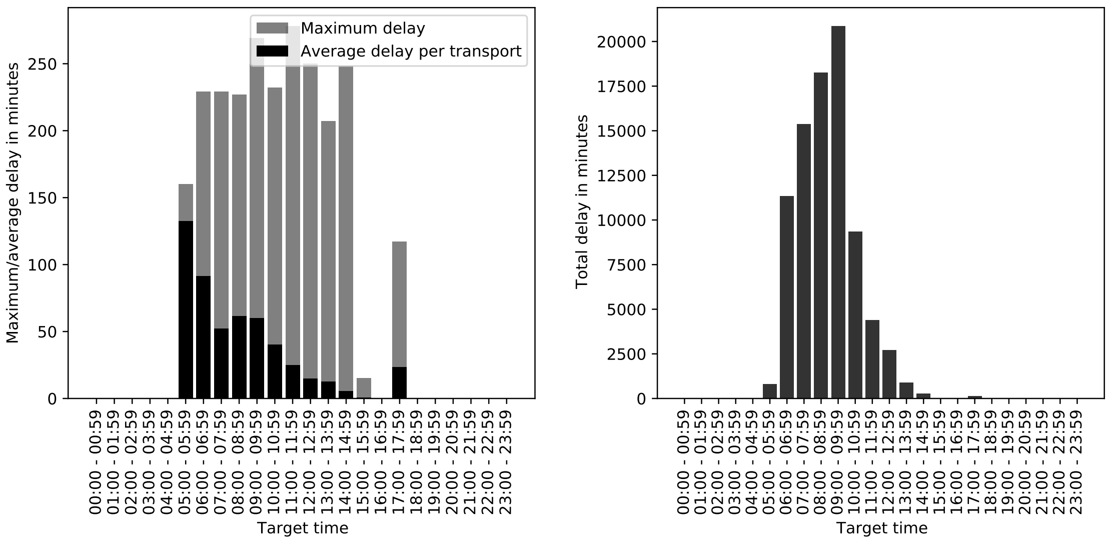


Figure 4: Improvement and gap for a larger county, similar to Figure 3

completely when applying our approach instead of a greedy approach. For the remaining problems, we decrease the maximum delay by up to 32 minutes, with an average improvement of around 5 minutes. In smaller counties this value is much higher.
In 289 of 364 instances, we found an optimal solution for at least 80% of all MIPs. An optimal solution for every MIP could be found in 267 instances. Consider the gaps in Figure 4 to evaluate solution qualities. Each day, we solve multiple MIPs and evaluate the MIP gap after either finding an optimal solution or meeting the 60-minute time limit. Similarly to Figure 3, the most significant gaps occur when drivers' shift times are difficult to meet. However, in this case, not all of the remaining instances could be solved to optimality. But it is worth noting that in most instances that were solvable to optimality within our time limit, whether they are weakly feasible or feasible, the optimal solution was found quite early in the solving process, although it took a long time for the solver to find an appropriate dual bound, i.e., to prove optimality. Thus, it is possible that in these cases with a large gap, we already discovered a (nearly) optimal solution but were unable to prove optimality.

5.3.2 Insights from the optimal solution

In the previous section we have evaluated the optimization results and discovered that they improve the (simulated) current course of action. Now, we investigate the optimized schedules in more detail.
We begin by examining how the delays vary over the course of the day. In Figure 5a, the maximum and average delay aggregated over the year is plotted for every hour. The average delay is computed using the number of transports. In Figure 5b, we further provide the total delay over the year.
The maximum delay is roughly equal throughout the day. From the late afternoon, there is nearly no delay, with one exception at 5pm. This is presumably caused by shift ends. The average delay is highest in the early morning and decreases continuously. We also consider the total delay on the right hand side. Here, we have high delays from 6am, not from 5am as for the maximum and average delay. While only a small number of transports are needed at this time, there are only few vehicles working a shift at this time, so the average delay for these patients becomes very high. This continues with the transports starting shortly after 6am. As the morning shift times do not start earlier than 6am and they first need to drive to the origin of a transport, it is not possible to reach these patients without a delay. In general, the night shifts end at 6am. Thus, they can probably not handle these transport requests without risking shift time violations.



(a) Maximum and average delay.

(b) Total delay.

Figure 5: Delays in the optimal schedules for instances of 2019, distinguished by target time.

A more detailed evaluation shows that the vehicles with earlier, i.e., morning shift times have to handle considerably more transports. Vehicles are working almost at full capacity at least until noon. Afterwards, less ad hoc transports are requested and the situation eases. This means that as soon as a vehicle becomes available, it will be assigned to some transport. As few vehicles are available at the beginning of a day, some transports cannot be handled on time. These delays then cause delays for later transports. This occurs because vehicles are occupied by earlier transports that are preferred as we minimize the maximum delay. In the afternoon, less transports are requested. Thus, more vehicles are available and the delays finally decrease.

5.3.3 Examples for semiplannable transports and further requirements

This section contains some sample schedules for our modeling approach extensions from Subsubsections 4.2.1 and 4.2.2. We have been unable to conduct a full numerical study due to a lack of data, but we were able to identify some exemplary days where our extensions worked well. For both, using dummy transports for semiplannable transports as well as incorporating Covid-19 requirements, we explain the problems with the given data before presenting some examples.

Using dummy transports for semiplannable transports While trying to handle semiplannable transports (Definition 3.6) using dummy transports (Definition 3.8), the following data issues have occurred: At the beginning of the optimization process, each transport T is assigned to the county of the origin location O_T. Often, the destination D_T is located in another county. In this case, patients have to be transported between different counties, causing outward and return trips optimized in different models. This leads to dummy transports that are not applied in the actual return trip. In practice, the dispatcher can assign such transports to the same county. Furthermore, in more recent data provided by the ILS, a type of dummy node is used for dialysis transports. As soon as a dialysis is requested, two transports are created, including the return trip on 23:59 on the same day. Its target time is corrected as soon as it becomes known. Thus, if a dialysis is created at least one day before, then it is a plannable transport but its return trip is not, although it is stated to be plannable in the data.
Due to these reasons, we provide an example day where the incorporation of a dummy transport leads to a significantly better result.

Example 5.1 (Advantages of using dummy transports)**.** *In Table 2, one can see two schedules that have been established on an example day in 2019. Table 2a is the current schedule shortly*

Table 2: A schedule before and after transport 20 has been replaced by its dummy node d28. The second dummy node, d24 for transport 24, does not correspond to a future transport and will thus not be replaced before the end of the day. The information about the return trip has become known at 11:00, all fixed transports are given below the line, so the remaining ones could be rescheduled. As can be seen, the delays remain the same for all transports. Note that there are relatively few transports in the afternoon as they are often ad hoc ones and thus are not known at this point.

(a) Schedule before 11:00.

ID (i)	t_i	k	start (y_i)	end	Δ_i
28	06:45	0	06:45	07:52	0
4	07:30	13	07:30	11:57	0
24	08:00	0	08:00	08:33	0
23	08:00	6	08:03	11:01	3.8629
25	08:30	0	08:33	10:00	3.2262
11	08:45	7	08:45	09:46	0
6	09:00	5	09:40	10:02	4.2565
0	09:00	8	09:40	11:10	4.2565
26	09:45	10	09:45	11:01	0
2	09:45	7	09:47	10:37	2.0320
13	10:00	10	10:45	11:59	45.0436
3	10:00	1	10:33	11:46	33.8629
12	10:00	0	10:00	10:58	0.5156
14	10:00	5	10:14	10:56	14.6932
15	10:00	7	10:37	11:48	37.1338
21	10:00	10	10:05	10:43	4.8170
7	10:15	5	10:59	12:18	44.0412
16	10:30	0	10:58	11:44	27.8051
9	10:30	11	11:03	12:30	33.8629
18	10:40	10	11:23	12:22	43.1832
1	11:00	0	11:44	12:31	44.0945
d28	11:52	1	11:52	12:59	0
d24	12:33	10	12:33	13:08	0
17	15:30	2	15:30	18:21	0

(b) Schedule at 11:00.

ID (i)	t_i	k	start (y_i)	end	Δ_i
28	06:45	0	06:45	07:52	0
4	07:30	13	07:30	11:57	0
24	08:00	0	08:00	08:33	0
23	08:00	6	08:03	11:01	3.8629
25	08:30	0	08:33	10:00	3.2262
11	08:45	7	08:45	09:46	0
6	09:00	5	09:40	10:02	4.2565
0	09:00	8	09:40	11:10	4.2565
26	09:45	10	09:45	11:01	0
2	09:45	7	09:47	10:37	2.0320
13	10:00	10	10:45	11:59	45.0436
3	10:00	1	10:33	11:46	33.8629
12	10:00	0	10:00	10:58	0.5156
14	10:00	5	10:14	10:56	14.6932
15	10:00	7	10:37	11:48	37.1338
21	10:00	10	10:05	10:43	4.8170
7	10:15	5	10:59	12:18	44.0412
16	10:30	0	10:58	11:44	27.8051
9	10:30	11	11:03	12:30	33.8629
18	10:40	9	11:23	12:22	43.1832
1	11:00	0	11:44	12:31	44.0945
20	12:00	7	12:00	13:10	0
d24	12:33	0	12:33	13:08	0
17	15:30	2	15:30	18:21	0

before 11:00. At 11:00, a new transport is requested. This transport is a return trip for a previous dialysis transport for which a dummy node has been created. The dummy node's target time was estimated based on the duration of previous dialysis appointments at the same location, i.e., in the same hospital. The outward journey has ID 28 and the dummy node is denoted by d28. It is determined that the new transport with ID 20 corresponds to this dummy node. Thus, it is deleted and replaced by the new transport using the new information, i.e., the actual target time.
Table 2b shows the schedule created at 11:00. As can be seen, the vehicles allocated to the transports not contained in $\mathcal{P}^{\sigma}_{fixed}$ for σ = 11:00, have changed. This is due to the fact that the transport is requested a little later than expected and vehicle 7 is available at that time. The schedule shows that it finishes its previous transport at 11:48, the new transport is requested at 12:00 instead of the presumed 11:52. So, this vehicle can reach the origin of the return trip in time.

The maximum delay in this example county is reduced by 20 minutes when compared to the schedule created without dummy nodes. As a result, the total delay decreases. In this case, we only create one dummy node, which is later replaced, and, the penalty weight was set to $\gamma = 0.5$ for dummy transports. Taking this into account, this is a very positive result, and an even greater improvement can be predicted if more dummies are used whenever appropriate data is available.

Table 3: Optimized pools for vehicles in one example area and the example shift times for Tuesdays. For each shift time, the number of handled transports, as well as the percentage of the allocation to each pool are given.

Shift time	Number of vehicles	Number of transports	Covid-19 vehicles	Non-Covid vehicles	Floater vehicles
06:00 - 14:00	1	189	2.9%	94.2%	2.9%
07:00 - 15:00	1	191	1.5%	91.2%	7.3%
08:00 - 16:00	1	161	2.9%	90.0%	7.1%
08:30 - 16:30	1	135	1.5%	93.9%	4.6%
09:00 - 17:00	2	237	0.8%	95.2%	4.0%
10:00 - 18:00	2	128	4.6%	94.3%	1.1%
10:30 - 18:30	1	119	2.9%	97.1%	0.0%
11:00 - 19:00	1	37	6.7%	93.3%	0.0%
14:00 - 22:00	1	69	7.3%	90.9%	1.8%
15:00 - 22:00	1	14	9.1%	90.9%	0.0%
16:00 - 24:00	1	20	10.5%	89.5%	0.0%
22:00 - 06:00	1	9	0.0%	100.0%	0.0%

Incorporating further requirements. Now, we consider schedules when dealing with Covid-19 transports. The dispatcher does not have a special course of action in the Covid-19 pandemic. Instead, it treats transports of infected patients in the same manner as other transports. As a result, evaluating our extension from Subsubsection 4.2.2 against the (simulated) reality is no longer useful, because explicit minimization of infection risks is not taken into account.
Instead, we aim to obtain new insights into how the pandemic requirements could be handled. The first example shows how a computed pool division can look, while the second one attempts to compare our two different approaches discussed in Subsubsection 4.2.2 to handling Covid-19 transports.
In a pool division, the first and second pools contain vehicles that transport either only Covid-19 patients or no Covid-19 patients at all. Floater vehicles make up a third pool. We evaluate the results and devise a strategy for implementing this pool division in practice.

Example 5.2 (Distribution of vehicles in an optimized pool division)**.** *We have collected all information about transports that took place in some county between 1st January, 2020 and 30th June, 2021. The percentage of Covid-19 transports is quite low, as days with very few (or even zero) transports are included. In Table 3, the pool division aggregated over all Tuesdays is given. Floater vehicles are mostly used in the earlier shifts. Later in the day, it is often possible to use Covid-19 vehicles. This might be caused by a higher number of transports in the morning, as discussed in Section 1.*

We can gain an intuition which vehicles are a good choice for fixed Covid-19 vehicles using such tables. Depending on the number of Covid-19 transports on one day, some of these vehicles should be assigned to the Covid-19 vehicle pool. Similar peculiarities are obtained for different combinations of counties and weekdays.
Because the exact distribution of transports, i.e., how many infected patients should be transported, is not known in advance, it is helpful to decide on the number of vehicles for each pool depending on the current number of Covid-19 cases. We assumed previously in Section 1 that these numbers can have a high correlation. Furthermore, it can be helpful to hold an additional floater vehicle in reserve to allow for any discrepancies.
We now compare the two options for handling Covid-19 transports that have been discussed in Subsubsection 4.2.2. In total, there are little data containing a high percentage of Covid-19 transports. Thus, we consider an example that yields the same maximum and total delay for both heuristic approaches, but the transports are handled very differently depending on the approach used. We will discuss these differences concerning the vehicle fleet.

Example 5.3 (Comparison of the approaches for handling Covid-19 transports)**.** *On our example day in April 2020, 45 transports are requested, 17 of which are for Covid-19 patients. Semiplannable transports are treated as ad hoc transports. For the second approach, limiting the number of Covid-19 transports per vehicle, we assume that each vehicle has two sets of protective clothing. In our example, there are 17 vehicles available. In both schedules, 15 of them are used. In particular, both approaches use the same vehicles. Those with a shift early in the morning are not required because no transport is requested.*

If we minimize the number of Covid-19 vehicles used, three of the fifteen vehicles are Covid-19 vehicles, with five additional vehicles serving as floater vehicles. There are four Covid-19 and six floater vehicles in the other case. Because the number of transports it can handle is limited, a larger pool is required (and not penalized). In fact, when minimizing the number of Covid-19 vehicles, some Covid-19 and floater vehicles carry out at least three Covid-19 transports, which is hardly penalized when distributing the clothing equally.
One thing that stands out in both schedules is that every vehicle that transports any infected patient ends its shift with all transports of infected persons. As a result, the penalty γ^{change} is never applied. The fact that this is possible with both approaches and also results in a relatively short delay is a positive result in terms of reducing infection risks. The longest delays occur later in the evening and cannot be avoided even if we do not include any Covid-19 requirements aside from increasing transport duration. This is due to a lack of available vehicles at the same time.

To summarize, both approaches have advantages and produce similar results especially when the number of Covid-19 transports is relatively low in practice. For example, the amount of protective clothing available may influence the dispatcher's approach. As this number grows, the need to consider this limitation diminishes and the dispatchers' focus may shift to minimizing contact between infected and non-infected patients and drivers. Another possibility is to combine both approaches, distributing protective clothing to all vehicles but providing more to those that are likely to have more Covid-19 transports. A fleet division, such as the one previously mentioned, could be useful in this regard.

6 Conclusion and future research

In this work, we have proposed a solution approach for scheduling patient transports that are not rescue transports. Information about these transports can be incomplete and may only be partly known several hours before they are required. Our objective is minimizing the delay for patients in a fair manner while respecting shift times. We apply a VRPGTW formulation that can then be solved by state-of-the-art MIP solvers.
We implemented the MIP formulation for the cases of full and incomplete information. We classify required transports into plannable transports (full information), semiplannable transports (almost full information but the target time is unknown) and ad hoc transports (no information about the transport at all). Ad hoc transports are incorporated by an iterative algorithm that solves Model (4) every time that full information about a transport becomes known. Semiplannable transports can, on the one hand, be treated like ad hoc transports or, on the other hand, by introducing dummy transports with an estimated target time. When using the second approach, they are treated like plannable transports. We have compared our modeling approach to the current scheduling practice of the dispatcher. Thereby, we have exemplarily observed that the waiting times in the optimized schedules are significantly lower than those obtained via a simulation of the current scheduling practice. To incorporate semiplannable transports where significant improvements can be seen, we require more data that includes such transports. Using the current data, we were only able to elaborate on some examples.
We have extended the model so that Covid-19 transports can be handled by different vehicle fleets. Still, the model remains solvable in real time and can be solved with MIP-based algorithms. We have outlined algorithmic approaches, which speed up the solution process.
In summary, we have proposed a formulation for the scheduling problem of patient transports that can be used in practice, also with further extensions to the pandemic situation. However, extensions are not limited to this application. We are able to decrease the delays for patients. Further, we can adhere to drivers' shift times more often than a simulation of the reality can, while, in almost all instances, even preserving smaller delays. With the availability of more data, it is expected that the proposed approach will work even better.
Several research directions are of interest for the future. As already mentioned, the usage of multi-objective optimization might be helpful as we have conflicting goals, e.g. minimizing delays and adhering to shift times. Another potential for improvement lies in incorporating semiplannable transports, where — assuming more data is available — other methods, e.g., further estimations of the duration and target time of transports, can be implemented. Furthermore, the usage of

dummy nodes can be extended, so that they can be created for more types of transport than those presented here for dialysis. Our approach can also be transferred to different scheduling or routing problems.

Acknowledgments

We are grateful for continuous support from the ILS, Nuremberg, and for many fruitful and stimulating discussions with Dr. Marc Gistrichovsky and Ludwig Fuchs (both ILS). Research reported in this paper was supported by project HealthFaCT under BMBF grant 05M16WEC. Furthermore, this paper has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 764759.
The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

Declarations

Funding

Research reported in this paper was supported by project HealthFaCT under BMBF grant 05M16WEC. Furthermore, this paper has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 764759.

Competing Interests

There are no competing interests.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Code availability

Not applicable.

Authors' contributions

Not applicable.

References

[1] T. Achterberg, T. Koch, and A. Martin. Branching rules revisited. *Operations Research Letters*, 33(1):42–54, 2005.

[2] M. Allen, A. Bhanji, J. Willemsen, S. Dudfield, S. Logan, and T. Monks. A simulation modelling toolkit for organising outpatient dialysis services during the covid–19 pandemic. *PLOS ONE*, 15(8):1–13, 08 2020.

[3] A. Beaudry, G. Laporte, T. Melo, S. Nickel, A. B. Jeppesen, G. Laporte, T. Melo, and S. Nickel. Dynamic transportation of patients in hospitals. *OR Spectrum*, 32(1):77–107, 01 2010.

[4] T. Bektas, P. P. Repoussis, and C. D. Tarantilis. *Chapter 11: Dynamic Vehicle Routing Problems*, pages 299–347. 2014.

[5] G. Berbeglia, J.-F. Cordeau, and G. Laporte. Dynamic pickup and delivery problems. *European Journal of Operational Research*, 202(1):8–15, 2010.

[6] E. Berhan, B. Beshah, D. Kitaw, and A. Abraham. Stochastic vehicle routing problem: A literature survey. *Journal of Information and Knowledge Management*, 13, 10 2014.

[7] D. Bertsimas, P. Jaillet, and S. Martin. Online vehicle routing: The edge of optimization in large–scale applications. *Operations Research*, 67(1):143–162, 2019.

[8] D. J. Bertsimas and G. van Ryzin. A stochastic and dynamic vehicle routing problem in the euclidean plane. *Operations Research*, 39:601–615, 8 1991.

[9] K. Bestuzheva, M. Besançon, W.-K. Chen, A. Chmiela, T. Donkiewicz, J. van Doornmalen, L. Eifler, O. Gaul, G. Gamrath, A. Gleixner, L. Gottwald, C. Graczyk, K. Halbig, A. Hoen, C. Hojny, R. van der Hulst, T. Koch, M. Lübbecke, S. J. Maher, F. Matter, E. Mühmer, B. Müller, M. E. Pfetsch, D. Rehfeldt, S. Schlein, F. Schlösser, F. Serrano, Y. Shinano, B. Sofranac, M. Turner, S. Vigerske, F. Wegscheider, P. Wellner, D. Weninger, and J. Witzig. The SCIP Optimization Suite 8.0. ZIB-Report 21-41, Zuse Institute Berlin, December 2021.

[10] D. Chen, S. Pan, Q. Chen, and J. Liu. Vehicle routing problem of contactless joint distribution service during covid-19 pandemic. *Transportation Research Interdisciplinary Perspectives*, 8:100233, 2020.

[11] Z.-L. Chen and H. Xu. Dynamic column generation for dynamic vehicle routing with time windows. *Transportation Science*, 40(1):74–88, 2006.

[12] J.-F. Cordeau, G. Desaulniers, J. Desrosiers, M. M. Solomon, and F. Soumis. *VRP with Time Windows*. 2002.

[13] J. F. Cordeau, G. Laporte, J. Y. Potvin, and M. W. Savelsbergh. Transportation on demand. *Handbooks in Operations Research and Management Science*, 14:429–466, 2007.

[14] J. F. Cordeau, G. Laporte, M. W. Savelsbergh, and D. Vigo. Vehicle routing. *Handbooks in Operations Research and Management Science*, 14:367–428, 2007.

[15] G. B. Dantzig and J. H. Ramser. The truck dispatching problem. *Management science*, 6(1):80–91, 1959.

[16] K. F. Doerner and R. F. Hartl. Health care logistics, emergency preparedness, and disaster relief: New challenges for routing problems with a focus on the austrian situation. In B. Golden, S. Raghavan, and E. Wasil, editors, *The Vehicle Routing Problem: Latest Advances and New Challenges*, pages 527–550. Springer US, Boston, MA, 2008.

[17] M. Ehrgott. *Multicriteria Optimization*. Springer-Verlag, Berlin, Heidelberg, 2005.

[18] F. Ferrucci, S. Bock, and M. Gendreau. A pro-active real-time control approach for dynamic vehicle routing problems dealing with the delivery of urgent goods. *European Journal of Operational Research*, 225(1):130–141, 2013.

[19] C. Fiegl and C. Pontow. Online scheduling of pick-up and delivery tasks in hospitals. *Journal of Biomedical Informatics*, 42:624–632, 8 2009.

[20] T. Flatberg, G. Hasle, O. Kloster, E. J. Nilssen, and A. Riise. Dynamic and stochastic vehicle routing in practice. *Operations Research/ Computer Science Interfaces Series*, 38:41–63, 2007.

[21] N. S. Gamchi, S. A. Torabi, and F. Jolai. A novel vehicle routing problem for vaccine distribution using sir epidemic model. *OR Spectrum*, 43:155–188, 3 2021.

[22] M. Gendreau, F. Guertin, J.-Y. Potvin, and E. Taillard. Parallel tabu search for real-time vehicle routing and dispatching. *Transportation Science*, 33(4):381–390, 1999.

[23] M. Gendreau, G. Laporte, and F. Semet. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27(12):1641–1653, 2001.

[24] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2021.

[25] F. T. Hanshar and B. M. Ombuki-Berman. Dynamic vehicle routing using genetic algorithms. *Applied Intelligence*, 27:89–99, 8 2007.

[26] S. C. Ho, W. Y. Szeto, Y. H. Kuo, J. M. Leung, M. Petering, and T. W. Tou. A survey of dial-a-ride problems: Literature review and recent developments. *Transportation Research Part B: Methodological*, 111:395–421, 5 2018.

[27] P. J. Hulshof, N. Kortbeek, R. J. Boucherie, E. W. Hans, and P. J. Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. *Health Systems 2012 1:2*, 1:129–175, 12 2012.

[28] T. Ibaraki, S. Imahori, M. Kubo, T. Masuda, T. Uno, and M. Yagiura. Effective local search algorithms for routing and scheduling problems with general time–window constraints. *Transportation Science*, 39(2):206–232, 2005.

[29] Y. Kergosien, C. Lenté, D. Piton, and J.-C. Billaut. A tabu search heuristic for the dynamic transportation of patients between care units. *European Journal of Operational Research*, 214(2):442–452, 2011.

[30] G. Kim, Y. S. Ong, T. Cheong, and P. S. Tan. Solving the dynamic vehicle routing problem under traffic congestion. *IEEE Transactions on Intelligent Transportation Systems*, 17:2367–2380, 8 2016.

[31] S. Kim, M. E. Lewis, and C. C. White. Optimal vehicle routing with real-time traffic information. *IEEE Transactions on Intelligent Transportation Systems*, 6:178–188, 6 2005.

[32] A. L. Kok, E. W. Hans, and J. M. Schutten. Vehicle routing under time-dependent travel times: The impact of congestion avoidance. *Computers & Operations Research*, 39:910–918, 5 2012.

[33] A. H. Land and A. G. Doig. An automatic method for solving discrete programming problems. In M. Jünger, T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey, editors, *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, pages 105–132. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[34] N. Leithäuser, D. Adelhütte, K. Braun, C. Büsing, M. Comis, T. Gersing, S. Johann, A. M. C. A. Koster, S. O. Krumke, F. Liers, E. Schmidt, J. Schneider, M. Streicher, S. Tschuppik, and S. Wrede. Decision-support systems for ambulatory care, including pandemic requirements: using mathematically optimized solutions. *BMC Medical Informatics and Decision Making*, 22(132), 2022.

[35] O. B. Madsen, H. F. Ravn, and J. M. Rygaard. A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of Operations Research 1995 60:1*, 60:193–208, 12 1995.

[36] J. T. Margolis, Y. Song, and S. J. Mason. A markov decision process model on dynamic routing for target surveillance. *Computers & Operations Research*, 141:105699, 5 2022.

[37] E. Melachrinoudis, A. B. Ilhan, and H. Min. A dial–a–ride problem for client transportation in a health–care organization. *Computers & Operations Research*, 34(3):742–759, 2007.

[38] S. Mitrović-Minić and G. Laporte. Waiting strategies for the dynamic pickup and delivery problem with time windows. *Transportation Research Part B: Methodological*, 38:635–655, 8 2004.

[39] J. Oyola, H. Arntzen, and D. L. Woodruff. The stochastic vehicle routing problem, a literature review, part ii: solution methods. *EURO Journal on Transportation and Logistics*, 6:349–388, 12 2017.

[40] J. Oyola, H. Arntzen, and D. L. Woodruff. The stochastic vehicle routing problem, a literature review, part i: models. *EURO Journal on Transportation and Logistics*, 7:193–221, 9 2018.

[41] J. Pacheco and M. Laguna. Vehicle routing for the urgent delivery of face shields during the covid–19 pandemic. *Journal of Heuristics*, 26(5):619–635, 2020.

[42] S. N. Parragh, J.-F. Cordeau, K. F. Doerner, and R. F. Hartl. Models and algorithms for the heterogeneous dial-a-ride problem with driver-related constraints. *OR Spectrum*, 34:593–633, 11 2010.

[43] V. Pillac, M. Gendreau, C. Guéret, and A. L. Medaglia. A review of dynamic vehicle routing problems. *European Journal of Operational Research*, 225(1):1–11, 2013.

[44] H. N. Psaraftis, M. Wen, and C. A. Kontovas. Dynamic vehicle routing problems: Three decades and counting. *Networks*, 67(1):3–31, 2016.

[45] U. Ritzinger, J. Puchinger, and R. F. Hartl. A survey on dynamic and stochastic vehicle routing problems. *International Journal of Production Research*, 54:215–231, 1 2015.

[46] Robert Koch Institute. Coronavirus disease 2019 (covid–19) – daily situation report of the robert koch institute, 2020.

[47] S. K. Savaşer and B. Y. Kara. Mobile healthcare services in rural areas: an application with periodic location routing problem. *OR Spectrum*, 44:875–910, 9 2022.

[48] M. Schilde, K. F. Doerner, and R. F. Hartl. Integrating stochastic time-dependent travel speed in solution methods for the dynamic dial-a-ride problem. *European Journal of Operational Research*, 238:18–30, 10 2014.

[49] N. Secomandi and F. Margot. Reoptimization approaches for the vehicle-routing problem with stochastic demands. *Operations Research*, 57:214–230, 8 2008.

[50] N. Soeffker, M. W. Ulmer, and D. C. Mattfeld. Stochastic dynamic vehicle routing in the light of prescriptive analytics: A review. *European Journal of Operational Research*, 298:801–820, 5 2022.

[51] B. W. Thomas. Waiting strategies for anticipating service requests from known customer locations. *Transportation Science*, 41:319–331, 8 2007.

[52] M. W. Ulmer, J. C. Goodson, D. C. Mattfeld, and B. W. Thomas. Dynamic vehicle routing: Literature review and modeling framework. 2017.

[53] M. W. Ulmer, J. C. Goodson, D. C. Mattfeld, and B. W. Thomas. On modeling stochastic dynamic vehicle routing problems. *EURO Journal on Transportation and Logistics*, 9:100008, 6 2020.

[54] P. L. van den Berg and J. T. van Essen. Scheduling non–urgent patient transportation while maximizing emergency coverage. *Transportation Science*, 53(2):492–509, 2019.

[55] T. Vidal, G. Laporte, and P. Matl. A concise guide to existing and emerging vehicle routing problem variants. *European Journal of Operational Research*, 286:401–416, 10 2020.

[56] D. Vigo and P. Toth, editors. *Vehicle Routing*. SIAM, Philadelphia, PA, 2014.

[57] A. Wang, A. Subramanyam, and C. E. Gounaris. Robust vehicle routing under uncertainty via branch-price-and-cut. *Optimization and Engineering*, 23:1895–1948, 12 2021.

[58] L. A. Wolsey. *Mixed Integer Programming*, pages 1–10. John Wiley & Sons, Ltd.

[59] L. A. Wolsey and G. L. Nemhauser. *Integer and Combinatorial Optimization*. John Wiley & Sons, Ltd, 1999.

[60] X. Yu, S. Gao, X. Hu, and H. Park. A markov decision process approach to vacant taxi routing with e-hailing. *Transportation Research Part B: Methodological*, 121:114–134, 3 2019.

[61] X. Yu, S. Shen, and H. Wang. Integrated vehicle routing and service scheduling under time and cancellation uncertainties with application in nonemergency medical transportation. *Service Science*, 13(3):172–191, 2021.

[62] J. Zhang, F. Liu, J. Tang, and Y. Li. The online integrated order picking and delivery considering pickers' learning effects for an o2o community supermarket. *Transportation Research Part E: Logistics and Transportation Review*, 123:180–199, 2019.

[63] J. Zhang, X. Wang, and K. Huang. On-line scheduling of order picking and delivery with multiple zones and limited vehicle capacity. *Omega*, 79:104–115, 2017.

[64] Y. Zhang, Z. Zhang, A. Lim, and M. Sim. Robust data-driven vehicle routing with time windows. *Operations Research*, 69:469–485, 3 2021.