A Capacity-Sharing Approach to Manage Jointly Transportation and Emergency Fleets at EMS Organizations

Yannick Kergosien, Valérie Bélanger, Angel Ruiz

A Capacity Sharing Approach to Manage Jointly Transportation and Emergency Fleets at EMS Organizations

Yannick Kergosien[a], Valérie Bélanger[b,d] and Angel Ruiz[c,d]

[a]Université de Tours, LIFAT EA 6300, CNRS, ROOT ERL CNRS 7002, 64 avenue Jean Portalis, 37200 Tours, France
[b]Département de gestion des opérations et de la logistique, HEC Montréal, 3000 chemin de la Côte Sainte-Catherine, Montréal (Québec), H3T 2A7, Canada
[c]Département opérations et systèmes de décision, Faculté des sciences de l'administration, Université Laval, Québec (Québec), G1K 7P4, Canada
[d]Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport (CIRRELT), Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal (Québec), H3C 3J7, Canada

ARTICLE HISTORY
Compiled November 12, 2021

ABSTRACT

Emergency Medical Services (EMS) first mission is to reach people requiring urgent medical attention and transport them to hospitals or care facilities. In many cases, EMS also provide a second mission, which concerns the non-emergency transportation of patients from one hospital to another, or between their home and medical facilities. These services have different characteristics and goals from a managerial standpoint (i.e. emergency requests are uncertain, while transport requests are known in advance and can be planned) and in practice, most EMS organizations split their fleet into two sub-fleets that are managed independently. However, both missions are in most of the cases carried out by the same types of ambulances and crews, suggesting that managing both fleets together might bring potential advantages. This study is one of the first ones, if not the first, to explore the potential advantages of a new management strategy that allows sharing resources between two separated ambulance fleets. In particular, the proposed strategy allows for dynamically modifying the size of each fleet considering that a subset of ambulances can change their mission during the day to better adapt to the system's state. This strategy offers an incomplete integration of the fleets, but has the worthy advantages of improving the overall system performance and being simple to implement by an EMS organization. Numerical experiments on realistic instances demonstrate, using a discrete event simulation tool, the feasibility and benefits of the proposed strategy.

KEYWORDS
Ambulance Management ; Medical Transportation ; Emergency Medical Services

1. Introduction

Since the pioneering work of Savas (1969), a considerable amount of research has been devoted to the development of decision models to support the management of Emergency Medical Services (EMS). Generally speaking, EMS around the world provide similar services, but important differences can be observed in the way those services

CONTACT Yannick Kergosien Email: yannick.kergosien@univ-tours.fr

are provided. Dick (2003) classified EMS practices into Anglo-American and Franco-German systems. In the Anglo-American model, on which this study is based, EMS organizations are mostly separated from the medical system, and offer solely paramedical care. Their aim is to respond to emergency calls as soon as possible and to transport the patient to the appropriate medical facility quickly and safely. In a number of cases, organizations providing EMS also provide transportation services for patients that need to be moved from one hospital to another, or between their home and medical facilities (Reuter-Oppermann, van den Berg, and Vile 2017).

To clearly distinguish non-emergency requests from emergency ones, non-emergency transport requests will be referred to as transport requests. Both services require similar resources, i.e. vehicles and crews, so one might expect some benefit from a joint planning of these common resources. Nonetheless, to the extent of our knowledge, EMS organizations often deal with their fleets in separated manners.

On the one hand, emergency requests are related most of the time to life-threatening and time-sensitive situations. They require the fast deployment of an ambulance from its location to the scene of the emergency, and often imply the patient transfer to a hospital to receive adequate care after their stabilization. Therefore, the strategic location of ambulances in the served territory is of paramount importance to keep response times as low as possible, or under given standards (Bélanger, Ruiz, and Soriano 2019). Specifically, the response time refers to the time elapsed from the reception of a request to the ambulance's arrival at the scene (Reuter-Oppermann, van den Berg, and Vile 2017). Coverage is another important measure used to evaluate EMS performance, which is related to the system's ability to respond to a percentage of requests within a predetermined time threshold (McLay and Mayorga 2010). By nature, emergency requests are highly random: decisions regarding the ambulance location is mostly based on expected demand values or probabilities.

On the other hand, transport requests are generally not as time-sensitive as emergency requests. Nonetheless, delays can cause patient discomfort, but can also impact the efficiency of health system performance, e.g. if a patient arrives late to a scheduled appointment (Beaudry et al. 2009). Consequently, transport services often focus on delays to measure their performance, which includes delay to requested appointments and the number of requests for which the vehicle arrived late. Contrarily to emergency requests, the response to transport requests can be planned: they are known in advance so the organization has some time to organize its resources (van den Berg and van Essen 2019). A transport request implies an ambulance traveling to the patient's pickup location, boarding the patient in the ambulance, and transporting them to the final destination where the team is discharged. In most cases, transports start or finish at a hospital. Therefore, vehicles assigned to transport requests are often managed by creating a sequence of movements between hospitals.

Although a number of resources or capacity sharing schemes have been proposed in the literature for a variety of industrial or service providing contexts, emergency and transport requests are generally handled separately by EMS organizations. But as a matter of fact, one may wonder if capacity sharing strategies might be useful in the context of EMS organizations, and even more, what should be the real extent of the improvement achieved by such strategies. Indeed, only few papers have considered interactions between emergency and transport fleets (Kiechle et al. 2009; van den Berg and van Essen 2019). To the best of our knowledge, this study is among the first ones, if not the first, proposing not only a realistic and potentially easy to implement strategy, but also a thorough empirical analysis of the improvement brought by two capacity sharing strategies.

More precisely, this study proposes a policy that exploits the hypothesis that each fleet answers random demands of different nature and thus uncorrelated. Therefore, it is reasonable to assume that, if at a given moment the demand for one type of service is higher than expected, the demand for the other service may be lower than expected. If it is the case, some capacity from the latter might be temporarily transferred to the former in order to help coping with the demand surge. The proposed policy takes the form of an online algorithm that, at each new event, evaluates the expected performance of each fleet with respect to their respective metrics, and decides whether or not a transfer of a vehicle, i.e. changing the vehicle's mission or in other words changing the type of requests it is assigned to, may improve the situation. If it is the case, a mathematical model is solved to identify the vehicle that will be reassigned from one fleet to the other. Notice that no transfer deteriorating one fleet expected performance will be accepted, particularly in the case of the emergency fleet. Moreover, based on the selected location model for the emergency fleet, demand zones are generally covered within standard times by more than one ambulance, so the transfer of a well chosen vehicle to perform other tasks might only have a minor impact (few seconds) on the expected response time, ensuring that in all the cases it will be kept under the standard. The proposed approach has been designed to be flexible enough so as to adapt to various EMS policies. Extensive empirical analysis using simulation highlights the potential of the proposed capacity sharing, but also its challenges when compared to the independent management generally observed in practice.

It is important to note that this study assumes that all - or a subset - of the vehicles and crews have the equipment and the training and skills to perform both emergency and transport requests as we have observed in some organizations under the Anglo-American model. However, since having over-equipped vehicles and over-trained crews may incur higher operational costs, the proposed capacity sharing approach has been designed to consider only a subset of predefined ambulances able to perform the two types of requests.

The paper is structured as follows. The next section reviews relevant contributions devoted to the management of emergency and non-emergency transport fleets, respectively, and then reports the works that have attempted to coordinate, to some extent, both fleets. Section 3 describes how emergency and non-emergency transport fleets are managed independently, along with the methods or tools used to do it. Then, Section 4 proposes a capacity sharing scheme to pool vehicles from both fleets. Results to extensive numerical experiments are reported in Section 5, which includes a discussion on the potential of the proposed capacity sharing advantages and challenges. The conclusion proposes further research avenues and closes the paper.

2. Literature Review

For more than 50 years, the management of emergency services has received a lot of attention from the Operations Research scientific community (Brotcorne, Laporte, and Semet 2003; Goldberg 2004; Reuter-Oppermann, van den Berg, and Vile 2017). Several decision problems ranging from demand forecasting to real-time fleet management have been studied and dealt with by various means, including optimization techniques and simulation methods (Ingolfsson, Erkut, and Budge 2003; Aringhieri et al. 2017). While most of the efforts have been directed towards the response to emergency and life-threatening situations, studies devoted to non-emergency patient transportation remain limited. In the following, we give an overview of diverse methods and approaches

that have been used to deal with both types of requests, particularly those that allow real-time fleet management. We will then review the few available works that seek to find ways to benefit from shared resources.

2.1. *Works Related to Emergency Transportation*

The main goal of an EMS is to respond to life-threatening situations as quickly as possible; the response time being highly influenced by both ambulance locations and dispatching rules (Bélanger, Ruiz, and Soriano 2019). Indeed, at any given time during the day, ambulances need to be located on the territory to serve. The ambulance location problem therefore aims to select the best standby sites where ambulances wait for incoming calls. The notion of coverage, which measures the number of zones (or possible demands) that can be covered within a predetermined time threshold, have been used extensively to formulate location models. Toregas et al. (1971) formulated what we know to be the first coverage model used to determine the number of ambulances to fully cover a region of interest. However, since many EMS organizations have to operate with a given fleet of ambulances, whose number is fixed, Church and ReVelle (1974) proposed instead to maximize the coverage with a given ambulance fleet. This idea further resulted in the development of many variants, which differ mostly by the way they account for the several sources of uncertainty. Multiple coverage models that seek to increase the number of vehicles covering each zone have been formulated to mitigate ambulance unavailability (Daskin and Stern 1981; Hogan and ReVelle 1986; Gendreau, Laporte, and Semet 1997). Despite their simplicity, these models have been used extensively to inform decision-making in various situations (Laporte et al. 2009), which highlight their potential application in practice. Several models have also been proposed to capture the dynamic and uncertain behavior of EMS in a more explicit manner using diverse approaches such as probabilistic modeling (Daskin 1982; Batta, Dolan, and Krishnamurty 1989), chance-constrained programming (ReVelle and Hogan 1989; Ball and Lin 1993; Beraldi and Bruni 2009), two-stage programming (Beraldi, Bruni, and Conforti 2004; Boujemaa et al. 2018; Bertsimas and Ng 2019), or robust optimization (Zhang and Jiang 2014; Bertsimas and Ng 2019).

All previous research assumed that each ambulance returns to its designed standby site after serving a request, regardless of the time of the day or the state of the system. However, the demand pattern typically changes over the day and can evolve in an unpredictable manner, which forces managers to modify ambulance locations in order to better adapt to the system's evolution (Bélanger, Ruiz, and Soriano 2019). The latter is referred to as ambulance relocation. An ambulance can be relocated at given time intervals, for instance, every two hours, or in real time, either at the end of an ambulance's mission to determine its next standby site (Maxwell et al. 2009; Schmid and Doerner 2010) or after an ambulance is dispatched to adapt to the new system's state (Gendreau, Laporte, and Semet 2001; Andersson and Värbrand 2007). Several models have been proposed to tackle variants of the relocation problem, which fits specific contexts, including multiperiod (Schmid and Doerner 2010; van Barneveld, Bhulai, and van der Mei 2017; Degel et al. 2015), offline (Gendreau, Laporte, and Semet 2006; Nair and Miller-Hooks 2009; Sudtachat, Mayorga, and McLay 2016) and online relocation models (Gendreau, Laporte, and Semet 2001; Jagtenberg, Bhulai, and van der Mei 2015; van Barneveld, van der Mei, and Bhulai 2017). It has been shown that adapting the system in real time can lead to significant improvements (Belanger et al. 2016).

Once we receive the calls, another important decision that will affect the system performance is to determine which ambulance will answer the call. Despite several recent studies that seek to propose new and improved dispatching rules, see e.g. McLay and Mayorga (2013), Bandara, Mayorga, and McLay (2014), andNasrollahzadeh, Khademi, and Mayorga (2018), the closest-idle policy, which always sends the closest ambulance to the call scene, remains the prevailing one.

2.2. *Works Related to Non-Emergency Transportation*

In many cases, EMS also provide transportation services for patients needing to go from one hospital to another, or between their home and medical facilities. Since these transport requests are known some time in advance, the main decision is therefore to plan ambulance routes in order to serve all transport requests efficiently and without delay, which includes the assignment of requests to ambulances and the sequencing of the transports to be performed. This problem is strongly related to the dial-a-ride problem (Ho et al. 2018), which consists in designing vehicle routes and schedules in a system of demand-dependent, collective people transportation (Cordeau and Laporte 2007; Molenbruch, Braekers, and Caris 2017). From a modeling perspective, dial-a-ride problems (DARP) are related to Vehicle Routing Problems (VRP) with pickups and deliveries. Cordeau and Laporte (2007) distinguished between the static and dynamic natures of transport requests' arrivals and identified two distinct modes for DARPs. The static case only serves the transport requests received ahead of a certain time limit, whereas the dynamic case considers requests throughout the day and updates the scheduling plan accordingly. In the healthcare context, specific constraints also need to be taken into account, such as special equipment, alternative loading modes and patient isolation (Beaudry et al. 2009). In addition, several objectives are simultaneously pursued, both from the organization perspective (e.g. minimizing traveling distance) and from the patient perspective (e.g. minimizing lateness).

The patient transport problem has been studied in different contexts, which leads to different variants of the DARP and diverse solution approaches. Early studies have dealt with a static version of the problem in which all transport requests are known when routes are designed. This situation is common for organizations that transport patients from their home to clinics to receive care or treatment for which the appointment time is known in advance, e.g. rehabilitation centers (Melachrinoudis, Ilhan, and Min 2007) or outpatient clinics (Parragh et al. 2012). However, most of the time, only a portion of requests are known a priori, the remaining being received in real time, so that the designed routes need to be adjusted as requests come in. Thus, the nature of DARP becomes dynamic. Therefore, fast heuristics are required to solve the problem in real time, which is again complicated by a set of context-specific constraints. To this end, several solution approaches have been built over time including insertion heuristics (Hanne, Melo, and Nickel 2009), tabu search (Beaudry et al. 2009), variable neighborhood search (Schilde, Doerner, and Hartl 2011), memetic algorithms (Zhang, Liu, and Lim 2015), and iterated local search (Lim, Zhang, and Qin 2017). In all aforementioned cases, it is assumed that multiple patients can travel together given that some constraints are respected and the required equipment is available (e.g. no more than one stretcher and one wheelchair), each vehicle having the capacity to accommodate several patients. However, in ambulance transportation, as the one we are dealing with in this paper, the capacity of the vehicle is limited to one patient. This special case has been tackled in Kergosien et al. (2011) who introduced a tabu heuristic to

schedule transport requests, allowing subcontracting to a private company if the fleet capacity is not enough to serve the demand.

2.3. *Works Related to Joint Management of Emergency and Non-Emergency Transportation Services*

Thus far, we have presented studies that focused on one type of requests and assumed a dedicated fleet of vehicles. Indeed, most of the works on EMS concentrated on the development of models and solution approaches to support the management of one fleet or the other. However, despite their different natures, the response to emergency calls and non-emergency transports can be performed by similar, if not the same resources. Therefore, one might think that a certain pool of resources can service both emergency and non-emergency transports. Kergosien et al. (2014) studied three fleet management strategies to deal with two types of requests, emergency and non-emergency. The first strategy is classic and consists in managing two fleets independently, one for each type of requests. The main idea of the second and third strategies is to treat both types of requests as if they were all emergencies, and manage a single fleet that serves them. In the second strategy, each transport request is modeled as a "dummy" emergency request that occurs at the patient earliest transportation date. The third strategy improves the second one by anticipating the best dates to perform the transport requests. The performance of each strategy is evaluated using a discrete-event simulation tool.

Recently, van den Berg and van Essen (2019) studied a system that utilizes two types of vehicles: basic life support (BLS) ambulances and advanced life support ambulances (ALS). BLS ambulances can only serve non-emergency transport requests, and ALS ambulances are generally reserved for emergency requests, but can be sent to serve non-emergency transport if the BLS fleet capacity is exhausted. The authors proposed a model that seeks to build BLS routes in such a way that the remaining coverage offered by ALS ambulances is maximized. Finally, Kiechle et al. (2009) studied a case where all the vehicles in a single fleet can serve all types of demands. When a request is received, the closest vehicle is sent to the call, and several heuristics are used to reorganize the routes planned to perform non-emergency transports. In other words, non-emergency requests are planned without taking into account emergency requests and, whenever an emergency request arrives, an ambulance is deployed and its other tasks are being redistributed among other ambulances. Simulation was used to evaluate the performance of each heuristic and their impact on coverage for potential emergency requests.

We can conclude that, despite of the diversity of methods and approaches proposed to manage emergency or transport fleets, very few works take into account collaboration between ambulances dedicated to emergency and non-emergency services. The goal of this paper is therefore to explore the potential of a collaborative scheme allowing capacity sharing between fleets. We would like to stress that this paper focuses on showing the feasibility and benefits of the proposed approach rather than comparing the performance of existing methods for handling two independent ambulance fleets. For this reason, two well-known methods from the literature have been adapted for the management of each fleet, one devoted to emergency requests, the other to transport requests. This will constitute our baseline. Then, the capacity sharing scheme will be added, allowing the transfer of vehicles between fleets. Since each fleet will still be managed according to its own methods and rules, we believe that this setting will

provide a fair assessment of the proposed capacity sharing scheme.

3. Managing Fleets Independently

We consider a single EMS organization operating two ambulances fleets, denoted EF and TF, which are dedicated to serving emergency and non-emergency transport requests, respectively. For the sake of simplicity, non-emergency transport requests are referred to as transport requests. Managing the fleets consists basically in assigning requests to ambulances (emergency or transport) and, in the case of EF, locating idle ambulances at standby sites in such a way that they will respond to emergency requests in a timely manner. Therefore, according to each fleet's vocation, the problem can be divided into two distinct subproblems, an ambulance location problem for deploying EF ambulances over the territory to serve, and a special variant of a dial-a-ride problem for assigning and scheduling transport requests to TF vehicles. The next sections propose and justify the choice of specific methods to handle and solve these specific sub-problems.

3.1. *Handling Emergency Requests: An Ambulance Location problem*

To cope with the arrival of emergency requests, a set of ambulances needs to be located at standby sites where they wait for incoming calls. The region to serve is divided into demand zones, each zone being characterized by the probability that the next incoming emergency request comes from it (this probability can be computed based on historical data and/or the population density of the zone). In addition, each emergency request is characterized by its location, the duration of the intervention at the scene, the travel time from the scene to the destination hospital, and the discharge time at the hospital. All this information is not known in advance and will be revealed in real time.

This situation can be modeled as a location problem, which aims to select ambulance standby sites in such a way that the expected demand coverage is maximized. As it was discussed in Section 2, several formulations can be proposed to model this problem. These variants distinguish mainly by the way they tackle explicitly or not the uncertain availability of ambulances at the moment a call arrives. In our case, and keeping in mind that the main goal of this research is to assess if a capacity sharing approach might be interesting in practice, we preferred to rely on deterministic models instead of using probabilistic models such as the maximum expected coverage problem (MEXCLP) Daskin (1982). Indeed, these models generally require to make hypothesis on probabilistic parameters such as the so called *busy fraction* whose setting might impact the system performance and therefore introduce some noise when analyzing the benefits of the capacity sharing scheme. Although deterministic models do not explicitly consider the fact that a vehicle might not be able to serve a call, some models such as the backup coverage model 2 (BACOP2) (Hogan and ReVelle 1986) seek to mitigate ambulance unavailability by ensuring the coverage of each demand point by two ambulances. In fact, the BACOP2 proposes a bicriteria formulation that, given a number of available vehicles, seeks to locate them to simultaneously maximize (1) the population covered at least once, and then (2) the population that is covered twice. By doing so, the model aims to avoid concentrating vehicles in zones that are easy to serve, and rather locates them in areas that seem critical. In our implementation, we replaced the population of a zone by the probability that the next call originates from

that zone. In the system under study, the BACOP2 is used to relocate ambulances in a dynamic manner to maximize the system performance in two specific situations: when an emergency request arrives or when an ambulance completes its task. On the one hand, each time a new request arrives, the closest available ambulance is dispatched to the call, so the system's coverage decreases. If it decreases in such a way that a demand zone is not covered, or that less than 50% of demand zones are covered twice, then the relocation of the remaining ambulances is launched in order to improve coverage. Ambulance relocation encompasses two steps. In the first step, the BACOP2 is executed so that the best locations for the remaining ambulances are found (note that the resulting solution does not give the information on which ambulance should be located to which standby location, but only the best locations). Then, the second step determines how to move ambulances from their current locations to the new ones. This problem is modeled as a task assignment problem where a task represents a new standby location to reach, machines correspond to ambulances, and the assignment costs - meant to be minimized - are related to the distances between the ambulance current and the new locations.

On the other hand, whenever an ambulance completes a task, two cases are to be considered. First, if at least one demand zone is not covered, the newly idle ambulance is sent to the standby site that maximizes the number of zones that are covered once. Otherwise, the ambulance is sent to the standby location that maximizes the number of zones that are covered twice.

3.2. *Scheduling Non-Emergency Requests: a Dial-a-ride problem*

The assignment and scheduling of transport requests among the available vehicles consist in solving a variant of a dial-a-ride problem (DARP) where the objective is to minimize the traveling costs and patient inconvenience. The latter is often measured as a delay on the appointment time. However, contrarily to a general DARP, we suppose that, in the case of transport requests, the ambulance can only carry one patient at a time. We also assume that each transport request is characterized by an origin and a destination, usually hospitals, although a patient's home can also be considered as an origin or a destination, and an appointment time that corresponds to the earliest time at which the patient is ready to be transported. If the transport starts after this time, a delay is incurred. Each transport request is also characterized by an estimation of the service time, which includes the time needed to travel from the vehicle position to the patient location, the time to board the patient, the transport time between the origin and the destination, and the time required to transfer the patient at the destination. Notice that the actual service time is only known when the transport is completed.

DARPs are very difficult to solve, and the proposed capacity sharing scheme needs to solve the DARP each time a new event occurs. Moreover, the dynamic nature of requests' arrival process makes that, even if a lot of effort is dedicated to finding a high quality solution, the arrival of a new request forces the system to solve a new DARP and therefore change and adapt the previously found solution. These arguments led us to the conclusion that a simple heuristic method able to produce good quality solutions in a short computational time might be the best approach to implement, and we propose a two-step constructive heuristic inspired by Jaw et al. (1986). In the first step, all known requests are sorted in ascending order of appointment time. In the second step, each of the sorted requests is assigned to an ambulance in the following manner. If the request can be performed without delay, then it is assigned to the ambulance that

minimizes the total travel distance. Otherwise, if none of the ambulances can avoid a delay responding to the request, the ambulance that can start the transport the soonest is selected. Finally, it is worth mentioning that, once an ambulance has completed a transport, it waits as much as possible before traveling to its next mission's departure location. This strategy could help avoid an empty ride if a last-minute request occurs with a departure from the hospital where the ambulance is located.

4. A Capacity Sharing Scheme for Ambulances Pooling

We consider now that the organization providing both emergency and transport services owns a fleet of identical ambulances. It is also assumed that all – or a subset – of the vehicles and crews are able to perform both kind of services. As described in Section 3, the fleet is split into two fleets EF and TF that answer both uncertain demands which are, or are assumed to be, uncorrelated. Therefore, it is fair to expect that in a number of cases, episodes of stress faced by one particular fleet may coincide with a lower than nominal workload of the other. We hypothesize that these unrelated variations in the fleet workloads might be used to improve the overall system performance if some permeability is allowed between both fleets.

Without loss of generality, we assume that both fleets have been dimensioned to achieve, under normal conditions, expected target performances α_0 and β_0, which are computed with respect to coverage and average delays on transport appointment, respectively. However, other performance metrics might be considered. Accordingly, n ambulances have been assigned to EF and are either busy responding to emergency requests or located at standby sites. The best standby sites to use can be determined by any ambulance location method. In our case, we use the BACOP2 method as described in Section 3.1. We also assume that, upon the arrival of an emergency request, the closest idle ambulance is always sent to the call. Meanwhile, m ambulances are reserved to perform transport requests. Transport requests are assigned to ambulances using a scheduling method, in our case, the method described in Section 3.2, which seeks to minimize both delays and the total travelled distance. Again, the particular method is not relevant to the study, since the goal is not to compare performances of specific scheduling methods, and other approaches might be used.

We assume that a capacity sharing scheme, which takes the form of a policy table, has been negotiated between the managers of the two fleets in such a way that, each time a new event occurs, i.e. the arrival or the completion of a request, and considering the system's state and the workload of both fleets, it is decided whether or not a transfer of a vehicle from one fleet to the other might be advantageous. Furthermore, if it is the case, the specific ambulance to transfer is selected and the necessary adjustments are made to the fleets. The next subsection presents and describes a policy table that helps decide ambulance transfers, while Section 4.2 explains how the actual transfer is implemented.

4.1. *A Policy Table to Decide Ambulance Transfers Between Fleets*

We assume that managers have agreed on a *policy table*, where lines $i \in I$ refer to possible levels of performance for EF, and columns $j \in J$ correspond to consecutive ranges of performance for TF. Table 1 illustrates such a generic policy table. Each cell (i, j) identifies a combination of performance ranges for EF and TF that can also be seen as different system's states. Therefore, each state (i, j) is limited by α^i and

α_i, and by β^j to β_j, which are the lower and upper performance bounds for EF and TF fleets, respectively. We define the first state in the matrix $(0, 0)$ as the target (or best) performance for both fleets. Finally, notice that $i + j$ sets a partial order on the potential system's states, so states $(1, 0)$ and $(0, 1)$ are considered equivalent, but preferred to any state for which $i + j > 1$. From a practical standpoint, the design of the policy table consists of setting the cardinality of I and J, and the right values for α_i and β_j, $\forall i \in I$ and $\forall j \in J$.

		TF Performance					
		$\beta \leq \beta_0$	$\beta^1 < \beta \leq \beta_1$	...	$\beta^j < \beta \leq \beta_j$	...	$\beta \leq \beta_J$
EF Performance	$\alpha \geq \alpha_0$						
	$\alpha^1 > \alpha \geq \alpha_1$						
	...				$(i-1, j)$		
	$\alpha^i > \alpha \geq \alpha_i$			$(i, j-1)$	(i, j)	$(i, j+1)$	
	...				$(i+1, j)$		
	$\alpha \geq \alpha_I$						

Table 1. A generic policy table

Without loss of generality, we consider that the system's state change upon two types of events: the arrival of a new request or the completion of an ongoing one. Each time one of those events happens, corresponding actions are taken according to each fleet management strategy, as it was discussed in Section 3, and the new system's state is estimated. For instance, if an EF ambulance completes serving an emergency request, then it is relocated according to the proposed relocation strategy. EF performance α is the expected coverage given the actual locations of the ambulances in the fleet. As per TF, a proxy of its performance β is computed as the expected average delay on the transport appointments that are planned to be served within the next two hours. Notice that other arbitrary approximations might be used to estimate TF's performance by using, for example, longer or shorter horizons, or even computing the average delay overall known transport requests not yet completed, but we believe that the proposed proxy is a fair one and right enough for the assessment of the TF. Assuming that $\alpha^i > \alpha \geq \alpha_i$ and $\beta^j < \beta \leq \beta_j$, the system's state is denoted by (i, j). If the new state is not the target state $i = j = 0$, then a potential vehicle transfer between fleets is evaluated. To this end, one must carefully assess the benefits of the transfer of a vehicle both from EF to TF and from TF to EF. Indeed, for each case, the benefit to the fleet that receives an additional vehicle may be lower or higher than the performance worsening of the fleet that releases it. The decision to transfer a vehicle or not thus depends both on the current performance of each fleet, but also on the expected improvement and deterioration resulting from the transfer. The policy table explicitly formalizes the tradeoffs that managers are ready to accept between both fleets as illustrated hereafter.

Let us assume that an event happens. The event is handled by the system according to the fleet's policy as it has been explained in Section 3, and the new system state (i, j) is estimated. Then, the effects on the fleet's performance of ambulance transfer from TF to EF and from EF to TF are evaluated. Let us first consider an ambulance transfer from TF to EF. Such a transfer is considered only if $j < J$ since it is assumed that managers do not wish to remove a vehicle from a fleet that is already at its worst performance, i.e. $j = J$. If the transfer improves EF's performance, but it is not enough as to move the fleet state to $i' < i$, then the transfer does not seem to be worthy and it is discarded. Otherwise, the transfer is considered, leading the system to a new state (i', j') such that $i' < i$ and $j' \geq j$. The transfer net contribution is computed as $\Delta_1 = (i - i') - (j' - j)$, where $(i - i')$ and $(j' - j)$ provide the "number of states"

gained by EF and lost by TF, respectively. Then, if $i < I$, an ambulance transfer from EF to TF is also studied. If we denote by (i'', j'') the new system's state, the transfer net contribution is computed as $\Delta_2 = (j - j'') - (i'' - i)$. The transfer offering the largest net and positive contribution is implemented. If both potential transfers have positive and equal net contributions, then the one bringing the number of vehicles in each fleet closer to the one at the target state is preferred and is implemented. Figure 1 illustrates the decisional processes that form the proposed capacity sharing scheme.

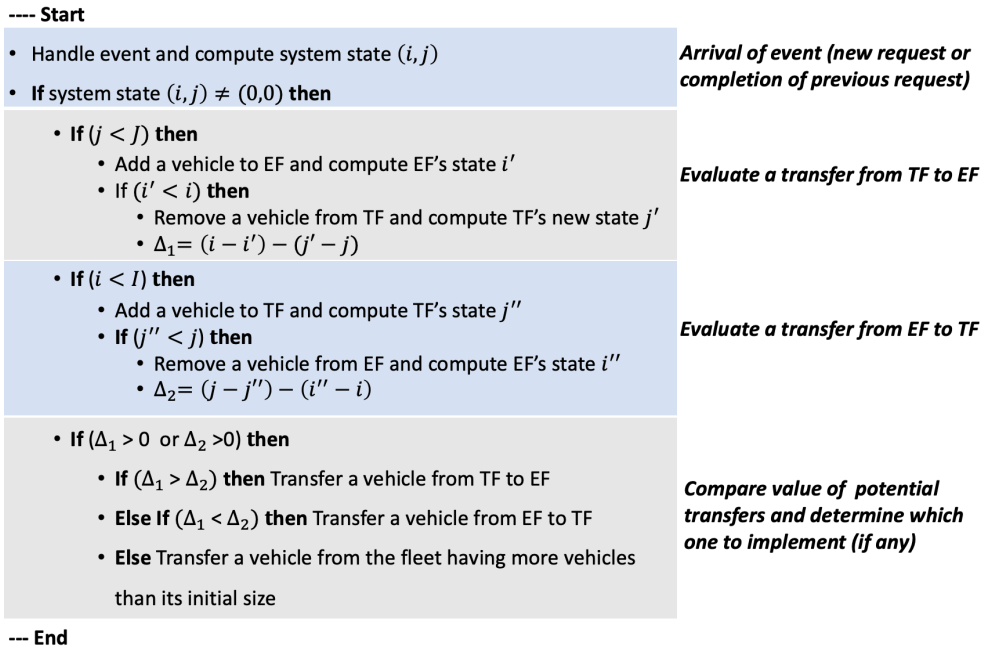

Figure 1. Evaluating the potential transfer of a vehicle

4.2. *The Transfer of Vehicles between Fleets*

Once the transfer of a vehicle has been decided, it is still necessary to select the vehicle to transfer and to make the adequate adjustments to the fleets to cope with the new situation. However, choosing the right vehicle to transfer is not an easy task. Although it is possible to allow any vehicle in the fleet to eventually change its mission, it can be more interesting from a managerial standpoint to limit the flexibility of the fleets to a relatively small number of vehicles $n^* \leq n$ and $m^* \leq m$. In this case, one must decide whether the n^* and m^* vehicles are identified a priori or not, meaning that only specific vehicles or any vehicle in the fleet may change its mission, respectively. The strategy in which the vehicles are identified seems easier to deploy and manage. Moreover, it is better suited to contexts in which practical constraints exist, including specific equipment or crew skills. The strategy in which the vehicles are not identified rather offers increased flexibility.

Let us define $\mathcal{P}$ as the set of ambulances that are allowed to change their mission. The selection of the vehicle to transfer from $\mathcal{P}$ unfolds two cases depending on the fleet that releases the vehicle.

Transfer from EF **to** TF. Let us assume that, at the current moment, EF owns n' vehicles whose locations are known. If none of the idle vehicles is in $\mathcal{P}$, then the transfer process is aborted. If only one of EF's idle vehicles belongs to $\mathcal{P}$, then this vehicle is selected for the transfer. If at least two of EF's idle vehicles belong to $\mathcal{P}$, an ambulance relocation problem is solved with $(n' - 1)$ vehicles to find a new location plan with $(n' - 1)$ standby sites using the BACOP2 model (see Section 3.1). To identify the ambulance to release, an assignment problem is formulated where n' ambulances of the current plan must be assigned to $(n' - 1)$ locations in the new plan in such a way that the total assignment cost, i.e. the total distance that ambulances drive from their current to their new location, is minimized and making sure that all the EF ambulances that not belong to $\mathcal{P}$ are assigned to a standby location. The ambulance which is not assigned to a new location is transferred to TF. Assuming that, at the current moment, TF owns m' vehicles, the requests scheduling method described in Section 3.2 is executed considering the $m' + 1$ ambulances and their current locations and missions. Notice that diversion (i.e. changing a vehicle away from its current destination to a new current destination) is allowed provided that no patient is on board.

Transfer from TF **to** EF. First, the candidates in TF to be transferred are identified. To this end, let us assume that at the current moment TF owns m' vehicles from which m'' are available (i.e. they do not have a patient on board or they are not driving to pick up a patient). If none of the m'' vehicles is in $\mathcal{P}$, then the transfer process is aborted. If only one of m'' available vehicles belongs to $\mathcal{P}$, then this vehicle is selected for the transfer. If at least two of the m'' available vehicles belongs to $\mathcal{P}$, the candidate set $\mathcal{P}'$ is built with them. Assume that, at the current moment, EF owns n' idle vehicles whose locations are known. An ambulance relocation problem is solved with $n' + 1$ vehicles to find a new location plan with $n' + 1$ standby sites. A new assignment problem is formulated where the n' ambulances of the current location plan plus the ambulances in $\mathcal{P}'$ must be assigned to $n' + 1$ locations, subjected to a constraint that limits to one the number of ambulances in $\mathcal{P}'$ that can be assigned to a standby point. As in the previous case, the assignment problem seeks to minimize the total travel distance by the ambulances in the relocation process. Finally, the requests for the TF are reassigned and rescheduled among the $m' - 1$ remaining vehicles.

5. Numerical Experiments and Results

This section reports results produced by a set of numerical experiments designed to assess the potential performance improvement as well as the downside brought by the proposed capacity sharing scheme, and discusses its potential use in practice.

It first explains how experiments were designed and how the instances were generated. Then, it reports numerical results to two sets of experiments. In the first set, it is assumed that, although demand is uncertain, it follows a stable pattern (i.e. call arrival rates are constant over the planning horizon). Contrarily, the second set of experiments assumes demand surges at specific moments in time. To be able to evaluate the performance of the proposed capacity sharing scheme in a dynamic context, a discrete event simulation (DES) model is used. We refer the interested reader to (Kergosien et al. 2015) for a detailed description of the simulation tool, from its design to its validation. The simulation model was implemented in C++. The different pro-

posed fleet management strategies were integrated in the simulation model. We used the commercial software CPLEX to solve the BACOP2 model whenever ambulance relocation is necessary. In our experiments, the time required by CPLEX each time it was called during the simulation was short (under 100 ms) which makes that the whole decision process (from evaluating the potential transfer of an ambulance, to the selection of the ambulance to transfer and the potential adjustments to the fleets) less than 2 seconds for all our experiments. Moreover, we solved several instances with the free solver GLPK and concluded that the computational times were comparable to the ones required by CPLEX, suggesting that a potential implementation of our approach wouldn?t require cutting-edge software.

5.1. *Experiments Design*

To perform the numerical experiments, an instance based on the city of Montreal (Canada) was considered. The system under study includes 440 zones, 10 hospitals, 1 depot and 38 potential sites, with the city center located in the east side. The geographic repartition of zones, hospitals, potential sites and depot are presented in Figure 2. In the case of the simulation, several replications are needed to compute average values and confidence intervals for the system performance. In our case, after some preliminary experiments to assess the variability of the observed performance indicators, 50 replications were generated to ensure reasonable confidence intervals during allexperiments. A replication s consists of different lists of requests along with their specific characteristics corresponding to a 10-hour working shift. The results reported in this section, however, correspond to the eight middle hours in order to remove the transient states corresponding to the first and last hour of the day. Each replication was generated a priori and stored in files so that the same request lists were used to fairly compare the considered management strategies.

To create realistic experiments, we gathered and merged several sources of information (annual reports of the local EMS organization, demographic statistics of the region, and information collected from the literature) to set the values of the parameters used to generate the numerical instances. Preliminary experiments allowed the fine tuning of some demand-related parameters to ensure a workload balance of the two fleets under stationary demands. Emergency requests were randomly generated in the following manner. Inter-arrival times were drawn from a Poisson distribution with an average of 5 minutes. Once a request is generated, it is associated with a specific zone following a discrete distribution where the probability of selecting a zone depends on its demographic weight. The hospital destination was randomly selected among 3 sites that can take in charge emergency requests, each with probability 1/3. Intervention time at scene varies uniformly between 5 and 10 minutes, and the time to discharge the patient at the hospital varies, also uniformly, between 10 and 15 minutes. As per the transport requests, inter-arrival times were drawn from a Poisson distribution with an average of 3 minutes, and they were assumed to be known by the system 30 to 240 minutes before the pickup time, allowing decision makers some time to schedule them and plan the routes of ambulances in advance. Origin (or destination) location was randomly determined to be a hospital with probability 2/3 (the specific hospital also being selected randomly) or a patient's home (with probability 1/3). In the latter case, the specific coordinates of the patient's home were generated uniformly among all the 440 zones. The time to take care of the patient at the hospital or at home varies uniformly between 10 and 15 minutes. In all the cases, traveling times between

the different sites (hospitals, depots, standby sites, zone centers) were generated as in (Kergosien et al. 2015).

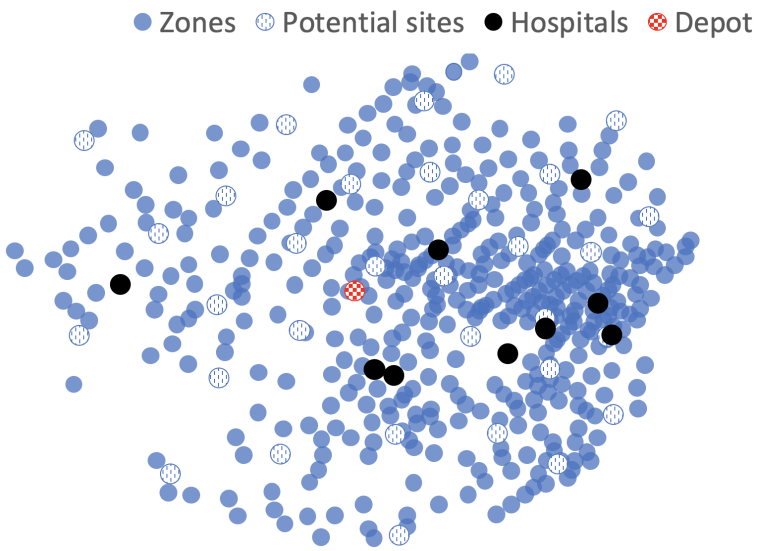

Figure 2. Geographic repartition of zones, hospitals, potential sites and depot

As it was mentioned before, the policy table is the key to the capacity sharing proposal. It is therefore necessary to build a credible and balanced table if one wishes to assess the potential of the approach in practice. This policy table can be obtained after consultation with different EMS decision makers and managers. From one organization to another, the emphasis on the quality of response given for one or the other type of requests, as well as the criteria for evaluation set by the local health authorities, may differ.

In our case, and without loss of generality, the policy table contains six performance levels for each fleet. The levels were chosen in the following manner. Firstly, we set the target performances (a single coverage of 100% with more than 80% of the regions covered twice, and average lateness under 5 min) that seemed realistic performances according to reports of real organizations. Preliminary experiments using the independent management methods in Section 3 demonstrated that fleets with $n = m = 15$ vehicles might keep these performance levels under normal conditions. Then, we proposed progressive and somehow homogeneous reductions in performance to set the following thresholds. In the case of EF, the second and third levels keep the percentage of regions covered at 100%, but reduce the percentage of regions covered twice by 2, from higher than 80% to between 40 and 80%, and from there to between 0 and 40%. Then, only single coverage was concerned. The fourth and fifth levels require coverage higher than 90% and 85%, respectively. Finally, the last level corresponds to percentages of coverage regions under 85%. As per TF, we simply increased the average delay by 5 minutes from level to level. The proposed policy table is given in Table 2.

The next subsection aims to assess the contribution of the proposed capacity sharing scheme to the fleets' performance when the demand is assumed to be stationary and the workload level of the fleet varies. Then, Subsection 5.3 tries to determine to which extent the proposed scheme may help the system cope with sudden variations in the demand faced by one fleet or the other, or even by the two fleets simultaneously.

% of coverage		Average lateness					
once	twice	< 5 min	[5 min, 10 min[	[10 min, 15 min[	[15 min, 20 min[	[20 min, . 25 min[	≥ 25 min
100%	$> 80\%$	(0,0)	(0,1)	(0,2)	(0,3)	(0,4)	(0,5)
100%	]40%,80%]	(1,0)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)
100%	[0%,40%]	(2,0)	(2,1)	(2,2)	(2,3)	(2,4)	(3,5)
[90%,100%]	-	(3,0)	(3,1)	(3,2)	(3,3)	(3,4)	(4,5)
[85%,90%]	-	(4,0)	(4,1)	(4,2)	(4,3)	(4,4)	(5,5)
$< 85\%$	-	(5,0)	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)

Table 2. Policy Table

5.2. *Stationary demand*

The goal of this first set of experiments is to assess the performance of the proposed capacity sharing approach when the demand is assumed to be stationary, and to compare it to the one produced by the independent management strategies. To this end, we executed the 50 replications presented previously under three management strategies. In the first strategy referred to as **Ind**, fleets are managed independently as described in Section 3. This strategy, which is the one generally observed in practice, constitutes the baseline of the study. The second and third strategies implement the capacity sharing scheme in different manners. The second strategy referred to as **ShP** assumes that only a subset P of ambulances are able to perform both emergency and transport requests, so they can change their mission during the day, with $|P| = (n' + m')$. In the experiments we set $n' = m' = 6$. In the last strategy, **ShAll**, all the ambulances are allowed to change their mission. At the beginning of the day, $n = m = 15$ ambulances are assigned to EF and TF.

To investigate the system's response to different levels of workload, four scenarios were considered, including $n+m = \{26, 28, 30, 32\}$. Table 3 shows the numerical results produced for each of those 4 scenarios. For each scenario, the results are computed over the same 50 replications. Results reported under **Emergency Requests** are #, the average number of answered emergency requests, **Response time**, the average response time for all the answered emergency requests (in seconds), and **% within 9 min**, the portion of requests that are served within a 9-minute threshold, which constitutes a *de facto* standard in practice. Results reported under **Transport Requests** are #, the average number of answered transport requests, **Lateness**, the average delay with respect to the patient appointment expressed in seconds, and # **Late**, the number of requests for which a delay is observed. Finally, we also computed two proxies related to the additional cost brought by ambulance transfers in terms of vehicle movements during changes of missions. To this end, columns under **Fleet Efficiency** report the percentage of empty travelled distance (column **% Empty travel**) and the number of times an ambulance changes its mission (column # **Changes**).

It is important to mention that, since results were computed over 50 replications, statistical tests were performed to ensure that the results produced by the capacity sharing strategies were significantly different from the ones in the baseline. Taking advantage of the fact that all three methods were executed using the same list of calls, we computed for each metric paired-t confidence intervals on the (**Ind** $-$ **ShP**) and (**Ind** $-$ **ShAll**) differences for each performance indicator and with $\alpha = 0.05$. Results for which the confidence interval missed "0" were reported to Table 3, which explains why the table contains some results indicated by '- -'. Therefore, all reported differences are significant. Finally, let us mention that since we use 95% confidence intervals, the overall confidence of our conclusions will be of at least 90%, since they will be based

on the combined probability that two the confidence intervals simultaneously hold the "right" value of the distribution average.

		Emergency Requests			**Transport Requests**			**Fleet efficiency**	
scenario	Strat.	#	Response Time	% within 9 min	#	Lateness	#Late	% Empty travel	#Changes
15+15	Ind	96.1	303.3	97.0	135.1	**94.6**	**8.7**	**19.5**	
	ShP		290.5	97.9		147.4	24.0	20.9	2.2
	ShAll		**288.3**	**98.3**		157.8	28.5	21.1	4.2
13+13	Ind	96.1	340.5	91.1	135.1	**177.9**	**33.3**	26.0	
	ShP		320.4	94.3		296.8	83.5	25.6	18.7
	ShAll		**311.9**	**95.5**		423.9	111.2	**24.9**	27.6
14+14	Ind	96.1	317.4	94.7	135.1	**116.5**	**14.6**	**22.1**	
	ShP		299.3	97.2		209.3	53.7	22.9	10.3
	ShAll		**297.9**	**97.7**		228.7	65.0	23.1	17.6
16+16	Ind	96.1	294.5	97.6	135.1	**85.5**	**6.8**	**18.2**	
	ShP		287.0	- -		- -	- -	18.6	1.3
	ShAll		**286.4**	**98.5**		- -	8.9	18.6	1.2

Table 3. Numerical results produced by the three fleet management strategies for different fleet sizes and assuming stationary demand

Keeping in mind the previous comments, let us look first at the results produced for scenario ($n = m = 15$) by strategy Ind. Unsurprisingly, managing the fleets independently leads to excellent performances. Indeed, average response time for emergency requests is just above 303 seconds and 97.0% of the requests were served within the 9 minutes threshold. On average, only 8.7 over more than 135 of the transport requests were served late and the average lateness was around 94 seconds. The empty travel time of ambulances for repositioning purposes, to get to the next patient location, or to come back to their standby sites after completing a mission, represents 19.5% of their total travel distance. Strategy ShP improves EF's performance with respect to Ind by reducing the response time to 290.5 seconds, an improvement of 12.8 seconds, and increasing the percentage of requests served within 9 minutes to 97.9%. However, this improvement is not achieved without a cost. Indeed, the performance of the TF is worsened slightly. With respect to the results of Ind, the average number of late requests is increased by 15.3 and the average lateness also increases by 52.8 seconds. Finally, it can be observed that, on average, only 2.2 ambulances changed their mission. Similar results are obtained for Strategy ShAll: it even further improves the results produced by ShP for EF, but also offers slightly worse results for TF.

Let us look at results obtained for smaller fleet sizes, i.e. $n = m = \{13, 14\}$, and larger ones, i.e. $n = m = \{16\}$. As one might expect, the overall performance of the fleets deteriorates as the size of the fleets is reduced, and it improves as they increase. For $n = m = \{13, 14\}$, we observe that the capacity sharing strategies always improve the performance compared to Ind for emergency requests. As per the transport requests, Ind shows a better performance. If the fleet size increases to $n = m = \{16\}$, fleets are over-dimensioned and they both have the capacity to handle requests with remarkable performance. Both capacity sharing strategies reduce the average response time for emergency requests without deteriorating the response of TF, in the case of ShP, or reducing both the number of late requests and lateness, in the case of ShAll.

Table 3 also shows that the three strategies led to very similar results with respect to the empty travel time, so the use of capacity sharing strategies does not reduce the

utilization rate of ambulances. It is worth noting that the number of mission changes increases as the fleet capacity decreases: this can be seen as an attempt to adapt to the variability of the request arrival processes. Indeed, when the fleet sizes is reduced to the smallest value, $n = m = \{13\}$, the average number of mission changes rises up to 18.7 and 27.6 for strategies ShP and ShAll, respectively.

To summarize, it is difficult to conclude on which management strategy, if any, is better. Strategies based on the capacity sharing scheme seem to favor the emergency fleet at the expense of the transfer one. This can be explained, at least partially, by the choice of the performance thresholds for TF in the policy table. In fact, since the target performance is set to an average lateness lower than 5 minutes, capacity sharing strategies do not try to improve situations where lateness remains under 300 seconds, which is the case for ShAll in all scenarios, but also for ShP, except the scenario with $13 + 13$ ambulances. Also, it is worth mentioning that the proposed policy table takes average lateness as the proxy for TF but does not account for the number of late requests. This fact might be at the origin of the higher number of late requests produced by the capacity sharing strategies. As mentioned before, the results on the percentage of empty travel times show that both capacity sharing strategies only increase empty travel marginally, and that the number of mission changes remains small.

5.3. *Response to Sudden Increases of Demand*

The capacity of both emergency and transport fleets are generally set according to the expected or nominal demand. However, in practice, fleets must face temporary demand variations that affect the system's ability to maintain target performances. The capacity sharing scheme proposed in this paper exploits the hypothesis that the emergency and the non-emergency transportation fleets answer random demands of different nature and thus uncorrelated. Therefore, it is reasonable to assume that, during a period of demand higher than nominal for one fleet, it may happen that the other one experiments a lower than nominal demand. If it is the case, some capacity from the latter might be transferred to the former in order to cope with the demand surge in such a way that the global performance might be improved. However, if the demand for both types of services increases at the same time, it is not conceivable that some capacity from one fleet could be transferred to the other. To evaluate this ability of the proposed capacity sharing strategies, we run a series of experiments in which one or several surges in demand is provoked. To do so, we reduce the requests inter-arrival time by 5%. The duration of a surge is arbitrarily set to 60 minutes. Eight different scenarios were elaborated according to the number of surges faced by each fleet (0, 1 or 2) and their combinations. It is worth mentioning that, in all cases, the timespan of surges was set in such a way that two surges do not happen simultaneously. Finally, fleet sizes were set to $n = m = \{15\}$. Numerical results for all scenarios, which consists of the average over the 50 replications, are presented in Table 4. Again, as described previously, only statistically significant results are presented, otherwise the symbol '--' is indicated.

The first four scenarios correspond to the cases where surges affect only one fleet. Let us analyze first cases $(1-0)$ and $(2-0)$, where EF faces one and two surges, respectively. In such cases, we can observe that the proposed capacity sharing strategies improve markedly both the response time and the percentage of regions covered within the 9 min standard. However, capacity sharing strategies slightly increase average lateness and the number of late transport requests. Nonetheless, as it was mentioned in the

Surges		Emergency Requests			Transport Requests			Fleet Efficiency	
EF-TF	Strat.	#	Response Time	% within 9 min	#	Average Lateness	#Late	% Empty travel	#Changes
	Ind		338.8	90.3		**84.9**	**8.6**	**20.5**	
1-0	ShP	105.9	305.5	94.9	135.4	177.4	34.8	22.1	5.2
	ShAll		**300.0**	**95.6**		181.8	41.1	22.2	7.5
	Ind		375.8	84.6		**85.1**	**8.9**	**21.4**	
2-0	ShP	115.4	323.3	92.3	135.1	190.7	41.9	23.4	6.8
	ShAll		**314.3**	**93.8**		205.8	50.6	23.4	11.3
	Ind		305.5	96.4		723.9	**66.7**	**21.0**	
0-1	ShP	96.0	- -	- -	152.3	**545.8**	- -	21.8	13.7
	ShAll		**298.8**	**97.2**		573.1	73.8	21.7	19.2
	Ind		**309.4**	**96.8**		778.9	104.7	**22.2**	
0-2	ShP	96.1	317.3	95.4	160.6	**594.0**	**91.1**	22.7	24.6
	ShAll		- -	- -		630.4	- -	22.3	33.8
	Ind		346.2	89.4		677.5	**51.1**	**22.1**	
1-1	ShP	106.1	319.7	92.8	145.1	**412.9**	60.5	23.2	16.9
	ShAll		**311.1**	**94.5**		456.2	72.4	23.0	22.6
	Ind		347.2	89.3		784.0	106.9	**23.5**	
1-2	ShP	105.8	338.3	91.4	159.8	**584.6**	**97.5**	24.1	28.4
	ShAll		**326.5**	**93.4**		625.0	115.3	24.0	38.8
	Ind		379.9	83.5		754.4	**68.1**	**23.1**	
2-1	ShP	115.6	340.5	89.7	152.6	**481.8**	82.0	24.5	19.4
	ShAll		**330.4**	**91.5**		535.6	95.6	24.2	32.2
	Ind		380.0	83.4		832.5	112.2	**23.6**	
2-2	ShP	115.6	351.4	88.1	164.9	**588.7**	**105.3**	25.1	28.3
	ShAll		**334.1**	**91.5**		639.8	122.0	24.8	42.4

Table 4. Numerical results produced by the three fleet management strategies for different number of surges in demand

previous subsection, average lateness is in all cases lower than 3.5 minutes, far below the 300 seconds threshold that was set for the best performance of TF in the policy table. In other words, it is assumed that those lateness values are not only acceptable, but excellent with respect to the target performances. If we compare both sharing strategies, ShAll leads to better results than ShP regarding EF's performance, but also produces slightly worse results for TF's.

Let us now look at cases $(0-1)$ and $(0-2)$, where TF faces on one and two surges, respectively. ShP reduces lateness achieved by Ind from 723.9 seconds (one surge) and 778.9 seconds (two surges) to 545.8 and 594.0 seconds, which represent reductions of 24.6% and 23.7%, respectively. It also produces the same [1] number of late requests than Ind for the one-surge case, but reduces the number of average late requests to 91.1 (i.e. 12.9%) for the two-surge experiments. At the same time, it leads to the same *statistical* or slightly worse performance as Ind with respect to EF. On its side, ShAll produces for the one-surge case a reduction in lateness of 20.8% with respect to the results of Ind and slight improvements in the response time (a reduction of 6.6 sec or 2.2%) and percentage of coverage, which becomes 97.2%. For the two-surge case, the reduction in lateness achieves 148.5 seconds (19.1%) and the rest of performance metrics are statistically equal to those produced by Ind. We can therefore conclude that capacity sharing strategies handle in a very interesting manner surges in TF requests with none or a very small deterioration on EF's performance.

Finally, let us move to cases where surges (one or two) affect both types of requests. Table 4 shows that the results produced by the sharing capacity strategies clearly dom-

[1] from a statistical standpoint: the confidence interval around the average of the difference between ShP and Ind contains the value 0.

inate the ones produced by the independent management. In particular, ShP reduces the response time, increases the percentage of emergency requests answered within the 9 min. threshold, and reduces average lateness with respect to independent management in all the cases. Improvements range from 2.6% up to 10.4% in response time, while the percentage of coverage increases from 2.4 up to 7.4%. As per TF performance, average lateness is reduced from 25.4 up to 36.1%. Moreover, in cases $(1-2)$ and $(2-2)$, it also improves the number of late transport requests.

ShAll produces the best average response time and the percentage of emergency requests answered within the 9 min. time limit in all cases, with reductions in response time ranging from 6.0% up to 13.0% and a significant increase in coverage. Indeed, for cases $(2-1)$ and $(2-2)$, the coverage increases from 83.5 and 83.4%, for Ind, up to 91.5% for ShAll. It also improves average lateness with respect to Ind in all cases, but reductions are slightly smaller than the ones produced by ShP.

Table 4 also shows that, when surges are more frequent, ambulances tend to travel more often empty. In all the cases, nonetheless, the increase in the percentage of empty travel produced by the capacity sharing strategies is fairly small with respect to independent management. However, the number of mission's changes increases with the number of surges, to reach 28.3 and 42.4 for ShP and ShAll strategies in the $(2-2)$ case.

To better illustrate how the capacity sharing scheme acts, Figure 3 shows, for the first instance of the $(2-2)$ set, the number of emergency and transport requests (left axis) as well as the number of ambulances included in each fleet (right axis). The figure covers an eight-hour period, which has been divided into 32 time intervals of 15 min. Three horizontal lines have been added to identify the initial number of ambulances in each fleet ($n = m = 15$), as well as the expected number of requests per 15 min to be served by each type of fleet (3 and 5 for emergency and transport requests, respectively).

Figure 3 illustrates how, at the beginning of the day, fleet sizes adjust to cope with the surge in emergency requests. To this end, up to three ambulances are transferred from TF to EF. Later, from periods 7 to 10, both emergency and transport requests are at their expected values before a first surge in transport demand occurs (periods 11 to 14). After, the number of ambulances assigned to each fleet stabilizes until period 18, when the arrival of the second surge in emergency demand provokes the transfer of three TF vehicles to EF. Fleets adjust again to cope with the second surge in transport requests in period 28, before the end of the simulation.

5.4. *Implementing capacity sharing strategies in real contexts*

The numerical results presented in the previous subsections confirm that, from a pure performance perspective, the capacity sharing strategies are very interesting, particularly when the system must handle variations in the arrival rate of requests. For these situations, the capacity sharing strategies take advantage of the flexibility granted by their ability to transfer ambulances between fleets. This allows for adapting the capacity of the fleets resulting in more robust responses. Indeed, the numerical experiments demonstrate that, in case of demand surges, both capacity sharing strategies mitigate performance worsening without impacting too much the performance of the other fleet. We therefore believe that in a practical context where variations in the arrival rate of requests might also translate into lower demand, capacity sharing strategies would perform even better due to their ability to pool resources. It is also important to recall

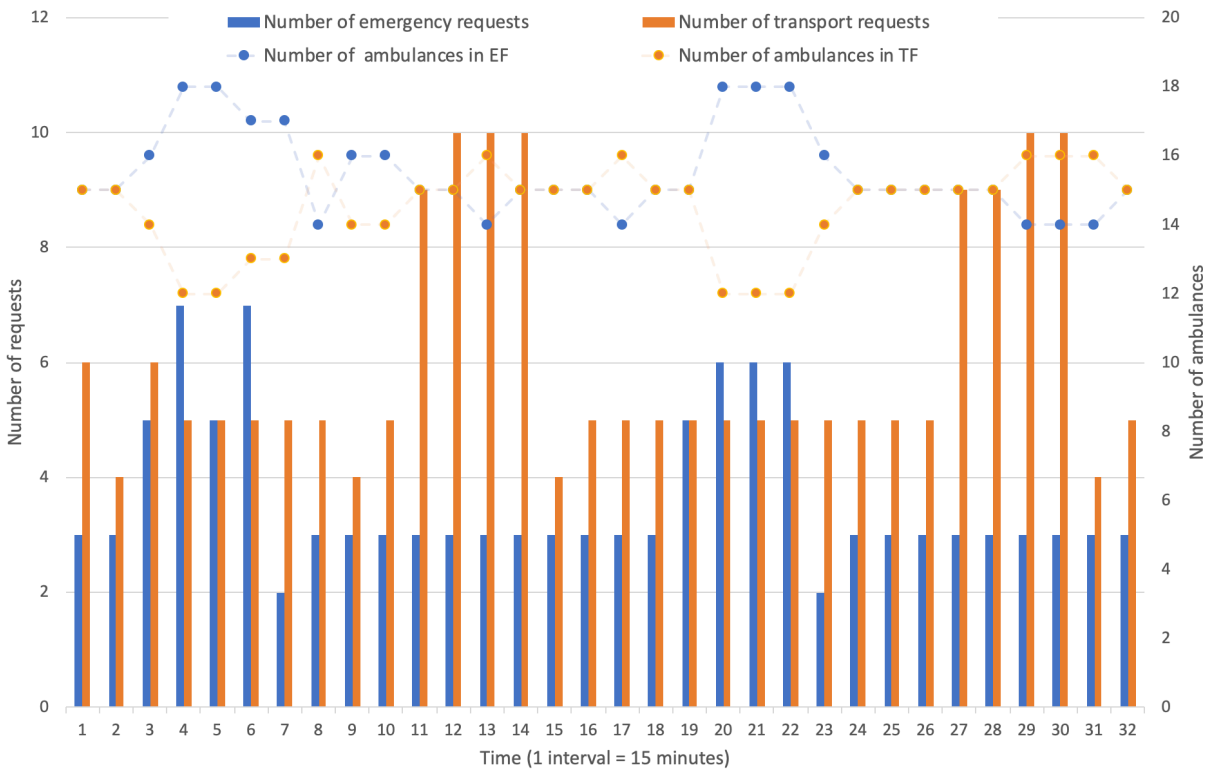

Figure 3. Variation of the number of requests and the number of ambulances in each fleet for a scenario with 4 surges in demand

that no move deteriorating the current performance of a fleet will be accepted by the capacity sharing mechanism. In particular, although one might think that in no case a transfer from the emergency fleet to the transportation fleet should be accepted, such a transfer can be done in specific occasions without deteriorating appreciably (several seconds) the emergency fleet's expected response time.

Nevertheless, we would like to insist on the fact that these numerical results were produced for a particular implementation of the capacity sharing scheme, with specific methods to manage the individual fleets and a given policy table, so one must be careful when generalizing them to other settings. Simulation tools as the one used in this research can be, in our opinion, very useful to assess the value of capacity sharing strategies in the context of a given organization. That being said, we believe that the proposed experiments provide a precise idea of the benefits that may be achieved by such strategies.

These compelling results lead us to discuss the practical application of the sharing schemes and query potential barriers. Firstly, from a computational perspective, models and algorithms supporting the decision-making process require a very short time, so we believe that they might be easily integrated into a real-time decision support system. Indeed, during the simulation, we note that for each event, all decision processes (from evaluating the potential transfer of an ambulance, to the selection of the ambulance to transfer and the potential adjustments to the fleets) were computed in less than 2 seconds. Secondly, from a practical standpoint, the implementation of the capacity sharing scheme seems feasible in modern EMS organizations. In fact, these organizations already own advanced communication systems connecting the vehicles to the dispatch center, so the crews already receive instructions in real-time. As per the potential cost of equipping vehicles and training crews to perform both kind of services, we believe it marginal in the context of the Anglo-American model. Moreover, the capacity sharing strategy offers excellent results even when only a reduced number of vehicles can perform both types of requests. In our opinion, the hardest obstacle for the implementation of capacity sharing schemes would be the reluctance of crews to perform both types of tasks and more specifically, to manage the impact on their remuneration. However, taking into account the potential improvement on the quality of the service provided to the population, we believe that organizations and unions will be able to negotiate and agree on more flexible models concerning the activity and the remuneration of crews.

6. Conclusion

More often than not, organizations providing EMS are also in charge of the non-emergency patient transports between medical facilities or between those and patients' homes. Although these *non-emergency* requests require very similar resources (vehicles and crew) or even the exact same resources as emergency requests, many organizations operate separate fleets to handle these two kinds of demands. In this paper, we explore the advantages that a strategy managing both fleets jointly may achieve. To this end, we propose a capacity sharing approach aiming at granting some permeability between fleets in an attempt to improve the overall system performance. The approach exploits the hypothesis that the fleets answer random demands of different nature and thus uncorrelated. Therefore, it is reasonable to assume that, if at given moment the demand for one type of service is higher than expected, it may happen that the demand for the other will be lower than expected. If it is the case, the transfer of one or more vehicles

may contribute to improve the overall performance. The approach takes the form of an online algorithm that, at each new event, evaluates the expected performance of each fleet in terms of their specific metrics and decides whether or not a "transfer", i.e. a change in the mission of a vehicle, may improve the situation. If it is the case, a mathematical formulation is solved to identify the vehicle that will be transferred from one fleet to the other. Two versions of the capacity sharing approach are considered. In the first version, only a restricted subset of specific ambulances is able to change their mission while, in the second version, all the ambulances can perform both tasks.

To assess the contributions of the capacity sharing approach, we have simulated the performance of a base case in which each fleet is managed independently using specific methods and then compared it to the case where the capacity sharing scheme is applied. In particular, non-emergency transport requests are scheduled by solving a dial-a-ride problem whilst the BACOP2 formulation is used to locate and relocate the ambulances of the emergency fleet in order to maximize the coverage. Two sets of experiments have been executed. In the first set, it is assumed that although both emergency and non-emergency requests are uncertain, they follow a stable pattern (i.e. the arrival rates are constant). In the second set of experiments, we assume a more realistic situation where demand surges happen at specific moments in time to represent the variations in requests arrivals observed in practice. All the experiments confirm the positive contribution of the two versions of the capacity sharing approach, particularly when the system faces surges in demand. Meanwhile, the proposed capacity sharing strategies do not seem to reduce the efficiency since the percentage of time that vehicles travel empty are not increased or increased very slightly. This capacity sharing approach offers an incomplete integration of the fleets but has the worthy advantages of improving the overall system performance and it is easily configurable to adapt to many EMS policies.

We believe that the practical implementation of the capacity sharing approach is feasible. The computational time required to execute all the decision process when a new request arrives, is less than 2 seconds. Moreover, EMS organizations already own advanced communication systems connecting the vehicles to the dispatch center and the potential cost of equipping vehicles and training crews to perform both kind of services is, in the context of the Anglo-American model, low or marginal. However, the contribution of the sharing approach depends on the particular characteristics of the demand faced by each organization, so preliminary simulation experiments should be run before launching the implementation in a given organization. In this sense, the capacity sharing approach lies on the values set in the policy table that rules the conditions for the transfer of vehicles between fleets. The policy table must be elaborated carefully since it represents tradeoffs between the different levels of performance for each fleet that are considered as comparable to managers. Last but not least, the potential sudden changes of mission cannot be envisaged without further studies on the crews' working and remuneration conditions that would need to be negotiated between unions and organizations.

Acknowledgment

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada through the Discovery Grants Program. The authors are also grateful for the remarks of the two anonymous referees, which greatly helped to improve the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Data availability statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Andersson, T., and P. Värbrand. 2007. "Decision support tools for ambulance dispatch and relocation." *Journal of the Operational Research Society* 58: 195–201.

Aringhieri, R., M. E. Bruni, S. Khodaparasti, and J. T. van Essen. 2017. "Emergency Medical services and beyond: Addressing new challenges through a wide literature review." *Computers & Operations Research* 78: 349–368.

Ball, M. O., and L. F. Lin. 1993. "A reliability model applied to emergency service vehicle location." *Operations Research* 41: 18–36.

Bandara, D., M. E. Mayorga, and L. A. McLay. 2014. "Priority dispatching strategies for EMS systems." *Journal of Operational Research Society* 65: 572–587.

Batta, R., J. M. Dolan, and N. N. Krishnamurty. 1989. "The maximal expected covering location problem : Revisited." *Transportation Science* 23: 277–287.

Beaudry, A., G. Laporte, T. Melo, and S. Nickel. 2009. "Dynamic transportation of patients in hospitals." *OR Spectrum* 32: 77–107.

Belanger, V, Y Kergosien, A Ruiz, and P Soriano. 2016. "An empirical comparison of relocation strategies in real-time ambulance fleet management." *Computers & Industrial Engineering* 94: 216–229.

Bélanger, V, A Ruiz, and P Soriano. 2019. "Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles." *European Journal of Operational Research* 272: 1–23.

Beraldi, P., and M. E. Bruni. 2009. "A probabilistic model applied to emergency service vehicle location." *European Journal of Operational Research* 196: 323–331.

Beraldi, P., M. E. Bruni, and D. Conforti. 2004. "Designing robust emergency medical service via stochastic programming." *European Journal of Operational Research* 158: 183–193.

Bertsimas, D., and Y. Ng. 2019. "Robust and stochastic formulations for ambulance deployment and dispatch." *European Journal of Operational Research* 279: 557–571.

Boujemaa, R., A. Jebali, S. Hammami, A. Ruiz, and H. Bouchriha. 2018. "A stochastic approach for designing two-tiered emergency medical service systems." *Flexible Services and Manufacturing Journal* 30 (1-2): 123 – 52.

Brotcorne, L., G. Laporte, and F. Semet. 2003. "Ambulance location and relocation models." *European Journal of Operational Research* 147: 451–463.

Church, R. L., and C. S. ReVelle. 1974. "The maximal covering location problem." *Papers of Regional Science Association* 32: 101–118.

Cordeau, J.-F., and G. Laporte. 2007. "The dial-a-ride problem : Models and algorithms." *Annals of Operations Research* 153: 29–46.

Daskin, M. S. 1982. "Application of an expected covering model to emergency medical service design." *Decision Sciences* 13: 416–439.

Daskin, M. S., and E. H. Stern. 1981. "A hierarchical objective set covering model for emergency medical service vehicle deployment." *Transportation Science* 15: 137–152.

Degel, D., L. Wiesche, S. Rachuba, and B. Werners. 2015. "Time-dependent ambulance allo-

cation considering data-driven empirically coverage." *Health Care Management Science* 18: 444–458.

Dick, W. F. 2003. "Anglo-american vs. franco-german emergency medical services system." *Prehospital and Disaster Medicine* 18: 29–37.

Gendreau, M., G. Laporte, and F. Semet. 1997. "Solving an ambulance location model by tabu search." *Location Science* 5: 75–88.

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

Gendreau, M., G. Laporte, and F. Semet. 2006. "The maximal expected relocation problem for emergency vehicles." *Journal of the Operational Research Society* 57: 22–28.

Goldberg, J. 2004. "Operations research models for the deployment of emergency services vehicle." *EMS Management Journal* 1: 20–39.

Hanne, T., T. Melo, and S. Nickel. 2009. "Bringing robustness to patient flow management through optimized patient transport in hospitals." *Interfaces* 39: 241–255.

Ho, Sin C, WY Szeto, Yong-Hong Kuo, Janny MY Leung, Matthew Petering, and Terence WH Tou. 2018. "A survey of dial-a-ride problems: Literature review and recent developments." *Transportation Research Part B: Methodological* 111: 395–421.

Hogan, K., and C. S. ReVelle. 1986. "Concepts and application of backup coverage." *Management Science* 34: 1434–1444.

Ingolfsson, A., E. Erkut, and S. Budge. 2003. "Simulation of single start station for Edmonton EMS." *Journal of the Operational Research Society* 54: 736–746.

Jagtenberg, C. J., S. Bhulai, and R. D. van der Mei. 2015. "An efficient heuristic for real-time ambulance redeployment." *Operations Research for Health Care* 4: 27–35.

Jaw, Jang-Jei, Amedeo R Odoni, Harilaos N Psaraftis, and Nigel HM Wilson. 1986. "A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows." *Transportation Research Part B: Methodological* 20 (3): 243–257.

Kergosien, Y., V. Bélanger, P. Soriano, M. Gendreau, and A. Ruiz. 2015. "A generic and flexible simulation-based analysis tool for EMS management." *International Joural of Production Research* 53 (24): 7299–7316.

Kergosien, Y., M. Gendreau, A. Ruiz, and P. Soriano. 2014. "Managing a fleet of ambulances to respond to emergency and transfer patient transportation demands." In *Health Care Systems Engineering*, edited by A. Matta, J. Li, E. Sahin, and E. Lanzarone, 303–315. Springer International Publishing.

Kergosien, Y., C. Lenté, D. Piton, and J.-C. Billaut. 2011. "A tabu search heuristic fot the dynamic transportation of patiens between care units." *European Journal of Operational Research* 214: 442–452.

Kiechle, G., K. F. Doerner, M. Gendreau, and A. Rutz. 2009. "Waiting strategies for regular and emergency transportation." In *Operations Research Proceedings*, edited by B. Fletschmann, K. H. Borgwardt, R. Klein, and A. Tuma, 271–276. Springer.

Laporte, G., F. V. Louveaux, F. Semet, and A. Thirion. 2009. "Applications of the double standard model for ambulance location." In *Innovations in Distribution Logistics*, edited by L. Bertazzi, M. G. Speranza, and J.A.E.E. van Nunen, 235–249. Berlin: Springer.

Lim, A., Z. Zhang, and H. Qin. 2017. "Pickup and Delivery Service with Manpower Planning in Hong Kong Public Hospitals." *Transportation Science* 51: 688–705.

Maxwell, M. S., M. Restepo, S. G. Henderson, and H. Topaloglu. 2009. "Approximate dynamic programming for ambulance redeployment." *INFORMS Journal on Computing* 22: 266–281.

McLay, L. A., and M. E. Mayorga. 2010. "Evaluating emergency medical service performance measures." *Health Care Management Science* 13: 124–136.

McLay, L. A., and M. E. Mayorga. 2013. "A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities." *IIE Transactions* 45: 1–24.

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

Molenbruch, Y., K. Braekers, and A. Caris. 2017. "Typology and literature review for dial-a-ride problems." *Annals of Operations Research* 259: 295–325.

Nair, R., and E. Miller-Hooks. 2009. "Evaluation of Relocation Strategies for Emergency Medical Service Vehicles." *Journal of the Transportation Research Board* 2137: 63–73.

Nasrollahzadeh, A. A., A. Khademi, and M. E. Mayorga. 2018. "Real-time ambulance dispatching and relocation." *Manufacturing & Service Operations Management* 20: 467–480.

Parragh, S., J.-F. Cordeau, K. F. Doerner, and R. F. Hartl. 2012. "Models and algorithms for the heterogeneous dial-a-ride problem with driver-related constraints." *OR Spectrum* 34: 593–633.

Reuter-Oppermann, Melanie, Pieter L van den Berg, and Julie L Vile. 2017. "Logistics for emergency medical service systems." *Health Systems* 6 (3): 187–208.

ReVelle, C. S., and K. Hogan. 1989. "The maximum availability location problem." *Transportation Science* 23: 192–200.

Savas, E. S. 1969. "Simulation and cost-effictiveness analysis of New York's emergency ambulance service." *Management Science* 15: 602–627.

Schilde, M., K. F. Doerner, and R. F. Hartl. 2011. "Metaheuristics for the dynamic stochastic dial-a-ride problem with expected return transports." *Computers & Operations Research* 38: 1719–1730.

Schmid, V., and K. F. Doerner. 2010. "Ambulance location and relocation problems with time-dependent travel times." *European Journal of Operational Research* 207: 1293–1303.

Sudtachat, K., M. E. Mayorga, and L. A. McLay. 2016. "A nested-compliance table policy for emergency medical service systems under relocation." *Omega* 58: 154–168.

Toregas, C., R. Swain, C. S. ReVelle, and L. Bergman. 1971. "The location of emergency service facilities." *Operations Research* 19: 1363–1373.

van Barneveld, T. C., S. Bhulai, and R. D. van der Mei. 2017. "A dynamic ambulance management model for rural areas." *Health Care Management Science* 20: 165–186.

van Barneveld, T. C., R. D. van der Mei, and S. Bhulai. 2017. "Compliance tables for an EMS system with two types of medical response units." *Computers & Operations Research* 80: 68–81.

van den Berg, P. L., and J. T. van Essen. 2019. "Scheduling Non-Urgent Patient Transportation While Maximizing Emergency Coverage." *Transportation Science* 52: 492–509.

Zhang, Z., and H. Jiang. 2014. "A robust counterpart approach to the bi-objective medical service design problem." *Applied Mathematics Modelling* 38: 1033–1040.

Zhang, Z., M. Liu, and A. Lim. 2015. "A memetic algorithm for the patient transportation problem." *Omega* 54: 60–71.