Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

EMS OPERATIONS MANAGEMENT: SIMULATION, OPTIMIZATION, AND NEW SERVICE MODELS

Nan Kong

Weldon School of Biomedical Engineering
Purdue University
206 S. Martin Jischke Dr.
West Lafayette, IN 47907, USA

Juan C. Paz
Xiaoquan Gao
School of Industrial Engineering
Purdue University
305 Grant Street
West Lafayette, IN 47907, USA

ABSTRACT

EMS is critical to health care industry. In this tutorial, we provide a glimpse of significant research achievements in EMS operations management. We focus on simulation modeling and their use in real-time ambulance dispatching, routing (ED selection), and redeployment/relocation decisions. We introduce optimization-based studies on ambulance management policies that have gained significant attention over the recent decade. We next describe our recent studies that optimize two emerging service models with the potential of revolutionizing EMS delivery, especially in areas with poor EMS access. Lastly, we describe prominent challenges at present, offer reflections on ongoing work, and outline future research.

1 INTRODUCTION

Emergency Medical Services (EMS) is a system that provides emergency medical care. Following an emergency call, once EMS is activated by an incidence that causes serious illness or injury, the focus of EMS is emergency care of the patient(s). As one of the most important health care services, EMS plays a vital role in saving people's lives and reducing the rates of mortality and co-morbidity. Further, EMS constitutes a significant portion of healthcare expenditure and a sizeable portion of healthcare workforce, EMS delivery is critical to operational success and sustainability of the entire healthcare system.

Access to EMS has been a serious challenge, especially in resource-deprived settings. Hence, effective management of EMS resources (mainly ambulances) is critical to excellent EMS delivery. From the perspective of operations research, the central tenet is to manage the resources such that emergency requests as many as possible can be transported to emergency departments (EDs) within a reasonable time-window. Further, a peak of emergency demand can severely overcrowd EDs, which are usually less prepared in the aforementioned settings. Analyses (often simulation based) are to help assess the impact of EMS resource management decisions on patient heath and EMS outcomes. Acknowledging their importance and sensitivity, OR researchers have studied these decision problems since the 1960s, as reported by Aringhieri et al. (2017), Reuter-Oppermann et al. (2017), and Bélanger et al. (2018). In addition, simulation is often embedded in a stochastic optimization framework for resource management optimization. Subsequently, OR researchers strive for a fine balance between simulation fidelity and optimization tractability. Increasingly, OR researchers expand the boundary of their simulation models to analyze patient health and EMS outcomes more systematically, given more detailed patient and service data, and more advanced computational power becoming available. Recently, OR researchers have actively applied simulation to predict the impact of emerging service models on augmenting patient health and EMS outcomes, which are realized by advances in transportation and information technologies. Examples of these emerging models include staffing drones in the ambulance fleet of EMS agencies and broadcasting emergency requests to nearby community-based citizen responders.

In this tutorial, we will review important models and solution approaches, not necessarily exhaustively, on EMS operations management. We will focus on real-time ambulance dispatching, routing, and redeployment/relocation decisions. In Section 2, we will provide details on a representative EMS simulation, and illustrate how to perform real-time ambulance management policy analyses and what to consider in terms of performance indices. Recently, many studies in the OR literature have embedded simulation to aid in the optimization of ambulance management policies. In Section 3, we will overview these optimization-based studies. Furthermore, technology innovation is a critical driver to the emergence of new service models that can augment EMS delivery and outcomes. In Sections 4 and 5, we will introduce two such models, following technology innovation in informatics and logistics. We will present our recent studies under these two emerging models. To conclude, we in Section 6 will provide reflections on several years of our research in this area and point out future research directions. We will focus on how simulation, and operations research in general, can really make a difference in real-world practice.

2 SIMULATION FOR AMBULANCE MANAGEMENT

To serve an emergency request, three main decisions should be addressed in a (near) online fashion, that is (1) which ambulance should be dispatched to serve an emergency request, (2) which ED facility should the patient be transported to, and (3) where to redeploy/relocate the ambulance after the service. To facilitate the management practice, proper real-time dispatching, routing and redeployment policies (DRRPs) are desired, often in such a way to maximize the number of emergency requests served within a time threshold, and to minimize the waiting times of patients. Comprehensive system analyses of DRRPs remain in great need, for which OR researchers have developed a variety of useful simulation models.

Aboueljinane et al. (2013) present a review of the simulation models in this area, many of which take a Discrete Event Simulation approach. First and foremost, they are able to replicate the dynamics of EMS operations (Figure 1). An emergency request enters an EMS agency when a citizen witness calls to report an emergency case for himself or for a third person. Dispatchers at the EMS operation center are in charge of answering the calls and assigning a color code to each request, based on the preliminary assessment on the severity of injury over the call. After the initial assessment (i.e. *initial triage*), the operator dispatches an ambulance following a given dispatching policy. Upon ambulance crew (paramedics, emergency technicians) arriving at the emergency scene, they may perform field assessment (i.e., *field triage*) jointly with the dispatcher. The crew then rescues the patient and, if necessary, transports him/her to a hospital. Usually the ambulance crew is responsible for the patient's care until he/she is handed to the hospital staff.

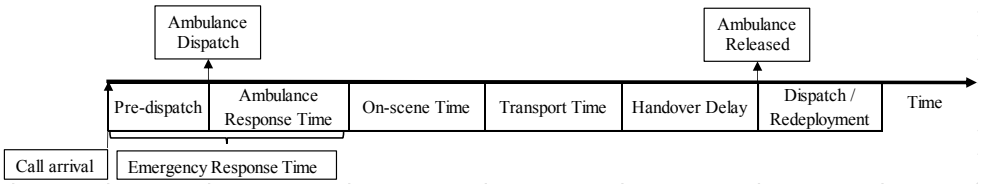

Figure 1. Representative EMS operational process.

2.1 Environment Generator

An environment in this context refers to a catchment area of an EMS agency. It is typically modeled with a planar graph $G = (N, E)$ with n nodes and m arcs. Each node is a centroid representing a small part of the whole area served by the EMS. Each arc models the connection between two nodes. There are three types of nodes, that is, the emergency demand nodes, the ambulance bases, and the ED facilities. There are further needs of manually adding or deleting nodes and arcs, and also moving nodes and, by consequence, all connected arcs. In addition, there are needs of further classifying the nodes as a proper way to generate the emergency demand. For example, for urban areas, it is also possible to characterize each node as residential, commercial, public utility space, and business offices. The graph $G = (N, E)$ is an undirected and labeled graph. The labels on the arcs (u, v) are the distance l_d among u and v, and the average speed l_s on that arc.

One can use such labels to compute travel times and/or distances among two nodes in *G*. To this end, on can use some ad-hoc version of the classic label-setting shortest-path algorithm (e.g., Dijkstra).

2.2 Simulation Input Specification

2.1.1 Spatial-Temporal Profile of Emergency Demand

As reported by many authors (e.g., Channouf et al. 2007; Setzler et al. 2009), emergency demand is not static, but, rather, fluctuates during a week and hour by hour in a day. A nodal demand table can be used to specify the relative demand fluctuation over a day with respect some normal demand at different urban areas (e.g., office nodes should have a higher relative demand during business hours of the day). Then according to the characteristics of each demand node, the model allows a negative (low) or positive (high) variation of some predefined (normal) demand arrival rate. Let w_u^i be the table entry on the demand scaler with respect to time interval i = 1, 2, 3, 4 (morning, afternoon, evening, night) and each node u.

One method to specify the generation rates of emergency demand at each node during each time interval is as follows. First, the model can take the total number of daily emergency requests just as an input, denoted by D. Then the total volume is distributed into different time intervals as $D = D_1 + D_2 + D_3 + D_4$, where these D_i's are the numbers of requests to be generated during the morning, afternoon, evening, and night time intervals, respectively. During each interval i, the D_i requests are further distributed over all the nodes belonging in N_D as follows. For each node $u \in N_D$, denote D_u^i to be the number of requests that should be generated at u during time interval i, and define it as $D_u^i = \frac{w_u^i D_i}{\sum_{v \in N_D} w_v^i}$, according to the characteristics of each nodal and its fluctuations over the day. Then the generation rate of node u is equal to D_u^i divided by the duration of time interval i. Finally, one can use the demand volume at each node to specify the total ambulance workload. Alternatively, one can specify the total ambulance workload first as a target percentage of full utilization, and in turns, specify the values of D_i's.

2.1.2 Capacity of the ED Facilities

The capacity of each facility located at node u can be derived from the total demand D plus the number of patients D' that arrive at the ED on their own. Let T_S be the average service time and n_E be the number of ED facilities, then one can compute the minimum necessary hourly capacity of the ED located at node $u \in N_E$ as $C_u = (D + D')T_s/(24n_E)$. The main assumption on the computation is to have patient requests evenly distributed among the ED facilities in such a way to have always one patient to utilize one unit of ED capacity as soon as it is released by another patient. Finally, given a scalar ranging in [1,2], the final capacity can be set in $[C_u, 2C_u]$.

2.1.3 Further Specifications

Independent of the demand generation, one can differentiate emergency requests by their urgency levels. Often the urgency is codified into categories and thus the urgency code is treated as a discrete random variable. As a result, probability distributions of ambulance rescue duration and patient release time (length of stay) at ED further differ by the urgency code. Moreover, increasing evidence in the emergency medicine literature on patient survival has allowed such further differentiation of patients and their urgency levels. Consequently, we have witnessed increasing emphasis on using patient survival to evaluate EMS performance in the OR/MS research literature. Finally, the modeling and analysis of ambulance operations management is often under the assumption that ED has sufficient capacity and thus the primary ED selection by the EMS can always be guaranteed without ambulance diversion, i.e., the initially contacted ED can always take in the emergency patient. However, in real practice, ED selection can be largely affected by

proactive ambulance diversion, which essentially prolongs the ambulance mission time and introduces additional endogenous uncertainty to the rescue process.

2.2 Modeling and Analysis of Real-Time Policies

The main aim with the developed simulation is to evaluate real-time policies (i.e., DRRPs) for the management of ambulances in terms of their impact on patient health related performance indices, such as ambulance transport time and rescue duration, and EMS system performance indices, such as occupancy of ambulances and overcrowding of ED facilities. Before introducing commonly considered DRRPs, we define an estimate of the number of ambulances needed at each base $b \in N_B$. We denote it to be A_b^e. Partition the entire node set *N* in such a way that each node *u* is assigned to the base closest to it. Let N_b be the subset of demand nodes assigned to base *b*. We use one interval for illustration, say morning for having the peak demand during a day. Thus, we remove the superscript 1 (indicating the morning interval) from w_u^1. Let W_b be the sum of the morning weights of the nodes in N_B, that is $W_b = \sum_{u \in N_b} w_u$. Then the number of ambulances needed at base $b \in N_B$ is estimated as $A_b^e = A \times W_b / (\sum_{u \in N_D \cup N_B} w_u)$.

2.2.1 Ambulance Dispatching

The most common dispatching policy in real settings is to assign an available ambulance from the closest base to the request (Cunninghame-Greene and Harries 1988). We refer to this policy, essentially static, as the **D-Closest** policy, as stated in Aringieri et al. (2018a). D stands for "Dispatching" and it applies in this subsection. This policy has been proven to perform, on average, uniformly better than the other dispatching rules by Larsen et al. (2002). Note that to ensure operational excellence by **D-Closest**, strategic location of ambulances is critical. When periodic redeployment is also considered for the ambulances, **D-Closest** can be slightly modified to improve the dispatching. See Lee (2014) for the centrality-based dispatching policy.

Alternatively, one can select to dispatch an ambulance from a list of *sufficiently close* bases, i.e., those close enough to reach the request within some time threshold. We refer to this policy as **D-LSCB**. Let L_B be such a list of sufficiently close bases. Then the **D-LSCB** policy selects a base from which to dispatch an ambulance as $\text{argmax}_{b \in L_B : A_b^a > 0} \{A_b^a - A_b^e\}$, where A_b^a is the number of ambulances available in *b* at the moment of the decision. **D-LSCB** is similar to those reported in Bandara et al. (2014), Haghani et al. (2004).

Two possible extensions of the two policies introduced above are the cutoff priority queue (**D-CPQ**) and the smart assignment (**D-SA**) policies when patient request urgency is differentiated. The **D-CPQ** involves temporarily suspending all emergency requests of lower urgency when available ambulances are fewer than a given threshold. The rationale here is that it is more desirable to ration the use of ambulances, i.e., freeing up potential ambulance resource to deal with the peaking emergency demand of higher urgency. Both **D-CPQ** and **D-SA** are introduced and discussed by Aringhieri et al. (2016), while **D-CPQ** is also analyzed by Yoon and Albert (2017). The **D-SA** involves dispatching not only ambulances at a base but even those who are under redeployment, i.e., moving from an ED to an ambulance base.

When **D-SA** is active, **D-LSCB** can be slightly modified accordingly. First, one can consider an additional list of destination bases, denoted by L_R^{AD}, of the redeploying ambulances that are capable to reach the request within the time threshold. Then the **D-LSCB** policy selects a base from which to dispatch an ambulance as $\text{argmax}_{b \in L_B \cup L_R^{AD} : A_b^a > 0} \{A_b^a - A_b^e\}$. That is, if the selected base $b \in L_B$, an ambulance is dispatched from the base; on the contrary, if $b \in L_R$, an ambulance during redeployment is dispatched. Further, if the selected ambulance $b \in L_R$ also belongs to L_B, the closest ambulance either under redeployment or available on base is dispatched. In real settings, this means to have some sort of tracking system on real-time ambulance locations. Note that in recent years, this sort of tracking system can be implemented on citizen responders in communities. This has given a vast additional dimension to real-time ambulance management and EMS delivery systems engineering in general.

2.2.2 ED Facility Selection

The **H-closest** policy selects the closest ED facility to the request. This is again a common choice in real settings. H stands for "Hospital" and it applies in this subsection. However, with this policy, ED managers in certain areas usually complain about their heavy workload and often occurred overcrowding incidence for uneven distributions of emergency demand in the catchment area of an EMS agency. They desire a more equitable distribution of the workload. Simulation-based analyses in the literature (e.g., Aringhieri et al. 2018b) has verified this for improvement on the overall efficiency of the ED facility network.

Two simple policies that are intended to address the problem of alleviating overcrowding at ED facilities are remarked in Nafarrate et al. (2010). The first policy, termed **H-SAQ**, aims to select the ED facility with the shortest admission queue counting only those that have same or higher urgency levels. The second policy, termed **H-BWL**, aims to balance the current workload of the ED facility and that needed to treat patients in the admission queue. To estimate the workload of an ED, one can use the multiplication of the average patient release time at ED and the number of patients inside the ED (both waiting for admission and under treatment). This estimate can be a weighted sum taking the urgency level into account. Note that **H-SAQ** takes the viewpoint of patients whereas **H-BWL** takes the viewpoint of the ED.

To counterbalance the effect of long travel times in cases with ED facilities being less crowded but far from most emergency requests, the **H-SAQ** and **H-BWL** polices can be modified slightly by comparing those facilities that are no farther than some radius threshold, which is often set to be half of the longest travel time between the emergency request and any ED facility.

2.2.3 Ambulance Redeployment/Relocation

A simple policy is to redeploy an ambulance to its original base once its current service is complete. We refer to this policy as **R-Base**. R stands for "Redeployment" and it applies in this subsection. Alternatively, EMS managers aim to make any ambulance ready upon its service completion. Thus, an alternative policy, termed **R-Closest**, is to redeploy the ambulance to the closest base. This alternative policy is used more often in real practice. A third policy is termed the **R-LCBT** policy, i.e., it redeploys the ambulance to the less covered base b within a given time threshold T^R as $\text{argmax}_{b \in L_R^{AR}}\{A_b^a - A_b^e\}$, where L_R^{AR} is the list of bases that can be reached from the current ED facility within T^R. Note that the parameter T^R is introduced to counterbalance the effect of longer travel times, as remarked in van Barneveld et al. (2018).

2.2.4 Quantitative Analysis

To evaluate a DRRP, we present the following list of performance indices as reference, which concerns both system-wide ambulance utilization and patient health/safety outcomes. Note that for the latter, time and/or distance-based proxies are currently used in the OR literature mostly. The indices include: Average time to reach the emergency scene (min); Average time to reach ED (min); Ambulance utilization considering only rescue mission time (%); Ambulance utilization considering also redeployment (%); ED utilization (%); Average waiting times of high- and low-level urgency at ED (min); and Fraction of patients of high- and low-level urgency reached within *x* mins. For the evaluation, one shall consider different policy combinations, as well as different scenarios, e.g., different ambulance workloads, different total ED capacities. After fixing a policy combination and a scenario, one must execute a sufficient number of runs over a time horizon with a sufficient warm-up period.

2.3 Various Models and Analyses in the Literature

Early simulation models were used to test different system configuration alternatives or to evaluate the performance of operational decision support tools. For example, Savas (1969) evaluated the impact of opening an additional ambulance station by simulation. From that point on, different decisions about EMS

design and operations have been studied using simulation. These studies include selection of station location (e.g., Berlin and Liebman 1989; Harewood et al. 2002), ambulance dispatch decision (e.g., Andersson et al. 2007; Carpentier 2006), and ambulance relocation strategies (e.g., Andersson et al. 2007; Rajagopalan et al. 2008). Later, simulation models have been embedded in decision support tools, e.g., Zhen et al. (2014) proposed a simulation–optimization framework for ambulance deployment.

More recently, generic EMS simulation models have been proposed. Henderson and Mason (2005) developed a simulation-GIS integrated model for Saint John, New Zealand. Kergosien et al. (2015) proposed a highly flexible DES model. Their conceptual framework can be replicated and extended to many different applications in EMS. Ridler et al. (2022) developed and published a generic open-source library for EMS simulation. Table 1 summarizes the characteristics of a review on simulation models that was initially presented by Kergosien et al. (2015), and extended in more recent simulation studies.

Table 1. Summary of simulation models (references not listed at the end, but attainable upon request).

	Savas (1969)	Swoveland et al. (1973)	Berlin and Liebman (1974)	Lubiez and Mielczarek (1987)	Fujiwara et al. (1987)	Trudeau et al. (1989)	Goldberg et al. (1990)	Repede and Bernardo (1994)	Gendreau et al. (2001)	Harewood et al. (2002)	Ingolfsson et al. (2003)	Henderson and Mason (2005)	Carpentier (2006)	Andersson et al. (2007a)	Rajagopalan et al. (2008)	Mason (2013)	Zhen et al. (2014)	Kergosien et al. (2015)	Aringhieri et al. (2016)	Aringhieri et al. (2018a)	Yang et al. (2019)	Yu et al. (2020)	Ridler et al. (2022)
Type of decisions or analysis																							
Location decisions	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■
Relocation strategies						■		■	■			■	■	■	■	■	■	■		■		■	■
Dispatch rules		■				■			■			■	■	■		■		■		■			■
Transfer decisions																		■					
Cruising decisions																						■	
Hospital selection																				■			
System characteristics																							
Districting		■		■			■			■													
Priority calls		■		■		■		■	■		■	■		■		■			■	■	■		■
Dispatching rules																							
Nearest ambulance	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■
Other								■					■	■		■		■		■			■
Input data																							
Demand arrival (inter-arrival times)																							
Historical data						■						■			■	■			■				■
Deterministic	■																						■
Empirical dist.																							■
Poisson process (exp. dist.)		■	■	■	■			■	■	■	■		■				■	■		■	■		■
Gaussian mixture model (spatial dist.)																					■		
Unspecified														■								■	
Travel times																							
Fixed matrix		■												■									■
Historical data																■	■						
Weibull dist.			■																				
Gamma dist.								■															
Linear regression							■												■				
Distance and speed*				■	■	■			■	■	■						■		■	■	■	■	
Complex computation												■				■							

Modified Euclidean																						
Unspecified																						
Intervention times																						
Deterministic																						
Empirical dist.																						
Exponential dist.																						
Normal dist.																						
Gamma dist.																						
Triangular dist.																						
Historical data																						
Unspecified																						

*Speed can vary by time and zone

3 REAL-TIME MANAGEMENT POLICY OPTIMIZATION

Recently, an increasing number of studies have focused on taking an optimization-based approach for real-time management policy optimization, ranging from optimal selection among static policies and further optimal parameter selection for some chosen policy, to offline dynamic optimization of state-dependent policies. In this section, we will review these two directions. Simulation has been used for either evaluating input policy configurations or (approximately) computing value functions in stochastic dynamic programming frameworks. With recent advancements on computing device and simulation techniques, the general research trend is to employ high-fidelity simulation for computing nonlinear outcome functions.

3.1 Static Policy Optimization

Work in this area includes ambulance relocation/redeployment optimization via integer programming (IP) and development of threshold policies and those hybrid ones with IP models, via simulation-optimization.

While IP-based approaches first appeared for static ambulance location decisions, such optimization engine can be called upon repeatedly to generate periodic relocation decisions efficiently. Ambulance location models are essentially extensions of the classic problems in location analysis. One prominent approach, originated by Church and ReVelle (1974), is to solve the so-called maximal covering location problem (MCLP). This model seeks to maximize the total demand attended, given a set of service facilities (corresponding to ambulances). Hogan and ReVelle (1986) proposed two extensions of the MCLP which consider a backup facility/ambulance. That is, the model seeks to assign two ambulances to each demand node. The first model, BACOP1, tries to maximize the fraction of demand that is covered by the two ambulances assigned. In the second model, BACOP2, the objective presents a balance between two fractions, one of which is the fraction covered by one ambulance assigned and the other of which is that covered by both ambulances. Yin and Mu (2012) also proposed two extensions of the MCLP, which takes into consideration the location of ambulance bases and the number of ambulances assigned to each base. Both model extensions aim to maximize the number of demand nodes covered while minimizing the average distance from the bases to the demand points not being covered. The above models heavily rely on point estimation of spatial-temporal demand, which may be challenging in the ambulance relocation/redeployment settings. In response, optimization-based approaches have been developed to better tradeoff the future demand uncertainty and the real-time nature of redeployment. Some stochastic/probabilistic optimization model is solved every time a decision needs to be made, either triggered by an event that changes the EMS system state or once every small time-interval.

One of the first models is known as the maximum expected coverage location problem (MEXCLP) in which the probability of no available ambulances is modeled explicitly by considering a busy fraction parameter derived from a queuing-theoretic model (Daskin 1983). Recently, McLay proposed the MEXCLP2 to efficiently deploy a mix of vehicles of different types to serve patients requiring different responses and/or of different urgency levels. Other important models include double standard model (Gendreau 1997) and maximum availability location problem (ReVelle and Hogan 1989). Alsalloum and Rand (2006) developed an extension of the MEXCLP that models the probability of covering the demand

required in a certain period. Sorensen and Church (2010) integrated the objective function of the MEXCLP into the maximum availability location problem to minimize the number of ambulances required to cover all demand zones. Several of these models have been extended to balance quality of service (QoS) indices, e.g., response time provided, and cost required to deliver such quality in the probabilistic sense, see, e.g., Lightner (2006) and Ingolfsson et al. (2008). To solve the above stochastic optimization models, simulation (e.g., the example model presented in Section 2) is employed to compute QoS indices, which have been pushed to measure differentiated patient survival and thus are highly nonlinear functions with respect to the ambulance location, in addition to common economic measures. The concern on embedding the simulation is not to present additional burden to the already verified efficient IP algorithms for the resultant less tractable optimization problems. More recent work includes Haghani and Yang (2007), Jagtenberg et al. (2015), Bélanger et al. (2016), van Barneveld et al. (2016), and van Barneveld et al. (2018).

Alternative, threshold-based policies are widely used for not only ambulance redeployment/relocation but also intended to help make decisions on ambulance dispatching and maybe ED facility selection as well (please referring back to Section 2 for a list of selected DRRPs). The attention is thus given to updating the threshold values whenever needed, which often relies on optimization of simulated outcomes with respect to the threshold variables. Among threshold-based policies, one prominent idea is to consider ambulance resource preparedness for probabilistically arriving requests in the near future. These indices can help more systematically assess the potential of an EMS system in response to satisfy future emergency requests based on forecasting. For example, Andersson and Värbrand (2007) proposed one decision procedure that involves a threshold value on some preparedness index. The authors employed an EMS simulation and performed simulation optimization. Enayati et al. (2018) proposed a hybrid procedure that involves solving two optimization models in sequence and considers workload restrictions for the ambulance crew. The first model maximizes coverage and the second one minimizes relocation. Redeployment is activated only when a percentage of improvement is greater than a minimum threshold, which is also to be optimized. Optimally designing these threshold-based policies requires careful threshold parameter tuning, which can be quite time-consuming and highly depend on characteristics of the focal catchment area.

3.2 Policy Optimization via Markov Decision Processes

In the recent decade, the focus has also been shifted towards offline dynamic dispatching and redeployment optimization, for which simulation has been heavily used to approximate value functions and assessing the learning efficacy in an approximate dynamic programming (ADP) or reinforcement learning (RL) based optimization framework. In an EMS system, patient demand is often highly uncertain, pre-planned scheduling or operational solutions may not optimally respond to fluctuating situations. Hence, real-time decision-making is required, which must consider system dynamics such as time-varying demands, time-varying traffic, and the different response times required by patients. With assumptions such as exponential service time and no-buffer request queue, McLay and Mayorga (2013a, 2013b), Jagtenber et al. (2017) built Markov Decision Processes (MDP) on real-time ambulance dispatching decision and solved them to optimality on small-scale instances. These exactly solved MDPs shed light on the value of the closest idle dispatching policy and how various equity formulations affect the underlying dispatching policies. Recent advances in ADP have increased our ability to solve large-scale problems efficiently. For example, Schmid (2012) and Jenkins et al. (2020) used a lookup table to approximate value functions. Maxwell et al. (2010) and Nasrollahzadeh et al. (2018) proposed novel basis functions based on the underlying problem structure to approximate value functions. Moreover, Chong et al. (2016), and Yoon and Albert (2020, 2021) constructed MDP models to optimize dispatch operations of multiple types of vehicles to patients of differentiated emergency response priorities. Note that the task of value function approximation needs to be performed repeatedly and intelligently to balance the learning efficacy for suboptimal policies and the searching efficiency of the overall optimal policy. In addition, the learning via simulation is also dependent up the existing knowledge on the influential factors/features to the value function.

4 REDEPLOYMENT DECISION PROCEDURE OPTIMIZATION WITH CRS

Various community-based programs are formed to recruit, train, and manage citizen responders. With training, citizen responders (CRs) are capable of recognizing common medical and non-medical emergencies and providing basic responses, such as hands-only cardiopulmonary resuscitation or automated external defibrillator operation for out-of-hospital cardiac arrests, naloxone spray administration for opioid overdoses, bleeding control for severe traumatic injuries, and epinephrine injection for allergic emergencies. As a result, patients in emergency situations could have a better chance to survive. Recently, increasing use of connected technology has made it possible to better engage community members into EMS practice in their communities. Example mobile applications include PulsePoint in the U.S. (Brooks et al., 2016), FirstAED in Denmark and Canada (FirstAED, 2021), and Heartrunner in Sweden (Heartrunner Sweden AB, 2021). CRs, if available to answer the response request of an emergency soon after its occurrence, are contacted by the dispatcher through a mobile application and tele-coached in real-time to provide the response. This has brought a new dimension to real-time EMS operations management.

From the above introduction, engaging citizen responders into EMS operations under the implementation of a community-based CR program, has emerged as a viable option to further improve patient survival in the emergency medicine literature (Scquizzato et al., 2020). In this section, we study a real-time ambulance redeployment procedure (Figure 2) to maximize patient survivals with consideration of CR geographic location and response availability.

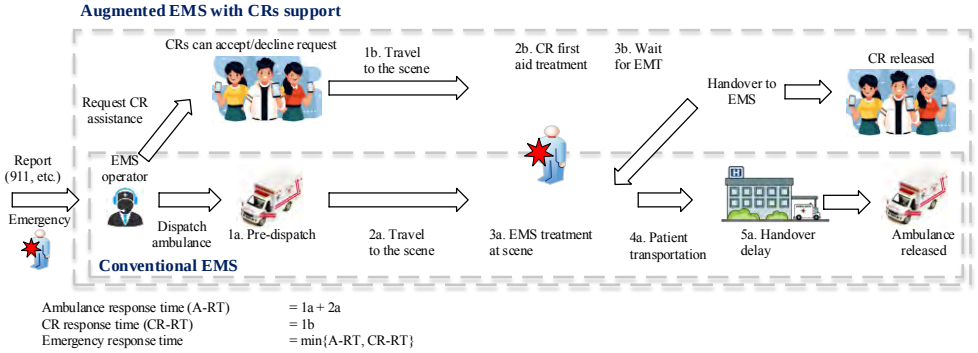

Figure 2. CR-augmented EMS vs. conventional EMS.

4.1 A Decision Procedure

The proposed real-time ambulance redeployment procedure (see Table 2) is evoked every time an ambulance becomes idle. This procedure mainly involves solving a MILP model, termed Idle Ambulances Redeployment with CRs (IAR-CR). This model assigns idle ambulances to locations when any event takes place at some discrete time t that makes some ambulance idle in the EMS system. For notation, see Table 3. Then this ambulance redeployment decision procedure is evaluated via a discrete-event simulation.

Table 2. Decision procedure of redeploying idle ambulances considering CRs.

Input: System state derived from an event with idle ambulances at discrete time t
Output: System state with all ambulances assigned to locations

set A^t of idle ambulances is identified
if $A^t = \{\}$ **then** end procedure 1
solve IAR-CR model
if the IAR-CR model has a feasible solution:
    assign optimal location of each $a \in A^t$, l_a^*, according to the IAR-CR optimal solution
else:
    assign each a to the nearest location site l
end procedure 1

In Procedure 1, a set of idle ambulances A^i is considered. The optimization model is intended to maximize the expected survival probability with CR support. Note that a maximally allowed displacement time is considered in the model, denoted by d^m, thus it is possible that no feasible solution can be found.

4.2 Survival Function

A survival function defines a patient's probability of survival with respect to the EMS response time. The OHCA survival function in De Maio *et al.*, (2003) is used to model the survival of patients with the EMS without CR support (i.e., under conventional practice), $S^A(t) = \frac{1}{1+\exp(0.679+0.262t)}$. A modified version of this function is used when patients can be attended by CRs. This new survival model includes a survival probability increase over the whole curve, which was obtained by adjusting the original scale parameter as follows. According to McNally *et al.*, (2011), which was a study related to bystander CPR, it is obtained a 4.2% survival increase at the point of 8 minutes for the U.S. average response time (Mell *et al.*, 2017), thus it is derived $S^{CR}(t) = \frac{1}{1+\exp(-0.073+0.262t)}$.

4.3 Optimization Model

The real-time availability of CRs represented by their individual probabilities of assistance to an emergency τ_c is introduced in the model by computing the probability of being assisted by CRs in each district d as $\alpha_d = 1 - \prod_{c \in C_d}(1-\tau_c)\ \forall d$, where C_d is the set of CRs that can reach district d in time, which is defined using the CRs' real-time geographical locations. At each distribution *d*, CR assistance probability α_d can be calculated this way assuming that a reasonable estimation of τ_c can be made through smartphone-enabled access to the GPS locations of CRs together with their historic response rates. Then the survival of an emergency in district d if an ambulance is dispatched from station l, denoted by s_{ld}, is calculated as the weighted average by α_d between the two survival functions defined early, and further multiplied by the time from stations to districts m_{ld}, $s_{ld} = \alpha_d \cdot S^{CR}(m_{ld}) + (1-\alpha_d) \cdot S^A(m_{ld})$. With the arrival rate of emergencies to each district λ_d, the system-wide total arrival rate is $\lambda = \sum_d \lambda_d$, and thus the probability of having an emergency at certain district *d* is $\pi_d = \lambda_d/\lambda$. Additionally, decision variables in the model can be defined as: i) R^1_{al}, 1 if idle ambulance $a \in A^i$ is assigned to location l, 0 otherwise, and ii) y^1_{ld}, 1 if district d is assigned to ambulances in l. We will present the model formulation in the tutorial session.

4.4 Simulation Model

In our simulation model, every time an event of *ambulances becoming idle* happens, procedure 1 (Table 2) or one of the benchmarks is run using the real-time system state as an input. The service area is represented in a 15*15 square grid with 225 nodes, each one with a one-mile squared area, these nodes serve as districts and also as ambulance stations (q_l: 15 ambulances). There are two hospitals evenly positioned. A fleet of 25 ambulances is randomly deployed at the beginning of the simulation, and they move in the aforementioned service area following deterministic movement times, which are calculated assuming an average speed of 30 mph (d^m: 12.5 min). Additionally, 225 CRs are randomly positioned in the districts, they move at an average speed of 5 mph and their response probability τ_c is assumed identical (h^m: 5 min), but different values are tested later. To model the spatial-temporal behavior of CRs, it is configured a "CR movement event" every 360 minutes in which they randomly change their geographic locations. Emergencies arrived at the system with a rate of nine patients per hour. They are assigned to districts following a uniform distribution and the on-site time is simulated according to an exponential distribution with a mean of 65 min. The nearest ambulance to emergencies is dispatched (including those in the middle of redeployments). If no ambulance is available, emergencies are put in queue, though the parameters used led to an uncongested system. An uncongested system was configured because it can get the most of relocations, in a congested system every time an ambulance became idle, it is immediately dispatched to the next patient, so there is little room for redeployments. It was assumed that the emergencies are critical, so all patients are required to transport to hospitals and no hospital diversion was allowed due to their criticality. Most of the parameter values are adapted from Enayati et al. (2018).

5 REAL-TIME UAV DISPATCHING AND RELOCATION TO AUGMENT EMS DELIVERY

The US is in the midst of an unprecedented opioid crisis resulting in 130 opioid overdose deaths daily (CDC 2021). Opioid overdose may result in respiratory depression and subsequent cardiac arrest without rapid intervention. Once this occurs, the chances of survival decline as much as 10% per minute (Stoesser et al. 2021). The acute treatment of overdose is the administration of naloxone. Despite the fact that naloxone has saved tens of thousands of lives, overdoses typically happen when no one else is present to administer it in time. Limited access to doctors and programs, on the other hand, might be a barrier to treatment (Volkow and Collins 2017). Narcan nasal spray, an intranasal naloxone formulation, is proven efficacious and easy to use (Volkow and Blanco 2021). Its availability now allows laypersons to administer naloxone on their own to revert overdoses. Nevertheless, reaching those who really need it on short notice remains a challenge, as there is considerable variability in the availability of naloxone by locality. In general, naloxone cannot be made publicly available at high-traffic or high-risk locations (Bennet and Elliot 2021). Hence, it is often impossible for trained first-responders to reach the victim in need to administer naloxone. Bystander-enabled Narcan UAVs are an emerging option to mitigate the challenge on access to naloxone (Figure 3). In such an augmented EMS system, 9-1-1 dispatchers trained as drone pilots could fly the drone and tele-coach the bystander to administer the Narcan to the patient while EMS personnel are on their way.

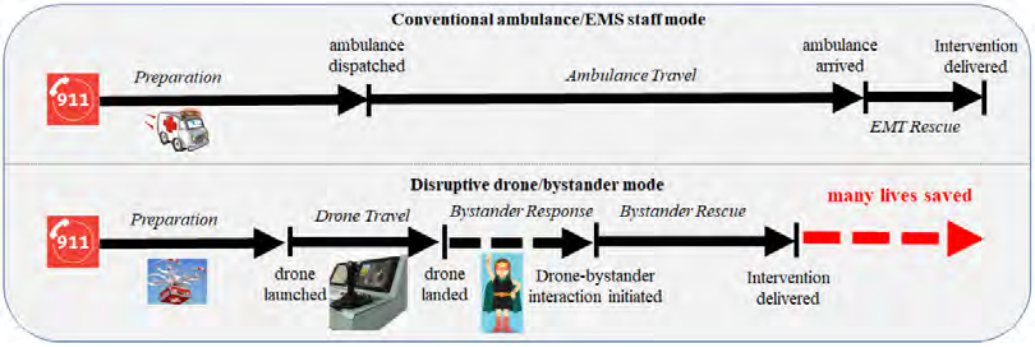

Figure 3. UAV-augmented EMS vs. conventional EMS.

From the above introduction, staffing drones to emergency care resource together with ambulances, has become a viable option to improve care access and patient survival in the emergency medicine literature (Cheskes et al., 2020; Rosamond et al. 2020). In this section, we study the real-time UAV dispatching and relocation policy optimization problem and derive a state-dependent policy to manage UAVs upon the completion of some rescue mission either by an ambulance or a UAV-bystander combo.

5.1 MDP Model

This section presents an infinite-horizon Markov decision process (MDP) formulation. At each decision epoch, it is possible that all the dispatching, relocation, and recharge actions are feasible in some states. This time-driven MDP modeling approach expands the decision space and increases the EMS system performance compared to the event-driven modeling used in the ambulance dispatching and relocation operations research (Maxwell et al. 2010; Nasrollahzadeh et al. 2018). By using this approach, we improve the decision quality by pooling requests and allowing dispatching, relocation and recharge decisions at discrete time points. In our implementation, we set a short decision epoch, which enables us to check the system state and make decisions frequently. This increases the potential to further improve the average response time of the EMS system. On the other hand, to alleviate the increased computational burden, we adopt an approximation dynamic programming solution approach, as shown in Section 5.2.

Let $\mathcal{N} := 1,2,\ldots,N$ be the set of demand nodes and $\mathcal{M} := \{1,2,\ldots,M\}$ be the set of UAV-charging stations. Then $\mathcal{N} \cup \mathcal{M}$ is the set of candidate relocation destinations of UAVs. We consider a total of L

UAVs and assume that the arrival of requests follows a Poisson distribution with mean λ. By Poisson splitting, arrival of requests at demand node i is modeled as $Poiss(\lambda_i)$ with $\lambda_i = P_i\lambda$, where $\sum_{i\in\mathcal{N}} P_i = 1$.

State Space. There are L UAVs in the EMS system. To simplify the presentation, we assume that we do not keep more than Kwaiting requests for the UAVs. This is not a restriction in the practical sense because Kcan be quite large. The state space is composed of three parts: vectors $B = (b_1, b_2, \dots, b_L)$, $C = (c_1, c_2, \dots, c_K)$ and $\boldsymbol{\gamma}$, where $b_l, l = 1, \dots, L$ contains information about the state of UAV $l, c_j, j = 1, \dots, J$, contains information about waiting request j, and $\boldsymbol{\gamma}$ denotes the busyness of the ambulance system. Naturally, the state of the UAVs and the requests evolve over time, but we omit this dependence for brevity.

The state of UAV l is given by $b_l = (d_l, r_l, e_l), l = 1, \dots, L$, where $d_l \in 1,2, \dots, N$ is the destination of each UAV, $r_l \in \{0,1,2,\dots, R\}$ is the remaining service time of each UAV, and $e_l \in 0,1,2, \dots, Z$ is the energy level of each UAV. The destination is determined when a dispatching/relocation/recharge decision is made. So is the remaining service time, which depends on the distance between the current UAV location and its destination. The remaining service time gradually decreases to zero as time evolves. Note that here we only consider dispatching, relocate and recharge decisions for idle UAVs, i.e., when the remaining service time is 0. Hence, we only need to know the destination and remaining service time of each UAV rather than the location, destination and start time of the movement as in previous work (Maxwell et al. 2010; Nasrollahzadeh et al. 2018). An unserved request j is represented by $c_j = (g_j, h_j), j = 1, \dots, K$, where $g_j \in 1,2, \dots, N$, is the request location, and $h_j \in 0,1,2, \dots, H$ is the time that the request being kept for each UAV. Once a request is admitted into the UAV request queue, it can only be served by a UAV. If a dispatching decision is made, the request would be marked as ``served" and immediately removed from the UAV request queue. In other words, we only keep track of unassigned requests.

Finally, we incorporate $\gamma \in [0,1]$ in the state space, $\gamma = 0$ indicates that all the ambulances are available and $\gamma = 1$ indicates that all the ambulances are busy. In our model, γ is viewed as exogenous information and hence we do not explicitly consider the ambulance staffing and its dispatching process in our model. Hence, we can use the tuple $s = (B, C, \gamma)$ to represent the state of the EMS system.

Action Space. We now describe the actions that we can take in a state. At each decision epoch t, for each unserved request in the UAV system, the administrator (i.e., decision-maker) first decides whether to admit the request into the UAV system or refer it to the ambulance system, i.e., define $R_j \in 0,1$, $j = 1,2, \dots, J$, where $R_j = 1$ if request j is referred to the ambulance system and then removed from the queue, $R_j = 0$ otherwise (i.e., admitted into the UAV system). Based on the status of the request queue and UAVs, the administrator makes dispatching, relocation and recharge decisions for each UAV in the system. Define $X_{l,j} = 1$ if UAV l is dispatched to request j, $X_{l,j} = 0$ otherwise, and $Y_{l,f} = 1$ if UAV l is redeployed to station f, $Y_{l,f} = 0$ otherwise. Finally, define $Z_{l,i} = 1$ if UAV l is relocated to demand node i, $Z_{l,i} = 0$ otherwise. We will provide the mathematical representation of the action space in the tutorial session.

Transitions. As stated early, we assume that request arrivals at location i follow a homogeneous Poisson process with rate λP_i. We further assume that travel times are deterministic and only linearly depend on the travel distance. Given the simplicity of the emergency rescue task, we assume fixed service times at the scene. Three types of events drive the state transitions: (1) new requests arrive; (2) the decision-maker makes an admission, dispatching, relocation or recharge decision; and (3) UAVs' remaining service time and their battery power naturally decrease as time evolves.

Cost Function. We consider minimizing the sum of the expected discounted cost associated with the response time of each request served by UAV and the number of requests referred to the ambulance.

5.2 Approximate Dynamic Programming

Standard discounted infinite-horizon MDP problems can be solved by value iteration, policy iteration, and linear programming (Puterman 2014). To solve a problem with any of the above algorithms, one needs to first parameterize the associated transition matrix ($\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to R$) and reward matrix ($\mathcal{S} \times \mathcal{A} \to R$). In our MDP model, the transition matrix has a dimension of $|\mathcal{S}|^2 \times |\mathcal{A}| = \left(\left((N+M+1)(R+1)(Z+1)\right)^L \cdot (N(H+1)^K)\right)^2 \times (2^K(N+M+1)^L)$. For an instance with 64 demand nodes and 2 UAVs, the dimension of the state space is about 2.96×10^{27}, which means that solving the Bellman equation to optimality is

impractical. We therefore employ an approximate policy iteration algorithmic approach to construct high-quality policies based on value functions. We will present an approximation policy iteration algorithm, and the carefully selected basis functions for the construction of the neural network in the tutorial session.

6 REFLECTION AND CONCLUDING REMARKS

In this tutorial, we review the OR/MS literature on EMS operations management with focus on real-time ambulance resource management. By far, this is still a very exciting and fertile research area, especially for the sense that it provides opportunities to develop computational techniques to analyze complex service system dynamics and solve large-scale stochastic optimization problems.

Reflecting on our work, we offer the following observations and outline future research directions accordingly. One, there is a need on applying reinforcement learning to further improve our ability of solving large-scale instances under the changing environment. With advancement of information technology, EMS agencies will desire continuous information updates. As new information arrives, our understanding on the service system changes continuously, which motivates us to identify more balanced acts between exploration (of uncharted territory of state-action space) and exploitation (of current knowledge being used to identify better polices). Two, it is time to start benchmarking our data sources and simulation models in this area. This is crucial to evaluate the performance (viability, reliability, and superiority) of new management strategies and solution techniques. We have witnessed recent developments on generic flexible simulation-based analysis tool for a variety of distinct and realistic contexts. Emphasis should be given to modules for incorporating GIS information, field-triage human factors, and a holistic view of the care spectrum spanning from EMS to ED and from ED to ICU, and over the entire hospital network with differentiated specialty and service characteristics. Further, the notion of digital-twins should be incorporated to investigate the partnership of dispatchers and AI tools in emergency response management. Three, as advanced information and autonomy technologies revolutionizing healthcare delivery. It is not surprising to OR/MS researchers on the benefit of teaming up with technology researchers and developers. We are fortunate to work with drone researchers and mobile app developers.

REFERENCES

Aboueljinane, L., E. Sahin, and Z. Jemai. 2013. "A Review on Simulation Models Applied to Emergency Medical Service Operations". *Computers & Industrial Engineering* 66(4):734–750.

Alsalloum, O. I. and G. K., Rand. 2006. "Extensions to Emergency Vehicle Location Models". *Computers and Operations Research* 33(9):2725–2743.

Andersson, T., S. Petersson, and P. Värbrand. 2007a. "Decision Support for Efficient Ambulance Logistics." *ITN Research Report LiTHITN-R-2005-1*. Linkopings Universiteit.

Andersson, T. and P. Värbrand. 2007. "Decision Support Tools for Ambulance Dispatch and Relocation". *Journal of the Operational Research Society* 58(2):195–201.

Aringhieri, R., G. Carello, and D. Morale. 2016. "Supporting Decision Making to Improve the Performance of an Italian Emergency Medical Service". *Annals of Operations Research* 236(1):131–148.

Aringhieri, R., M. Bruni, S. Khodaparasti, and J. van Essen. 2017. "Emergency Medical Services and Beyond: Addressing New Challenges through a Wide Literature Review". *Computers and Operations Research* 78:349-368.

Aringhieri, R., S. Bocca, L. Casciaro, and D. Duma. 2018a. "A Simulation and Online Optimization Approach for the Real-Time Management of Ambulances". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 2554 – 2565. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

Aringhieri, R., D. Dell'Anna, D. Duma, and M. Sonnessa. 2018b. "Evaluating The Dispatching Policies for a Regional Network of Emergency Departments Exploiting Health Care Big Data". In International Conference on Machine Learning, Optimization, and Big Data, edited by G. Nicosia, P. Pardalos, G. Giuffrida, and R. Umeton, Volume 10710 of Lecture Notes in Computer Science, 549–561. Cham, Switzerland: Springer International Publishing.

Bandara, D., M. Mayorga, and L. McLay. 2014. "Priority Dispatching Strategies for EMS Systems". *Journal of the Operational Research Society* 65(4):572–587.

Brooks, S.C., G. Simmons, H. Worthington, B.J. Bobrow and L.J. Morrison. 2016. "The PulsePoint Respond Mobile Device Application to Crowdsource Basic Life Support for Patients With Out-Of-Hospital Cardiac Arrest: Challenges for Optimal Implementation". *Resuscitation* 98: 20–26.

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 Operations Research* 272(1):1-23.

Bennett, A. S. and L. Elliott. 2021. "Naloxone's Role in the National Opioid Crisis — Past Struggles, Current Efforts, and Future Opportunities". *Translational Research* 234:43–57.

Berlin, G. N., and J. C. Liebman. 1974. "Mathematical Analysis of Emergency Ambulance Location". *Socio-Economic Planning Sciences* 8: 323–328.

Carpentier, G. 2006. *La Conception et la Gestion d'un Réseau de Service Ambulancier* [Ambulance Service Network Design and Planning]. Mémoire de maitrise: Université Laval.

CDC 2020. *Mortality in the United States, 2020*. National Center for Health Statistics Data Brief No. 427. December 2021. Available from https://www.cdc.gov/nchs/products/databriefs/db427.htm.

Channouf, N., P. L'Ecuyer, A. Ingolfsson, and A. Avramidis. 2007. "The Application of Forecasting Techniques to Modeling Emergency Medical System Calls in Calgary, Alberta". *Health Care Management Science* 10(1):25-45.

Cheskes, S., S. L. McLeod, M. Nolan, P. Snobelen, C. Vaillancourt, S. C. Brooks, K. N. Dainty, T. C. Y. Chan, and I. R. Drennan. 2020. "Improving Access to Automated External Defibrillators in Rural and Remote Settings: A Drone Delivery Feasibility Study". *Journal of American Heart Association* 9(14):e016687.

Chong, K. C., S. G. Henderson, and M. E. Lewis. 2016. "The Vehicle Mix Decision in Emergency Medical Service Systems". *Manufacturing & Service Operations Management* 18(3):347–360.

Church, R. and C. ReVelle. 1974. "The Maximal Covering Location Problem". In *Papers of the Regional Science Association*, 101–118.

Cuninghame-Greene, R., and G. Harries. 1988. "Nearest-neighbour Rules for Emergency Services". *Zeitschrift fur Operations Research* 32(5):299–306.

Daskin, M. S. 1983. "A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution". *Transportation Science* 17(1):48-70.

De Maio, V. J., I. G. Stiell, G. A. Wells and D. W. Spaite. 2003. "Optimal Defibrillation Response Intervals for Maximum Out-of-Hospital Cardiac Arrest Survival Rates". *Annals of Emergency Medicine* 42(2): 242–250.

Enayati, S., O. Y. Özaltın, M. E. Mayorga and C. Saydam. 2018. "Ambulance Redeployment and Dispatching Under Uncertainty with Personnel Workload Limitations". *IISE Transactions* 50(9): 777–788.

FirstAED. 2021. FirstAED. https://firstaed.com/, accessed 12th October, 2021.

Gendreau, M. 1997. "Solving an Ambulance Location Model by Tabu Search". *Location Science* 5(2):75–88.

Haghani, A., Q. Tian, and H. Hu. 2004. "Simulation Model for Real-Time Emergency Vehicle Dispatching and Routing". *Transportation Research Record* 1882(1):176–183.

Haghani, A. and S. Yang. 2007. "Real-Time Emergency Response Fleet Deployment: Concepts, Systems, Simulation and Case Studies". In *Dynamic Fleet Management*, 133-162.

Harewood, S. I., S. Budge, and E. Erkut. 2002. "Emergency Ambulance Deployment in Barbados: A Multi-Objective Approach". *Journal of the Operational Research Society* 53:185–192.

Heartrunner Sweden AB. 2021. HeartRunner. https://heartrunner.com/, accessed 10th December, 2021.

Henderson, S. and A. Mason. 2005. "Ambulance Service Planning: Simulation and Data Visualisation". *Operations Research and Health Care* 70:77–102.

Hogan, K. and C. ReVelle. 1986. "Concepts and Applications of Backup Coverage". *Management Science* 32(11):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.

Ingolfsson, A., S. Budge, and E. Erkut. 2008. "Optimal Ambulance Location with Random Delays and Travel Times". *Health Care Management Science* 11(3):262–274.

Jagtenberg, C. J., S. Bhulai, and R. van der Mei. 2015. "An Efficient Heuristic for Real-Time Ambulance Redeployment". *Operations Research for Health Care* 4:27–35.

Jagtenberg, C. J., S. Bhulai, and R. D. van der Mei. 2017. "Dynamic Ambulance Dispatching: Is the Closest-Idle Policy Always Optimal?". *Health Care Management Science* 20(4):517–531.

Jenkins, P. R., M. J. Robbins, and B. J. Lunday. 2020. "Approximate Dynamic Programming for Military Medical Evacuation Dispatching Policies". *INFORMS Journal on Computing* 33(1):2–26.

Kergosien, Y., V. Bélanger, P. Soriano, M. Gendreau, and A. Ruiz. 2015. "A Generic and Flexible Simulation-Based Analysis Tool for EMS Management". *International Journal of Production Research 53*(24):7299-7316.

Lee, S. 2014. "Role of Parallelism in Ambulance Dispatching". *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44(8):1113–1122.

Lightner, C., A. Tavakoli, and Y. Fathi. 2006. "Developing a Mathematical Model for Locating Facilities and Vehicles to Minimize Response Time". *Journal of Applied Business Research* 22(2):17-24.

Maxwell, M. S., M. Restrepo, S. G. Henderson, and H. Topaloglu. 2010. "Approximate Dynamic Programming for Ambulance Redeployment". *INFORMS Journal on Computing* 22(2):266–281.

Nafarrate, A., J. Fowler, and T. Wu. 2010. "Bi-criteria Analysis of Ambulance Diversion Policies". In Proceedings of the 2010 Winter Simulation Conference, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Ycesan, 2315–2326. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

McLay, L. A. 2009. "A Maximum Expected Covering Location Model with Two Types of Servers," *IIE Transaction* 41(8): 730–741.

McLay, L. A. and M. E. Mayorga. 2013a. "A Dispatching Model for Server-to-Customer Systems that Balances Efficiency and Equity". *Manufacturing & Service Operations Management* 15(2):205–220.

McLay, L. A. and M. E. Mayorga. 2013b. "A Model for Optimally Dispatching Ambulances to Emergency Calls with Classification Errors in Patient Priorities". *IIE Transactions* 45(1):1–24.

Nasrollahzadeh, A. A., A. Khademi, and M. E. Mayorga. 2018. "Real-time Ambulance Dispatching and Relocation". *Manufacturing & Service Operations Management* 20(3):467–480.

Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York.

Rajagopalan, H. K., C. Saydam, and J. Xiao. 2008. "A Multiperiod Set Covering Location Model for Dynamic Redeployment of Ambulances". *Computers & Operations Research* 35:814–826.

Reuter-Oppermann, M., P. L. van den Berg, and J. L. Vile. 2017. "Logistic for Emergency Medical Services Systems." *Health Systems* 6(3):187-208.

ReVelle, C. and K. Hogan. 1989. "The Maximum Availability Location Problem". *Transportation Science* 23(3):192–200.

Ridler, S., A. J. Mason, and A. Raith. 2022. "A Simulation and Optimisation Package for Emergency Medical Services". *European Journal of Operational Research 298*(3):1101-1113.

Rosamond, W. D., A. M. Johnson, B. M. Bogle, E. Arnold, C. J. Cunningham, M. Picinich, B. M. Williams, and J. K. Zegre-Hemsey. 2020. "Drone Delivery of an Automated External Defibrillator". NEJM 383(12):1186-1188.

Savas, E. S. 1969. "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service". *Management Science* 15:602–627.

Schmid, V. 2012. "Solving the Dynamic Ambulance Relocation and Dispatching Problem Using Approximate Dynamic Programming". *European journal of operational research* 219(3):611–621.

Scquizzato, T., O. Pallanch, A. Belletti, A. Frontera, L. Cabrini, A. Zangrillo, and G. Landoni. 2020. "Enhancing Citizens Response to Out-of-Hospital Cardiac Arrest: A Systematic Review of Mobile-Phone Systems to Alert Citizens as First Responders". *Resuscitation* 152(February):16–25.

Setzler, H., C. Saydam, and S. Park. 2009. "EMS Call Volume Predictions: A Comparative Study". *Computers & Operations Research* 36(6):1843-1851.

Sorensen, P. and R. Church. 2010. "Integrating Expected Coverage and Local Reliability for Emergency Medical Services Location Problems". *Socio-Economic Planning Sciences* 44(1):8–18.

van Barneveld, T. C., S. Bhulai, and R. van der Mei. 2016. "The Effect of Ambulance Relocations on the Performance of Ambulance Service Providers". *European Journal of Operational Research* 252(1):257–269.

van Barneveld, T. C, C. Jagtenberg, S. Bhulai, and R., van der Mei. 2018. "Real-time Ambulance Relocation: Assessing Real-time Redeployment Strategies for Ambulance Relocation". *Socio-Economic Planning Sciences* 62:129–142.

Volkow, N. D. and C. Blanco. 2021. "The Changing Opioid Crisis: Development, Challenges and Opportunities". *Molecular Psychiatry* 26(1):218–233.

Volkow, N. D. and F. S. Collins. 2017. "The Role of Science in Addressing the Opioid Crisis". *New England Journal of Medicine* 377(4):391–394.

Yin, P. and L. Mu. 2012. "Modular Capacitated Maximal Covering Location Problem for the Optimal Siting of Emergency Vehicles". *Applied Geography* 34:247–254.

Yoon, S. and L. A. Albert. 2018. "An Expected Coverage Model with a Cutoff Priority Queue". *Health Care Management Science* 21:517-533.

Yoon, S. and L. A. Albert. 2020. "A Dynamic Ambulance Routing Model with Multiple Response". *Transportation Research Part E: Logistics and Transportation Review* 133:101807.

Yoon, S. and L. A. Albert. 2021. "Dynamic Dispatch Policies for Emergency Response with Multiple Types of Vehicles". *Transportation Research Part E: Logistics and Transportation Review* 152:102405.

Zhen, L., K. Wang, H. Hu, and D. Chang. 2014. "A Simulation Optimization Framework for Ambulance Deployment and Relocation Problems". *Computers & Industrial Engineering* 72: 12–23.

AUTHOR BIOGRAPHIES

NAN KONG is a Professor in the Weldcon School of Biomedical Engineering at Purdue University. He earned his Ph.D. in the Department of Industrial Engineering at the University of Pittsburgh. He is Associate Director for Health Systems in the Purdue's Regenstrief Center for Healthcare Engineering. He is a Senior Member of IEEE. His research interests include stochastic programming, stochastic dynamic programming, integer programming, and healthcare and logistics applications of simulation and optimization methodologies. His e-mail address is nkong@purdue.edu, and website is http://engineering.purdue.edu/BASO.

JUAN C. PAZ is a PhD student in the School of Industrial Engineering at Purdue University. He received his bachelor and master's degree in Industrial Engineering from Universidad del Valle (Colombia) and Universidad Javeriana Cali (Colombia), respectively. He is a recipient of the Fullbright-Minciencias scholarship (U.S. - Colombia). His email address is paz3@purdue.edu.

XIAOQUAN GAO is a PhD student in the School of Industrial Engineering at Purdue University. She received her bachelor degree in Theoretical and Applied Mechanics from Peking University. Her email address is gao568@purdue.edu.