Production and Operations Management

Robust Data-Driven Design of a Smart Cardiac Arrest Response System

Weiliang Liu, Qiuzhuang Sun, Loon Ching Tang, Zhisheng Ye
Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576

Abstract: This paper studies the data-driven design of a smart emergency response system for out-of-hospital cardiac arrest (OHCA) that involves drones for automatic external defibrillator delivery and community responders alerted via a mobile application, in addition to ambulances. Our study is motivated by the widespread exploration of drones for delivery service, and the emergence of mobile applications that crowdsource community for emergency response. Based on a historical OHCA dataset with community responders' response records from Singapore, we develop a robust joint deployment model of drone and ambulance to maximize the survivability of the response system while accounting for data uncertainty in OHCA occurrence and responders' behavior. We discretize the planning area into finite demand regions, and allow different regions to have different OHCA demand rates, alert response probabilities and alert response time distributions from responders. Each of these attributes is only known to reside in an uncertainty/ambiguity set constructed from historical data. Our objective is to maximize the worst-case demand-weighted survival rate in the presence of uncertainty. We reformulate the resulting robust deployment model as a mixed-integer linear program, which can be efficiently solved by a proposed row-and-column generation algorithm with convergence guarantee. We illustrate our model and solution approach using real data from Singapore. We find that (i) hedging against uncertainty leads to a higher survival rate of the response system, compared to a sample average approximation deployment approach; (ii) while adding more drones/ambulances to the system exhibits diminishing return, a few drones are sufficient to increase the survival rate dramatically; and (iii) the impact of the behavior of responders on survival outcomes is more significant than that of simply adding drones/ambulances. We also discuss several managerial insights from the numerical experiments.

Key words: Emergency response, robust optimization, facility location, smart city operations, drones

1 Introduction

Being one of the most time-sensitive emergency incidents, out-of-hospital cardiac arrest (OHCA) accounts for approximately 400,000 deaths in North America every year, resulting in an annual economic productivity loss of 10.2 billion dollars (Coute et al. 2021). These figures are expected to rise owing to increasing population aging (Leeson 2018).

Successful treatment of OHCA depends crucially on times to three critical interventions, i.e., cardiopulmonary resuscitation (CPR), defibrillation via an automated external defibrillator (AED), and advanced care when patients are transported to hospitals by ambulances. The survival rate of OHCA has been estimated to decrease by 10% for each minute without these interventions (Larsen et al. 1993). However, traditional

emergency response systems for OHCA have suffered from a low CPR rate from bystanders and low utilization rate of static AEDs, which can be attributed in part to crowd's inadequate awareness of OHCA emergency, underdeveloped life-saving skills, and poor accessibility of static AEDs (Delhomme et al. 2019, Virani et al. 2021). This results in a low survival rate for OHCA in many countries, e.g., 3%–6% in Singapore (Lai et al. 2015, White et al. 2021). Improvement of the survival rate calls for a new concept of operations of the emergency response system.

Motivated by the recent advancement of smart city initiatives for healthcare innovation, this paper investigates the design of a smart response system for OHCA that integrates traditional ambulance dispatching with a first responder initiative (Scquizzato et al. 2020) and a drone-delivered AED initiative (Claesson et al. 2016). The former initiative employs a mobile application (App) to alert registered community responders of nearby OHCA incidents so as to start the chain of survival before an ambulance arrives, while the latter makes use of specially designed drones to deliver AEDs to incident sites. Upon receiving an OHCA report from a witnessing bystander, the emergency medical services (EMS) provider simultaneously (i) sends out an ambulance from the nearest ambulance base, (ii) sends out a drone loaded with an AED from the nearest drone base, and (iii) alerts a group of registered community responders nearby the site via a mobile App. As is the current practice of the *myResponder* App in Singapore (Ng et al. 2021), responders tap "accept" if they are willing to respond to the alert, after which they will be guided to the site with a detailed map, and they further tap "arrived" upon arrival to notify the EMS center. Instead of searching for a nearby AED, the responders can immediately perform CPR and watch out for the drone-delivered AED if they arrive earlier than the drone. Currently, the first responder programs via Apps have been adopted in cities like Singapore, London and San Francisco (Brooks et al. 2016, Smith et al. 2017). On the other hand, Sweden has recently conducted a prospective clinical trial on AED delivery through drone (Schierbeck et al. 2022). We are not aware of the use of both initiatives in either theory or applications. Without drones or responders, a traditional OHCA response system has to rely solely on static AEDs and bystanders within line-of-sight before ambulance arrival. In contrast, the proposed smart system is expected to greatly shorten the intervention time to CPR/defibrillation and thus significantly improve OHCA survival rate.

While the proposed smart system involving ambulances, drones and responders is conceptually appealing, the design and operations of such a complex system in a city scale are challenging. Since responders' behavior is exogenous, our primary objective is to optimize the deployment locations of ambulances and drones. [1] This deployment should be city-dependent because OHCA occurrences and responder locations are highly dependent on the city layout which determines the spatial density of population (Warden et al. 2012). Exact spatial densities of OHCA incidents and responders are generally unavailable, but they can

[1] Although ambulances respond to different types of emergencies, we consider deploying them to meet OHCA incidents because OHCA calls are among the highest priority EMS calls; moreover, this consideration allows us to present a more versatile facility location model capable of deploying both ambulances and drones.

be estimated from historical data, e.g., data extracted from the first responder Apps. For cities without such a first responder program, a pilot implementation is needed to understand responders' behavior before optimizing the system deployment. However, the use of such data during deployment should be cautious as future OHCA and responder patterns can deviate from the status quo due to changes in demographics, pandemics, progressive responders' recruitments and trainings (SingHealth 2022), etc. Moreover, the spatial densities computed from historical data are subject to data uncertainty, especially when the pilot program only collects limited data. These challenges necessitate a robust deployment approach. The development of such a robust approach, however, is not straightforward due to the structural missing pattern in historical data. In particular, responder arrival times are subject to right-censoring by ambulance arrival times, as expounded in Section 4.1. Another factor that complicates the deployment decision is the competing nature of the three rescue forces, i.e., ambulances, drones and responders. As a consequence, the deployment of ambulance and drone needs to be jointly decided while taking into account responders' behavior. However, since the intervention times to CPR and defibrillation become nonlinear functions of the response times of the three forces (due to competition), performance metrics commonly used in EMS optimization, such as total/average response time, are inappropriate in our joint deployment model. The deployment decision is further complicated by the need for high model fidelity. Since the golden response time for OHCA is circa six minutes (Huang et al. 2021), we require a fine-grained discretization of the city's spatial domain to capture varying densities of OHCA incidents and responders, and to accurately calculate ambulance/drone response times. Combined with the above robustification goal and the presence of nonlinear intervention times, it is highly likely that the deployment program would be computationally expensive.

This paper aims to address the aforementioned challenges faced by the proposed smart OHCA response system. We develop a robust deployment model that hedges against data uncertainty in OHCA demand and responders' behavior to jointly determine the locations of ambulances and drones in the system. The objective is to maximize the worst-case expected survivals of the system in view of the uncertainty using given numbers of ambulances and drones, while meeting a target constraint on ambulance's maximum allowable response time. By exploiting this target constraint and the strong duality theory of conic linear programming, we reformulate the high fidelity deployment program as a large-scale mixed-integer linear program (MILP). To overcome the computational hurdles, we develop an exact row-and-column generation algorithm for efficiently solving the MILP. We further apply our model and solution algorithm to conduct a case study using real data from Singapore. We summarize our contributions as follows.

1. To the best of our knowledge, we provide the first joint deployment model of ambulance and drone that incorporates responders' behavior in a robust data-driven manner to guide the design of a smart OHCA response system. Our model leverages historical data on OHCA demand and responder response while utilizing robust optimization to hedge against data uncertainty. The uncertainty in responders' behavior, which is essentially supply-side uncertainty, is rarely considered in EMS optimization literature. The proposed

uncertainty/ambiguity sets are tailored to the structural missing pattern in historical data, including right-censored responder response times. We show that the robust deployment model, which considers both the demand-side uncertainty in OHCA occurrence and the supply-side uncertainty in responders' behavior, can be reformulated as a large-scale MILP.

2. We develop an efficient row-and-column generation algorithm for solving the resulting large-scale MILP, and design a speedup strategy to accelerate convergence of the algorithm. The benefit of this strategy is both empirically observed and theoretically justified by a convergence analysis.

3. Using real data from Singapore, we conduct extensive numerical experiments to gain managerial insights into the smart response system. We find that: (i) Hedging against data uncertainty leads to system deployments that yield higher survival rates under a wide range of possible realizations of future OHCA incidents; (ii) while adding more drones/ambulances to the response system exhibits diminishing return, a few drones are sufficient to dramatically reduce the defibrillation time and increase the survival rate, compared to a system with no drone; (iii) the impact of responders' behavior on OHCA survival is significant. In the ideal scenario where all collapsed patients are immediately treated by responders, the survival rate can be improved by up to 15% compared with the status quo responder response. Such an improvement almost triples the largest possible improvement obtained by simply adding drones/ambulances; and (iv) a less stringent target on ambulance response time leads to a higher efficiency of the system in saving more lives, but a more imbalanced distribution of rescue resources among the city.

This paper is organized as follows. Section 2 reviews the literature. Section 3 introduces the joint deployment model. Section 4 presents the robust deployment approach. Section 5 analyzes the robust deployment problem and derives its equivalent MILP formulation. Section 6 proposes a row-and-column generation algorithm for solving the MILP. Section 7 analyzes a real dataset from Singapore to set up the deployment model. Section 8 presents our numerical results and discusses qualitative insights. Finally, we conclude this paper in Section 9. All proofs and supporting materials are relegated to the electronic companion.

2 Literature Review

Our study draws on the vast literature on facility location design for a fixed amount of emergency resources (e.g., ambulances, trauma centers, public access AEDs) on a nodal network. A majority of papers in this field can be categorized based on their objectives, where the first category maximizes the demand coverage (Cho et al. 2014, van den Berg and Aardal 2015, Chan et al. 2016), and the second minimizes the total/average travel time (Rajagopalan and Saydam 2009, Toro-Díaz et al. 2013). Our paper differs from the first category in that we encode spatial information of the nodal network in a condensed travel time matrix rather than a sparse 0-1 coverage matrix; and differs from those in the second in that we essentially minimize a weighted sum of intervention times due to their different importance in OHCA survival, which are nonlinear functions of travel times. Our paper adopts a similar objective function as that of Erkut et al.

(2008), who also linked OHCA survival rate to intervention times of CPR, defibrillation, advanced care, all of which were performed by ambulance. In contrast, intervention times in our model are determined by ambulance travel time, drone travel time and responders' behavior in a nonlinear way, and we take into account various sources of uncertainty.

Our paper contributes to the literature on facility location under uncertainty. Most research in this area has focused on hedging against the demand-side uncertainty. For example, Atamtürk and Zhang (2007) proposed a two-stage robust network optimization problem, and use a cardinality-restricted uncertainty set and a budget uncertainty set to capture the demand weights uncertainty; Baron et al. (2011) considered a robust fixed-charge network location problem and compared the resulting solutions when the demand weight vector lies within a box uncertainty set and an ellipsoid uncertainty set. In the context of emergency response, Chan et al. (2018) studied the problem of AED deployment under cardiac arrest location uncertainty, where the demand weight vector representing the probability of next cardiac arrest arrival lies in a polyhedral uncertainty set. In contrast to the above, our paper employs a scenario-based uncertainty set to model demand uncertainty. More notably, in addition to the demand-side uncertainty, our paper considers the supply-side uncertainty in responders' behavior.

Our paper also contributes to the operations research of OHCA responses. The majority of work in this area focused on deploying static AEDs in public area, with a view to either maximize demand coverage (Siddiq et al. 2013, Chan et al. 2013, 2016, Sun et al. 2016), minimize risk (Derevitskii et al. 2020), or minimize the tail of travel time distribution (Chan et al. 2018). In view of the low utilization rate of static AEDs (Delhomme et al. 2019), some recent studies explored the use of drones to deliver AED (Pulver et al. 2016, Claesson et al. 2016, Boutilier et al. 2017, Boutilier and Chan 2022). Among the existing literature, our study is particularly inspired by Chan et al. (2018) and Boutilier and Chan (2022), who laid the groundwork for using an optimization-based approach to deploy OHCA-related medical resources, such as AEDs and drones. However, our work differs from theirs in several key ways and presents original contributions: (i) Neither Chan et al. (2018) nor Boutilier and Chan (2022) considered the behavior of responders in their deployment models, whereas we explicitly characterize responders' uncertain and geographically diverse response probabilities and response time distributions, allowing us to reveal the significant impact of responders' behavior on OHCA survival, as shown in the numerical study in Section 8. (ii) Both Chan et al. (2018) and Boutilier and Chan (2022) constructed their models based on traditional OHCA datasets that only recorded historical OHCA locations. In contrast, our model is data-driven by a novel dataset that records responders' behaviors in addition to exact OHCA locations. This complex dataset calls for innovation in modeling approach and offers new insights, as illustrated in Sections 4 and 7. (iii) Chan et al. (2018) only considered the uncertainty in OHCA demand and proposed a demand uncertainty set that is based on complete data; by contrast, we consider uncertainties in both demand and supply, and present uncertainty/ambiguity sets that are adapted to the structural missing pattern in historical data. (iv) Boutilier and

Chan (2022) did not consider data uncertainty and treated ambulance's response time as known to solely optimize for drone deployment; the known ambulance's response time is the key to their models' computational tractability. This leaves the question of how ambulance network can be adapted to integrate drones in a response system unresolved. Our work fills in this gap by presenting the first robust joint deployment model of ambulance and drone.

Last but not least, our paper contributes to the emerging operations research literature on drone applications. The majority of this literature considered hybrid truck-drone delivery models, which was initiated by Murray and Chu (2015) and extended to numerous variants (Agatz et al. 2018, Carlsson and Song 2018, Wang and Sheu 2019). However, there is relatively scant literature on the applications of drone in a specific healthcare area. The exceptions are Pulver et al. (2016), Claesson et al. (2016), Boutilier et al. (2017), Kim et al. (2017), Scott and Scott (2019) and Boutilier and Chan (2022).

Notation. We use boldface lowercase letters (e.g., $\boldsymbol{y}$) to represent vectors, and use calligraphic uppercase letters (e.g., $\mathcal{Q}$) to represent sets. We use $[I] \equiv \{1,\ldots,I\}$ to denote the set of running indices up to I, where I is a known positive integer. A random variable $\tilde{\boldsymbol{u}}$ is associated with a tilde sign, and we use $\tilde{\boldsymbol{u}} \sim \mathbb{P}$, $\mathbb{P} \in \mathcal{P}_0(\mathcal{U})$ to denote that $\tilde{\boldsymbol{u}}$ follows a probability distribution $\mathbb{P}$ with support $\mathcal{U}$, where $\mathcal{P}_0(\mathcal{U})$ is the set of all probability distributions supported on $\mathcal{U}$. For a joint distribution $(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}}) \sim \mathbb{Q}$, $\mathbb{Q} \in \mathcal{P}_0(\mathcal{U}\times\mathcal{V})$, we use $\Pi_{\tilde{\boldsymbol{u}}}\mathbb{Q}$ to denote the marginal distribution of $\tilde{\boldsymbol{u}}$. For an event $\mathcal{E}$, we use $\Pi_{\tilde{\boldsymbol{u}}|\mathcal{E}}\mathbb{Q}$ to denote the marginal distribution of $\tilde{\boldsymbol{u}}$ conditioned on that event. This is extended to an ambiguity set $\mathcal{Q} \subseteq \mathcal{P}_0(\mathcal{U}\times\mathcal{V})$, where we define $\Pi_{\tilde{\boldsymbol{u}}}\mathcal{Q} = \bigcup_{\mathbb{Q}\in\mathcal{Q}}\{\Pi_{\tilde{\boldsymbol{u}}}\mathbb{Q}\}$ and $\Pi_{\tilde{\boldsymbol{u}}|\mathcal{E}}\mathcal{Q} = \bigcup_{\mathbb{Q}\in\mathcal{Q}}\{\Pi_{\tilde{\boldsymbol{u}}|\mathcal{E}}\mathbb{Q}\}$.

3 The Joint Deployment Model

Consider a continuous planning area over which OHCA incidents arrive. We discretize the planning area into I demand regions and assume that OHCA demand concentrates in region centers. Denote $\boldsymbol{d}^* \triangleq (d_1^*, d_2^*, \ldots, d_I^*)$ as the demand rate of OHCA incidents in these I regions. There are J candidate locations for drones and K candidate locations for ambulances in the planning area, which we index by $j \in [J]$ and $k \in [K]$, respectively. Let t_{ij}^{D} (resp. t_{ik}^{A}) be the time needed for a drone (resp. an ambulance) located at location $j \in [J]$ (resp. $k \in [K]$) to travel to region $i \in [I]$.

Let P^{D} and P^{A} be the numbers of drones and ambulances that are available for deployment. Because OHCA incidents are rare events, we do not consider system congestion and assume that each candidate drone (ambulance) location can be deployed at most one drone (ambulance); it can also be interpreted that we are primarily interested in determining the locations of drone/ambulance bases. Our objective is to optimize the smart OHCA response system by properly deploying the P^{D} drones and P^{A} ambulances. Given a deployment scheme, we introduce a binary vector $\boldsymbol{y}^{\mathrm{D}} \in \{0,1\}^J$ (resp. $\boldsymbol{y}^{\mathrm{A}} \in \{0,1\}^K$), where $y_j^{\mathrm{D}} = 1$ (resp. $y_k^{\mathrm{A}} = 1$) if candidate location $j \in [J]$ ($k \in [K]$) is chosen as a drone base (resp. ambulance base).

3.1 Event Timeline and Performance Measure

Whenever an OHCA incident occurs at region $i \in [I]$, a bystander witnesses it and notifies the EMS provider immediately. Upon notification, the EMS provider simultaneously sends out an AED-loaded drone and an ambulance from the nearest drone base $j(i) \in [J]$ and the nearest ambulance base $k(i) \in [K]$, and the travel times of the drone and ambulance are $t^{\mathrm{D}}_{i,j(i)}$ and $t^{\mathrm{A}}_{i,k(i)}$, respectively. Here, $j(i)$ and $k(i)$ are known once the deployment decisions $\boldsymbol{y}^{\mathrm{D}}$ and $\boldsymbol{y}^{\mathrm{A}}$ are fixed. In addition, the EMS provider also alerts registered responders nearby through the mobile App, e.g., *myResponder*. Let $\tilde{\delta}_i$ be an indicator, where $\tilde{\delta}_i = 1$ if at least one responder responds to the alert, and $\tilde{\delta}_i = 0$ otherwise. When $\tilde{\delta}_i = 1$, let $\tilde{u}_i$ be the travel time of the *first responder* to arrive at the site, which we refer to as the *alert response time*. We make the following assumptions on the three interventions.

- CPR is performed by either the responder or the paramedics on ambulance, whoever arrives earlier.
- Defibrillation can be performed by the bystander and the responder as long as the AED-loaded drone arrives. It can also be performed by the paramedics on ambulance.
- Advanced care is performed at the nearest hospital, and it takes a deterministic duration of time o_i for the ambulance to travel from the incident site to the nearest hospital.

The first assumption is based on the observation that CPR rates initiated by bystander without CPR training have been consistently low from 2006 to 2019, according to the heart disease and stroke statistics in Virani et al. (2021). For the second assumption, AEDs are handy to use and they are designed to be used by members of the general public regardless of training (Gundry et al. 1999). Therefore, it is reasonable to assume that the bystander, with audio/video guidance of the EMS provider through the communication module in the AED-loaded drone, can perform defibrillation properly. In the last assumption, the ambulance travel time o_i is determined by the distance from the nearest hospital to region i. Because the geographical locations of the hospitals are fixed, o_i is more or less fixed for each region and varies from region to region. We realize that in reality, not all OHCA incidents are witnessed and not all patients have shockable rhythms, but these are not captured in the deployment model in order to maintain a concise presentation.

With the above assumptions, we can characterize the respective intervention times to CPR, defibrillation and advanced care for the incident at region $i \in [I]$ by

$$\begin{aligned} T_i^{\mathrm{CPR}}(\tilde{\delta}_i, \tilde{u}_i, \boldsymbol{y}^{\mathrm{A}}) &\triangleq \tilde{\delta}_i \min\left\{\tilde{u}_i, t^{\mathrm{A}}_{i,k(i)}\right\} + (1-\tilde{\delta}_i) t^{\mathrm{A}}_{i,k(i)}; \\ T_i^{\mathrm{DEF}}(\boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}}) &\triangleq \min\left\{t^{\mathrm{D}}_{i,j(i)}, t^{\mathrm{A}}_{i,k(i)}\right\}; \\ T_i^{\mathrm{AC}}(\boldsymbol{y}^{\mathrm{A}}) &\triangleq t^{\mathrm{A}}_{i,k(i)} + o_i. \end{aligned} \tag{1}$$

The survival rate of the OHCA patient depends heavily on these three intervention times. According to the empirical study in Larsen et al. (1993), the survival rate can be reasonably approximated using a linear regression of the three intervention times as

$$s_i(\tilde{\delta}_i, \tilde{u}_i, \boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}}) = \beta_0 - \beta_1 T_i^{\mathrm{CPR}}(\tilde{\delta}_i, \tilde{u}_i, \boldsymbol{y}^{\mathrm{A}}) - \beta_2 T_i^{\mathrm{DEF}}(\boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}}) - \beta_3 T_i^{\mathrm{AC}}(\boldsymbol{y}^{\mathrm{A}}), \quad i \in [I], \tag{2}$$

where $\beta_0 \in (0,1)$ represents the baseline survival rate if professional EMS can be applied immediately upon collapse, and $\beta_1, \beta_2, \beta_3$ are positive regression coefficients that capture the decrease in survival rate per unit of time passed without implementing the three interventions in an OHCA rescue. Values of these parameters have been empirically determined in Larsen et al. (1993) using a large dataset on 1,667 cardiac arrest patients who had received CPR and defibrillation before arrival of EMS personnel. We do not pursue a more complicated model to approximate survival here, e.g., a logistic regression model with explanatory variables including patient's age, sex and health history, because these variables are beyond the control of the EMS system and because nonlinear survival model would render subsequent performance optimization intractable.

The above survival rate (2) is random as the alert response indicator $\tilde{\delta}_i$ and the alert response time $\tilde{u}_i$ are random. In an emergency response system, the primary interest is usually the overall survival in the planning area. Therefore, we use the demand-weighted expected survival rate

$$\mathbb{E} \sum_{i\in[I]} d_i^* s_i(\tilde{\delta}_i, \tilde{u}_i, \boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}})$$

as the overall performance measure of a joint deployment decision $(\boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}})$. This measure directly reflects the expected number of OHCA patients that can be saved via the smart response system. Using this measure as our objective is more intuitive than risk measures like Value at Risk or Conditional Value at Risk. Because $\beta_0, \beta_1, \beta_3$ and o_i, d_i^* are constants independent of the drone/ambulance deployment decision, maximizing the above demand-weighted survival rate is equivalent to minimizing the following objective:

$$\mathbb{E} \sum_{i\in[I]} d_i^* \left(\tilde{\delta}_i \min\left\{\tilde{u}_i, t^{\mathrm{A}}_{i,k(i)}\right\} + (1-\tilde{\delta}_i) t^{\mathrm{A}}_{i,k(i)} + \rho_1 \min\left\{t^{\mathrm{D}}_{i,j(i)}, t^{\mathrm{A}}_{i,k(i)}\right\} + \rho_2 t^{\mathrm{A}}_{i,k(i)} \right), \tag{3}$$

where $\rho_1 \triangleq \beta_2/\beta_1,\ \rho_2 \triangleq \beta_3/\beta_1$. Objective function (3) can be interpreted as the demand-weighted survival churn rate (normalized by β_1) due to untimely interventions.

To evaluate (3), we need to model the alert response indicators $\tilde{\delta}_i$ and the response times $\tilde{u}_i$, $i \in [I]$. Due to the heterogeneity of responder density across regions, a region with a higher responder density tends to have a higher chance of getting at least one responder, i.e., a higher probability of $\tilde{\delta}_i = 1$. This is also evidenced from the real data visualized in Figure 2 ahead. As such, we model $\tilde{\delta}_i$ as a Bernoulli random variable, and let p_i^* be the *alert response probability* that $\tilde{\delta}_i = 1$. We allow p_i^* to be region-dependent to capture such heterogeneity. Once at least a responder responds, the alert response time $\tilde{u}_i$, $i \in [I]$, is a nonnegative continuous random variable with distribution $\mathbb{P}_i^*$. The response time distribution $\mathbb{P}_i^*$ is also region-dependent. For example, a responder may take more time to reach the incident site in regions with dense high-rise buildings and/or complex road structures. While in an open region like a national park, a responder can easily locate the incident site. Let $\mathbb{Q}^*$ be the true distribution of $\tilde{\boldsymbol{u}} \triangleq (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_I)$. Then the objective (3) can be written as

$$\mathbb{E}_{\mathbb{Q}^*} \sum_{i\in[I]} d_i^* \left(p_i^* \min\{\tilde{u}_i, t^{\mathrm{A}}_{i,k(i)}\} + \rho_1 \min\{t^{\mathrm{D}}_{i,j(i)}, t^{\mathrm{A}}_{i,k(i)}\} + (1 - p_i^* + \rho_2) t^{\mathrm{A}}_{i,k(i)} \right). \tag{4}$$

Next, we specify constraints for the joint deployment decision $\boldsymbol{y}^{\mathrm{D}}$ and $\boldsymbol{y}^{\mathrm{A}}$, and also quantify the nearest drone base $j(i)$ and ambulance base $k(i)$ to region i under a given deployment.

3.2 Constraints on Drone/Ambulance Deployment

First, the deployment is subject to the budget constraint that up to P^{D} drones and P^{A} ambulances are available. Aside from this, the deployment of ambulances need to satisfy a target constraint on the maximum allowable response time, i.e., every incident should be attended by an ambulance within a target τ units of response time. This constraint highlights the equity concern of an emergency response system and is in accordance with the prevailing EMS regulations. For example, the US EMS Act requires that at least 95% of the emergency requests should be served within 10 minutes in urban areas. By taking this extra constraint on ambulance into consideration, the feasible set for drone deployment is given by

$$\mathcal{Y}^{\mathrm{D}} \triangleq \left\{ \boldsymbol{y}^{\mathrm{D}} \in \{0,1\}^{J} \middle| \textstyle\sum_{j\in[J]} y_j^{\mathrm{D}} = P^{\mathrm{D}} \right\}, \tag{5}$$

and the feasible set for ambulance deployment is

$$\mathcal{Y}^{\mathrm{A}} \triangleq \left\{ \boldsymbol{y}^{\mathrm{A}} \in \{0,1\}^{K} \middle| \textstyle\sum_{k\in[K]} y_k^{\mathrm{A}} = P^{\mathrm{A}}, \sum_{k\in\mathcal{K}_i} y_k^{\mathrm{A}} \geq 1, i \in [I] \right\}, \tag{6}$$

where $\mathcal{K}_i \triangleq \{k \in [K] \mid t_{ik}^{\mathrm{A}} \leq \tau\}$ is the index set of candidate ambulance bases that cover region i within response time τ.

Next, we quantify the nearest drone base $j(i)$ and ambulance base $k(i)$ to region i for a given joint deployment decision $\boldsymbol{y}^{\mathrm{D}}$ and $\boldsymbol{y}^{\mathrm{A}}$. To this end, we introduce an assignment variable x_{ij}^{D} (resp. x_{ik}^{A}) that is equal to 1 if region i is assigned to be served by drone base j (resp. ambulance base k). Let $\boldsymbol{x}^{\mathrm{D}}$ and $\boldsymbol{x}^{\mathrm{A}}$ be the collection of x_{ij}^{D} and x_{ik}^{A}, respectively. The feasible sets of the assignments $\boldsymbol{x}^{\mathrm{D}}$ and $\boldsymbol{x}^{\mathrm{A}}$, given the joint deployment decision $\boldsymbol{y}^{\mathrm{D}}$ and $\boldsymbol{y}^{\mathrm{A}}$, can be expressed as

$$\mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}) \triangleq \left\{ x_{ij}^{\mathrm{D}} \middle| \textstyle\sum_{j\in[J]} x_{ij}^{\mathrm{D}} = 1, 0 \leq x_{ij}^{\mathrm{D}} \leq y_j^{\mathrm{D}}, i \in [I], j \in [J] \right\}, \tag{7}$$

and

$$\mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}) \triangleq \left\{ x_{ik}^{\mathrm{A}} \middle| \textstyle\sum_{k\in[K]} x_{ik}^{\mathrm{A}} = 1, 0 \leq x_{ik}^{\mathrm{A}} \leq y_k^{\mathrm{A}}, i \in [I], k \in [K] \right\}, \tag{8}$$

respectively. The feasible set (7) stipulates that each region is served by exactly one drone base, and a candidate drone base j is available for service only if a drone is deployed on it, i.e., $y_j^{\mathrm{D}} = 1$. The same interpretation applies to the feasible set (8) for ambulance assignment. Since the travel times t_{ij}^{D} and t_{ik}^{A} are known, the nearest drone base $j(i)$ and ambulance base $k(i)$ to region i, as well as the corresponding travel times, can be determined through the following programs

$$t_{i,j(i)}^{\mathrm{D}} = \min_{\boldsymbol{x}^{\mathrm{D}} \in \mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}})} \sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}} \tag{9}$$

$$t_{i,k(i)}^{\mathrm{A}} = \min_{\boldsymbol{x}^{\mathrm{A}} \in \mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}})} \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}. \tag{10}$$

The above are two linear programs with polyhedral constraints. The optimal solutions are at some extreme point of the feasible set so the optimal assignment variables will automatically be binary. Therefore, we do not need to explicitly make the binary declarations in (7) and (8), which reduces the complexity of the optimization problem proposed next.

3.3 The Nominal Joint Deployment Problem

With the above set-up, the nominal joint drone/ambulance deployment problem that minimizes (4) can be written by combining (4) and (7)–(10), which is given by

$$\begin{aligned}\min_{\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}}} \ & \mathbb{E}_{\mathbb{Q}^*}\left[\sum_{i\in[I]} d_i^*\left(p_i^*\min\{\tilde{u}_i,\sum_{k\in[K]} x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}+\rho_1\min\{\sum_{j\in[J]} x_{ij}^{\mathrm{D}}t_{ij}^{\mathrm{D}},\sum_{k\in[K]} x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}\right.\right.\\ & \left.\left.+(1-p_i^*+\rho_2)\sum_{k\in[K]} x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\right)\right] \qquad \text{(N–JD)}\\ \text{s.t.} \ & \boldsymbol{y}^{\mathrm{D}}\in\mathcal{Y}^{\mathrm{D}},\ \boldsymbol{y}^{\mathrm{A}}\in\mathcal{Y}^{\mathrm{A}}\\ & \boldsymbol{x}^{\mathrm{D}}\in\mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}),\ \boldsymbol{x}^{\mathrm{A}}\in\mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}).\end{aligned}$$

In the above program, the objective function is obtained by substituting $t_{i,j(i)}^{\mathrm{D}}$ and $t_{i,k(i)}^{\mathrm{A}}$ derived in (9) and (10) into (4). We call it the nominal problem because it uses the true demand rate $\boldsymbol{d}^*$, the true alert response probability $\boldsymbol{p}^*$, and the true probability distribution $\mathbb{Q}^*$ of the alert response time $\tilde{\boldsymbol{u}}$. In practice, however, it is generally impossible to obtain perfect information on $\boldsymbol{d}^*$, $\boldsymbol{p}^*$ and $\mathbb{Q}^*$, and any estimate of them is likely to subject to uncertainty. In the next section, we adopt the robust deployment approach to hedge against this uncertainty.

4 The Robust Deployment Approach

As argued, we address the issue of limited information on $\boldsymbol{d}^*$, $\boldsymbol{p}^*$ and $\mathbb{Q}^*$ by using the robust deployment approach. With this approach, instead of assuming perfect information on the demand rate vector $\boldsymbol{d}$, the alert response probability vector $\boldsymbol{p}$, and the alert response time distribution $\mathbb{Q}$, we assume that (i) $\boldsymbol{d}$ is only known to reside in an uncertainty set $\mathcal{D}\subseteq\mathbb{R}^I$, (ii) $\boldsymbol{p}$ is only known to reside in an uncertainty set $\mathcal{P}(\boldsymbol{d})\subseteq\mathbb{R}^I$ that is dependent on the realization of $\boldsymbol{d}$, (iii) $\mathbb{Q}$ is only known to belong to an ambiguity set $\mathcal{Q}\subseteq\mathcal{P}_0(\mathbb{R}_+^I)$. This assumption aligns with the fact that we are, at best, only aware of several structural properties of $\boldsymbol{d}$, $\boldsymbol{p}$ and $\mathbb{Q}$ that are consistent with the observed data. Then the robust joint deployment problem which minimizes (4) over the worst-case scenario is given by

$$\begin{aligned}\min_{\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}}} \ \max_{\boldsymbol{d}\in\mathcal{D},\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d})} \ \max_{\mathbb{Q}\in\mathcal{Q}} \ & \mathbb{E}_{\mathbb{Q}}\left[\sum_{i\in[I]} d_i\left(p_i\min\{\tilde{u}_i,\sum_{k\in[K]} x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}\right.\right.\\ & \left.\left.+\rho_1\min\{\sum_{j\in[J]} x_{ij}^{\mathrm{D}}t_{ij}^{\mathrm{D}},\sum_{k\in[K]} x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}+(1-p_i+\rho_2)\sum_{k\in[K]} x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\right)\right] \qquad \text{(R–JD)}\\ \text{s.t.} \ & \boldsymbol{y}^{\mathrm{D}}\in\mathcal{Y}^{\mathrm{D}},\ \boldsymbol{y}^{\mathrm{A}}\in\mathcal{Y}^{\mathrm{A}}\\ & \boldsymbol{x}^{\mathrm{D}}\in\mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}),\ \boldsymbol{x}^{\mathrm{A}}\in\mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}).\end{aligned}$$

The uncertainty set $\mathcal{P}$ for the alert response probability $\boldsymbol{p}$ is expected to be dependent on the demand rate $\boldsymbol{d}$ because both $\boldsymbol{p}$ and $\boldsymbol{d}$ are governed by the population density of a region. The sets $\mathcal{D}$, $\mathcal{P}$ and $\mathcal{Q}$ should be carefully designed so that they can be constructed from historical data on OHCA and responder response, and that the resulting robust program is tractable and not over-conservative due to the multiple layers of robustness. In the following, we provide a data-driven recipe to the construction of these sets.

4.1 Ambiguity Set for Alert Response Time Distribution

Our construction of $\mathcal{Q}$ is motivated by a unique feature of the responder response data, that is, right censoring. Right censoring is common in many applications, e.g., clinical trial study, reliability testing, and retail demand estimation. It refers to the scenario that instead of observing the actual value of a random variable, we only know that it is larger than a certain value. In our historical response dataset with 292 incidents, the arrival time of the ambulance was recorded for every incident, but the arrival time of a responder was recorded only if the responder arrived earlier than the ambulance. This is because when the ambulance arrived at the site, all responders still on the way would receive a message from the mobile App notifying that their services were no longer needed. Furthermore, if more than one responder arrived at the incident site, only the earliest arrival time would be recorded, which is a built-in feature of the App. Therefore, the random ambulance response time actually serves as a censoring variable to the alert response time. Under this arbitrary censoring scenario, a commonly used nonparametric estimator for the distribution $\Pi_{\tilde{u}_i}\mathbb{Q}$ of the alert response time $\tilde{u}_i$ in region i is the Kaplan-Meier estimator (Kaplan and Meier 1958). However, this estimator is subject to estimation uncertainty, and it estimates $\Pi_{\tilde{u}_i}\mathbb{Q}$ only up to the maximum of the observed/censored alert response times in the dataset.

To address the two issues, recall that we have set a maximum allowable response time τ for the ambulance deployment in Section 3. As such, the exact alert response time distribution $\Pi_{\tilde{u}_i}\mathbb{Q}$ beyond τ is unimportant. In view of this, we use τ to divide the support of $\Pi_{\tilde{u}_i}\mathbb{Q}$ into two intervals $[0,\tau]$ and (τ,∞). The estimation uncertainty of $\Pi_{\tilde{u}_i}\mathbb{Q}$ over (τ,∞) can be easily quantified through the probability of $\tilde{u}_i$ falling in this interval, i.e., $\Pi_{\tilde{u}_i}\mathbb{Q}[\tilde{u}_i > \tau] \in [\underline{q}_i, \bar{q}_i]$, where $\underline{q}_i, \bar{q}_i$ represent lower and upper uncertainty intervals for the actual probability. On the other hand, we propose to quantify the estimation uncertainty of $\Pi_{\tilde{u}_i}\mathbb{Q}$ in the first interval using a mean-dispersion ambiguity set. The mean and dispersion of $\tilde{u}_i$ over the entire real line can be uncomputable due to censoring. Conditional on $\tilde{u}_i \le \tau$, however, the mean and dispersion of $[\tilde{u}_i|\tilde{u}_i \le \tau]$ are computable as long as the maximum of the observed/censored alert response times in region i is not smaller than τ. Inspired by this fact, we let $\breve{u}_i$ and ε_i be the respective conditional mean and absolute dispersion of $[\tilde{u}_i|\tilde{u}_i \le \tau]$, both of which can be computed from the Kaplan-Meier curve. Then, we capture the ambiguity of $\Pi_{\tilde{u}_i}\mathbb{Q}$ in the first interval by allowing it to be any distribution that is consistent with this information. That

is, $\Pi_{\tilde{u}_i}\mathbb{Q}$ should satisfy $\mathbb{E}_{\Pi_{\tilde{u}_i}\mathbb{Q}}[\tilde{u}_i|\tilde{u}_i \le \tau] = \check{u}_i$ and $\mathbb{E}_{\Pi_{\tilde{u}_i}\mathbb{Q}}[|\tilde{u}_i - \check{u}_i|\,|\tilde{u}_i \le \tau] \le \varepsilon_i$. Consolidating the above results, we construct the ambiguity set for the alert response time distribution as

$$\mathcal{Q} \triangleq \left\{ \mathbb{Q} \in \mathcal{P}_0(\mathbb{R}_+^I) \,\middle|\, \begin{array}{l} \Pi_{\tilde{u}_i}\mathbb{Q} = \mathbb{P}_i,\ i \in [I] \\ \mathbb{P}_i\left[\tilde{u}_i > \tau\right] \in [\underline{q}_i, \bar{q}_i] \\ \mathbb{E}_{\mathbb{P}_i}\left[\tilde{u}_i \middle| \tilde{u}_i \le \tau\right] = \check{u}_i \\ \mathbb{E}_{\mathbb{P}_i}\left[|\tilde{u}_i - \check{u}_i| \middle| \tilde{u}_i \le \tau\right] \le \varepsilon_i \end{array} \right\}. \tag{11}$$

When the historical dataset is not large enough and/or when the planning area is divided into many small regions, the above construction can lead to a large ambiguity set for $\mathbb{Q}$, and the parameters $(\underline{q}_i, \bar{q}_i, \check{u}_i, \varepsilon_i)$ cannot even be computed if the maximum of the observed/censored alert response times in region i is smaller than τ. In this case, we suggest clustering $\Pi_{\tilde{u}_i}\mathbb{Q}$ of the I demand regions based on similarities in their Kaplan-Meier curves and geographical locations. The alert response time distributions of regions in the same cluster are homogeneous and thus can be treated as identical, and the common distribution can be estimated more accurately by aggregating response data in these regions. This strategy is adopted in our case study, as detailed in Appendix EC.4.

4.2 Uncertainty Sets for Demand Rate and Alert Response Probability

The uncertainties in $\boldsymbol{d}$ and $\boldsymbol{p}$ come from sampling errors and possible pattern shifts in OHCA demand and responders' behavior. A common way to capture the estimation uncertainty is through confidence intervals. For a given discretization scheme of the planning area, however, the dimension of $\boldsymbol{d}$ and $\boldsymbol{p}$ is usually high, and interval estimation for a high-dimensional vector with limited data is well-known to be difficult and even impossible (Javanmard and Montanari 2014). For the demand rate $\boldsymbol{d}$, this study adopts a resampling-based approach that fits a two-dimensional distribution to the locations of historical OHCA incidents using kernel density estimation (KDE), and then resamples from this fitted density. Through resampling from the KDE, we can generate OHCA incidents of the same size as the training sample and use them to compute a bootstrapped estimate of $\boldsymbol{d}$, which is known as a *demand scenario* in the robust optimization literature. The above procedure can be repeated B times to form a scenario-based uncertainty set for $\boldsymbol{d}$ as

$$\mathcal{D} \triangleq \left\{ \boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)}, \ldots, \boldsymbol{d}^{(B)} \right\}. \tag{12}$$

In the above, each scenario $\boldsymbol{d}^{(b)}$, $b \in [B]$, has a dimension of I, i.e., the number of demand regions. Due to the high dimensionality of $\boldsymbol{d}$, a large number of demand scenarios should be generated from the fitted density for sufficient uncertainty quantification.

As for the set $\mathcal{P}$, it is difficult to apply the above resampling procedure, because (i) some regions have few OHCA incidents, and estimating $\mathcal{P}$ in these regions can be inaccurate, (ii) the alert responses in two neighboring regions can be correlated as they might be handled by responders residing nearby, and the

correlation makes it difficult to write the likelihood function for estimating $\boldsymbol{p}$ based on historical response data. In view of these issues and the high dimensionality of $\boldsymbol{p}$, we capture its uncertainty from an aggregate level. That is, we aggregate the I demand regions into several clusters, and estimate an average alert response probability for OHCA incidents in the cluster level. Compared with estimating each p_i, estimating the average response probability over a cluster of multiple regions is more stable, i.e., a narrower confidence interval, as aggregation increases the sample size. Apart from this, if we construct the uncertainty interval for each p_i individually, i.e., $p_i \in [\underline{p}_i, \bar{p}_i]$ for $i \in [I]$, the greedy nature of the robust technique enforces that in the worst-case scenario, all p_i will take the lower bound $\underline{p}_i$ in that interval. Then by the nature of optimization, the deployment program will mostly use the rank of $\underline{p}_i$, $i \in [I]$, to select ambulance locations, which can lead to a non-robust solution. On the other hand, if the uncertainty interval is constructed at the cluster level, the worst-case scenario of $\boldsymbol{p}$ will depend on the location decision of ambulances (see Proposition 3 and the discussion therein), and the deployment program will try to hedge against this uncertainty. To better support our argument, we consider in Appendix EC.2.1 a simplified system with two regions, and show that when the uncertainty interval is constructed for individual p_i, but the order of the estimated lower bounds $\{\underline{p}_i\}_{i\in[I]}$ does not agree with the order of the true values $\{p_i^*\}_{i\in[I]}$, the deployment program with uncertainty interval constructed at the cluster level is guaranteed to produce a better deployment.

To enact the above idea, suppose the I demand regions are aggregated into V (possibly overlapping) clusters $I_{(1)}, I_{(2)}, \ldots, I_{(V)}$ with $\bigcup_v I_{(v)} = [I]$. In cluster v, we first use historical data to construct an uncertainty interval $[\underline{p}_{(v)}, \bar{p}_{(v)}]$ for the average alert response probability, e.g., using the Wilson score method. With this interval, we construct the uncertainty set for $\boldsymbol{p}$, the alert response probabilities for all the I regions, to be dependent on the bth demand scenario $\boldsymbol{d}^{(b)}$ as

$$\mathcal{P}(\boldsymbol{d}^{(b)}) \triangleq \left\{ \boldsymbol{p} \in [0,1]^I \middle| \frac{\sum_{i\in I_{(v)}} d_i^{(b)} p_i}{\sum_{i\in I_{(v)}} d_i^{(b)}} \in [\underline{p}_{(v)}, \bar{p}_{(v)}], v \in [V] \right\}. \tag{13}$$

The above construction is reasonable because under demand scenario $\boldsymbol{d}^{(b)}$, the average alert response rate in cluster v is $\sum_{i\in I_{(v)}} d_i^{(b)} p_i / \sum_{i\in I_{(v)}} d_i^{(b)}$. We require $\mathcal{P}$ to be dependent on $\boldsymbol{d}^{(b)}$ because both the alert response probability and the demand rate are generally increasing with the population density in a region, and thus they are functionally dependent. In Appendix EC.2.3, we show that if we instead use the estimated demand rate $\hat{\boldsymbol{d}}$ in the above construction of $\mathcal{P}$, it is possible that there is not any alert response under some extreme demand scenario $\boldsymbol{d}^{(b)}$, i.e., $p_i = 0$ if $d_i^{(b)} > 0$. Such over-conservatism is because the functional dependence between the set $\mathcal{P}$ and the demand scenario $\boldsymbol{d}^{(b)}$ is ignored. It can be avoided by using $\mathcal{P}(\boldsymbol{d}^{(b)})$ that directly requires $\sum_{i\in I_{(v)}} d_i^{(b)} p_i$ to be lower bounded by $\underline{p}_{(v)} \sum_{i\in I_{(v)}} d_i^{(b)}$.

The remaining issue is the design of the aggregation scheme $\{I_{(v)}\}_{v\in[V]}$. Firstly, there is a tradeoff in the choice of the cluster number V. On one hand, increasing the cluster number equips the set $\mathcal{P}$ with more uncertainty intervals $[\underline{p}_{(v)}, \bar{p}_{(v)}]$, $v \in [V]$. This decreases the size of the uncertainty set $\mathcal{P}$, so the robust

deployment program tends to produce a more aggressive decision. Such a decision can even be optimal to the nominal problem (N–JD) in some cases when the estimated intervals satisfy certain conditions related to the true $\boldsymbol{p}^*$ (see Appendix EC.2.1). On the other hand, increasing V inevitably reduces the sample size in each cluster, which can make the estimated intervals less accurate and fail to satisfy those conditions. In this case, the aggressive decision can instead have worse performance compared to the one with fewer clusters (see also Appendix EC.2.1). Therefore, the cluster number V should be carefully chosen based on the availability of historical data. Secondly, when the number of clusters is fixed, the demand regions should ideally be aggregated based on their spatial proximity. This argument is supported by our finding in Appendix EC.2.2, where we compare different aggregation schemes in a simplified system with three regions. We rigorously show in this simplified setting that if the cluster is formed by spatially dispersed regions, the deployment program tends to conservatively locate the ambulance at the center point of these dispersed regions regardless of ranges of the uncertainty intervals. The insight behind is that under such an aggregation scheme, the worst-case nature tends to assign a low alert response probability p_i to regions far away from the ambulance's location, and thus deploying the ambulance at any non-center point tends to yield a higher objective value. Such conservatism can be avoided by aggregating spatially nearby regions as a cluster. The above two considerations should be taken into account when devising a data-driven aggregation scheme $\{I_{(v)}\}_{v\in[V]}$, as demonstrated in Appendix EC.5.

5 Robust Deployment Problem Analysis and Reformulation

This section investigates analytical properties of the robust joint deployment problem (R–JD), and shows that it admits an equivalent MILP reformulation under the uncertainty/ambiguity sets proposed in the previous section. We start with the evaluation of the inner maximization problem over $\mathbb{Q}\in\mathcal{Q}$ in (R–JD). The results are summarized as follows.

PROPOSITION 1. *Let the ambiguity set $\mathcal{Q}$ be defined as in (11), where the parameters satisfy $0\le \underline{q}_i\le \bar{q}_i\le 1$, $0\le \check{u}_i\le \tau$, and $\varepsilon_i\ge 0, i\in[I]$. Suppose that $\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\le\tau$ for $i\in[I]$. Then the maximization problem over $\mathbb{Q}\in\mathcal{Q}$ in (R–JD), i.e.,*

$$\max_{\mathbb{Q}\in\mathcal{Q}}\ \mathbb{E}_{\mathbb{Q}}\sum_{i\in[I]} d_i p_i \min\{\tilde{u}_i,\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\},$$

is independent of the dispersion upper bounds $\{\varepsilon_i\}_{i\in[I]}$ and evaluates to

$$\sum_{i\in[I]} d_i p_i\left((1-\bar{q}_i)\min\{\check{u}_i,\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}+\bar{q}_i\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\right).$$

Proposition 1 states that with the target constraint on ambulance's maximum allowable response time (i.e., $\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\le\tau$), the inner maximization problem over $\mathbb{Q}\in\mathcal{Q}$ in (R–JD) has a closed-form evaluation. The closed form reveals that the worst-case distribution of $\tilde{u}_i$ in the marginal conditional ambiguity set $\Pi_{\tilde{u}_i|\tilde{u}_i\le\tau}\mathcal{Q}$

is exactly the Dirac measure placing a unit mass on $\check{u}_i$. This closed-form evaluation is fundamental to a tractable reformulation of (R–JD). If we were to adopt a different (and non-trivial) specification for the ambiguity set Q, e.g., a Wasserstein ambiguity set, such a closed form would be unavailable, which would make (R–JD) an intractable mixed-integer nonlinear program given the other two layers of maximization over $(\boldsymbol{d}, \boldsymbol{p})$. In addition, the above closed form is independent of the dispersion upper bounds $\{\varepsilon_i\}_{i\in[I]}$. The phenomenon that the optimal solution of a distributionally robust optimization program finally becomes independent of certain parameters in the ambiguity set has also been observed in Mohajerin Esfahani and Kuhn (2018). We further discuss this issue in Appendix EC.3.

Conditional on the event that at least one responder responds to the alert, Proposition 1 reveals not only the worst-case distribution of the alert response time, but also the corresponding intervention time of CPR. Recall that the intervention of CPR is performed by either the responder or the paramedics on ambulance, whoever arrives earlier. In the worst-case scenario, we have (i) with probability $\bar{q}_i$, the alert response time $\tilde{u}_i$ from the fastest responder in region i exceeds τ, and thus the CPR time is just the ambulance's response time $\sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}$ in that region, since the ambulance will arrive earlier for sure because of constraint (6); and (ii) with probability $1-\bar{q}_i$, the alert response time $\tilde{u}_i$ is exactly equal to $\check{u}_i$, and thus the intervention time of CPR is the minimum of $\check{u}_i$ and $\sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}$. Hence, the closed form in Proposition 1 is actually the demand-weighted intervention times of CPR in the worst-case scenario, with weight $d_i p_i$ being the expected number of OHCA incidents with alert response in region i.

With the above closed form, we can proceed to unveil the second layer of maximization over $\boldsymbol{p} \in \mathcal{P}(\boldsymbol{d})$ in (R–JD). The key is the following proposition that concerns with dualization.

PROPOSITION 2. *Let $\mathcal{P}(\boldsymbol{d})$ be defined as in (13), where the parameters satisfy $0 \le \underline{p}_{(v)} \le \bar{p}_{(v)} \le 1$, $v \in [V]$. Suppose that $\mathcal{P}(\boldsymbol{d})$ is nonempty. Then for any $w_i \in \mathbb{R}$, $i \in [I]$, the following maximization problem over $\boldsymbol{p} \in \mathcal{P}(\boldsymbol{d})$, i.e.,*

$$\max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d})} \sum_{i\in[I]} d_i w_i p_i \tag{14}$$

is equivalent to the minimization problem

$$\begin{aligned} \min_{\underline{\boldsymbol{\lambda}},\bar{\boldsymbol{\lambda}},\boldsymbol{\mu}} \quad & \sum_{v\in[V]} \sum_{i\in I_{(v)}} d_i(\bar{p}_{(v)}\bar{\lambda}_{(v)} - \underline{p}_{(v)}\underline{\lambda}_{(v)}) + \sum_{i\in[I]} \mu_i \\ \text{s.t.} \quad & d_i \sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}} (\bar{\lambda}_{(v)} - \underline{\lambda}_{(v)}) + \mu_i \le d_i w_i, \ i \in [I], \\ & \underline{\boldsymbol{\lambda}} \in \mathbb{R}^V_+, \ \bar{\boldsymbol{\lambda}} \in \mathbb{R}^V_+, \ \boldsymbol{\mu} \in \mathbb{R}^I_+. \end{aligned} \tag{15}$$

Under a mild condition on the nonemptiness of $\mathcal{P}(\boldsymbol{d})$, Proposition 2 states that the maximization problem (14) over $\boldsymbol{p} \in \mathcal{P}(\boldsymbol{d})$ can be equivalently dualized as a linear minimization program. This nonemptiness condition is easy to satisfy as long as the uncertainty intervals $[\underline{p}_{(v)}, \bar{p}_{(v)}]$, $v \in [V]$, are properly estimated from

the historical data, and it is guaranteed to be satisfied if the V clusters are non-overlapping (see Appendix EC.2.4). In view of the closed form in Proposition 1, we can apply Proposition 2 to (R–JD) by setting

$$w_i \triangleq (1-\bar{q}_i)\Big(\min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} - \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\Big),\ i\in[I]. \tag{16}$$

Conditional on alert response from responders, the above w_i represents the intervention time improvement for CPR in region i under the worse-case scenario. The value of w_i is upper bounded by 0, and it becomes smaller if the probability of $\tilde{u}_i \le \tau$ (i.e., $1-\bar{q}_i$) becomes larger, or the alert response time $[\tilde{u}_i|\tilde{u}_i \le \tau]$ becomes shorter (i.e., a more negative $\check{u}_i - \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}$). When the V clusters are non-overlapping, the above problem structure can be exploited and more insights can be obtained, as summarized in the following proposition.

PROPOSITION 3. *Consider $\mathcal{P}(\boldsymbol{d})$ in (13) such that $0 \le \underline{p}_{(v)} \le \bar{p}_{(v)} \le 1$ for $v\in[V]$. Suppose that the clusters $\{I_{(v)}\}_{v\in[V]}$ are non-overlapping, i.e., $I_{(v)} \cap I_{(v')} = \emptyset$ for all $v \ne v'$, and that the coefficients satisfy $d_i w_i \le 0$, $i\in[I]$. Then the maximization problem (14) over $\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d})$ is equivalent to the following minimization problem*

$$\begin{aligned} \min_{\boldsymbol{\lambda},\boldsymbol{\mu}} \quad & \sum_{v\in[V]}\sum_{i\in I_{(v)}} d_i \underline{p}_{(v)} \lambda_{(v)} + \sum_{i\in[I]} \mu_i \\ s.t. \quad & d_i\lambda_{(v)} + \mu_i \ge d_i w_i,\ i\in[I], \\ & \boldsymbol{\lambda}\in\mathbb{R}^V,\ \boldsymbol{\mu}\in\mathbb{R}^I_+. \end{aligned}$$

Moreover, for each cluster $v\in[V]$, let $i_1^v, i_2^v, \ldots, i_{|I_{(v)}|}^v$ be a permutation of the regions in the index set $I_{(v)}$ so that $d_{i_1^v} w_{i_1^v} \le d_{i_2^v} w_{i_2^v} \le \cdots \le d_{i_{|I_{(v)}|}^v} w_{i_{|I_{(v)}|}^v}$. Then there exists a worst-case alert response probability vector $\boldsymbol{p}^\star \in \mathcal{P}(\boldsymbol{d})$ attaining the maximum of (14) which satisfies $p^\star_{i_1^v} \le p^\star_{i_2^v} \le \cdots \le p^\star_{i_{|I_{(v)}|}^v}$ and $\sum_{i\in I_{(v)}} d_i p_i^\star / \sum_{i\in I_{(v)}} d_i = \underline{p}_{(v)}$ for each cluster $v\in[V]$.

By exploiting the mutual exclusivity of the V clusters and the nonpositivity of $\{d_i w_i\}_{i\in[I]}$, the above dual program involves fewer decision variables and simpler constraints compared with (15). Moreover, it is independent of the upper bounds $\{\bar{p}_{(v)}\}_{v\in[V]}$ of the uncertainty intervals. This is because (i) $\sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}} = 1$ for all $i\in[I]$ by the exclusivity, and (ii) the above conditions ensure that there exists an optimal solution $\boldsymbol{p}^\star$ to (14) which attains the lowest possible average response probability in each cluster, and thus the inequality constraints in $\mathcal{P}(\boldsymbol{d})$ can be replaced by equality constraints, i.e., $\sum_{i\in I_{(v)}} d_i p_i / \sum_{i\in I_{(v)}} d_i = \underline{p}_{(v)}$ for each $v\in[V]$. When some clusters are overlapping, however, the lower bounds of uncertainty intervals $\{\underline{p}_{(v)}\}_{v\in[V]}$ may not be simultaneously attainable by a feasible $\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d})$. As such, the upper bounds of the uncertainty intervals $\{\bar{p}_{(v)}\}_{v\in[V]}$ also play a role in deciding $\boldsymbol{p}^\star$ and they appear in the dual program (15).

Apart from the above, Proposition 3 also provides insight into the worst-case alert response probabilities $\boldsymbol{p}^\star \in \mathcal{P}(\boldsymbol{d})$. Since the negative w_i represents the intervention time improvement of CPR in region i for each alert response, the magnitude of $d_i w_i$ represents the overall effectiveness of the responder program

conditional on alert response. Proposition 3 reveals that the worst-case $\boldsymbol{p}^\star$ corresponds to the scenario that a lower alert response probability is assigned to a region that benefits more from alert responses. In this way, the effectiveness of the responder program is compromised to the utmost extent. As an attempt to hedge against this worst case, the robust deployment program (R–JD) tends to adjust the values of the deployment variables x_{ik} to level the values of $\{w_i\}_{i\in[I]}$ in (16), which is equivalent to balancing the responsiveness of ambulances and responders in the I regions.

With Propositions 1 and 2, the maximization problem over $\mathbb{Q}\in\mathcal{Q}$ and $\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d})$ in (R–JD) can be reformulated as a single-layer minimization program that is linear in the dual variables $(\underline{\boldsymbol{\lambda}},\bar{\boldsymbol{\lambda}},\boldsymbol{\mu})$, but still nonlinear in the deployment variables $(\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}})$. With some more efforts, we can obtain a MILP reformulation for (R–JD) as follows.

THEOREM 1. *The robust joint deployment problem (R–JD) with ambiguity set $\mathcal{Q}$ defined by (11), uncertainty set $\mathcal{D}$ defined by (12), and uncertainty set $\mathcal{P}(\boldsymbol{d})$ defined by (13) is equivalent to the following MILP, where $M\triangleq\max\{\max_{i\in[I],j\in[J]}t^{\mathrm{D}}_{ij},\max_{i\in[I],k\in[K]}t^{\mathrm{A}}_{ik}\}$ is a large enough positive constant:*

$$
\begin{aligned}
\min_{\substack{\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}},\\ \boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\nu},\underline{\boldsymbol{\lambda}}^{(b)},\bar{\boldsymbol{\lambda}}^{(b)},\boldsymbol{\mu}^{(b)},\zeta}}\quad & \zeta\\
\text{s.t.}\quad & \zeta\geq\sum_{i\in[I]}d_i^{(b)}\big(\rho_1\phi_i+(1+\rho_2)\sum_{k\in[K]}x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}\big)+\sum_{v\in[V]}\sum_{i\in I_{(v)}}d_i^{(b)}(\bar{p}_{(v)}\bar{\lambda}^{(b)}_{(v)}-\underline{p}_{(v)}\underline{\lambda}^{(b)}_{(v)})+\sum_{i\in[I]}\mu_i^{(b)},\ b\in[B],\\
& d_i^{(b)}\sum_{v\in[V]}\mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}^{(b)}_{(v)}-\underline{\lambda}^{(b)}_{(v)})+\mu_i^{(b)}\geq d_i^{(b)}(1-\bar{q}_i)(\psi_i-\sum_{k\in[K]}x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}),\ i\in I_{(v)},v\in[V],b\in[B],\\
& \underline{\boldsymbol{\lambda}}^{(b)}\in\mathbb{R}^V_+,\ \bar{\boldsymbol{\lambda}}^{(b)}\in\mathbb{R}^V_+,\ \boldsymbol{\mu}^{(b)}\in\mathbb{R}^I_+,\ b\in[B],\\
& \phi_i\geq\textstyle\sum_{j\in[J]}x^{\mathrm{D}}_{ij}t^{\mathrm{D}}_{ij}-M\nu_{i1},\ \phi_i\geq\sum_{k\in[K]}x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}-M\nu_{i2},\ i\in[I],\\
& \psi_i\geq\check{u}_i-M\nu_{i3},\ \psi_i\geq\textstyle\sum_{k\in[K]}x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}-M\nu_{i4},\ i\in[I],\\
& \nu_{i1}+\nu_{i2}=1,\ \nu_{i3}+\nu_{i4}=1,\ i\in[I],\\
& (\nu_{i1},\nu_{i2},\nu_{i3},\nu_{i4})\in\{0,1\}^4,\ i\in[I],\\
& \boldsymbol{y}^{\mathrm{D}}\in\mathcal{Y}^{\mathrm{D}},\ \boldsymbol{y}^{\mathrm{A}}\in\mathcal{Y}^{\mathrm{A}},\\
& \boldsymbol{x}^{\mathrm{D}}\in\mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}),\ \boldsymbol{x}^{\mathrm{A}}\in\mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}).
\end{aligned}
\tag{17}
$$

In the above MILP, ζ is an epigraph decision variable used to help enumerate demand scenarios in the demand uncertainty set $\mathcal{D}$, $(\phi_i,\nu_{i1},\nu_{i2})$ are auxiliary variables used to linearize the term $\min\{\sum_{j\in[J]}x^{\mathrm{D}}_{ij}t^{\mathrm{D}}_{ij},\sum_{k\in[K]}x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}\}$, and $(\psi_i,\nu_{i3},\nu_{i4})$ are auxiliary variables used to linearize the term $\min\{\check{u}_i,\sum_{k\in[K]}x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}\}$. Dual variables $\underline{\boldsymbol{\lambda}}^{(b)},\bar{\boldsymbol{\lambda}}^{(b)}$ and $\boldsymbol{\mu}^{(b)}$ are defined for each demand scenario $b\in[B]$. We note that if the V clusters are non-overlapping, the above MILP can be further simplified using Proposition 3.

As argued, we recommend a large number B for the demand scenarios to capture the uncertainty in OHCA demand rate. Due to the replication of the dual variables $\underline{\boldsymbol{\lambda}}^{(b)},\bar{\boldsymbol{\lambda}}^{(b)},\boldsymbol{\mu}^{(b)}$, $b\in[B]$, and the associated constraints

in the MILP (17), however, it quickly becomes computationally intractable to solve (R–JD) using solvers as B increases. This can be overcome through a scenario generation algorithm that scales gracefully with B, which is the focus of the next section. Before moving to the algorithm, we make the following observation on the solution structure of (R–JD), which will be useful in the algorithmic design.

PROPOSITION 4. *Let* $(\boldsymbol{y}^{\mathrm{D}}_\star, \boldsymbol{y}^{\mathrm{A}}_\star)$ *be an optimal deployment decision of drone/ambulance to* (R–JD)*, and let* $\boldsymbol{d}^{(b^\star)} \in \mathcal{D}$ *and* $\boldsymbol{p}^\star \in \mathcal{P}(\boldsymbol{d}^{(b^\star)})$ *be the corresponding worst-case demand scenario and alert response probability vector, respectively. Given* $(\boldsymbol{y}^{\mathrm{D}}_\star, \boldsymbol{y}^{\mathrm{A}}_\star)$*, denote the nearest drone base and ambulance base to region* i *by* $j_\star(i) \triangleq \arg\min_j \left\{t^{\mathrm{D}}_{ij}, j \in [J] : y^{\mathrm{D}}_{\star j} = 1\right\}$ *and* $k_\star(i) \triangleq \arg\min_k \left\{t^{\mathrm{A}}_{ik}, k \in [K] : y^{\mathrm{A}}_{\star k} = 1\right\}$*, respectively. Given* $(\boldsymbol{d}^{(b^\star)}, \boldsymbol{p}^\star)$*, denote the subset of regions with positive demand and alert response by* $\mathcal{M}^{\mathrm{d}} \triangleq \{i \in [I] : d_i^{(b^\star)} > 0\}$ *and* $\mathcal{M}^{\mathrm{a}} \triangleq \{i \in [I] : d_i^{(b^\star)} p_i^\star > 0\}$*, respectively. If* $(\boldsymbol{y}^{\mathrm{D}}_\star, \boldsymbol{y}^{\mathrm{A}}_\star, \boldsymbol{x}^{\mathrm{D}}_\star, \boldsymbol{x}^{\mathrm{A}}_\star, \boldsymbol{\phi}_\star, \boldsymbol{\psi}_\star)$ *is optimal to the MILP reformulation* (17) *of* (R–JD)*, then we have*

(i) *In region* $i \in \mathcal{M}^{\mathrm{d}}$*, the optimal assignment* $x^{\mathrm{D}}_{\star ij}$*,* $x^{\mathrm{A}}_{\star ik}$ *and the optimal* $\phi_{\star i}$ *are unique. In particular,* $\phi_{\star i} = \min\left\{t^{\mathrm{D}}_{ij_\star(i)}, t^{\mathrm{A}}_{ik_\star(i)}\right\}$*, and* $x^{\mathrm{D}}_{\star ij} = x^{\mathrm{A}}_{\star ik} = 1$ *when* $j = j_\star(i)$ *and* $k = k_\star(i)$*, and zero otherwise.*

(ii) *In region* $i \in \mathcal{M}^{\mathrm{a}}$*, the optimal* $\psi_{\star i}$ *is unique and equals* $\min\left\{\breve{u}_i, t^{\mathrm{A}}_{ik_\star(i)}\right\}$.

(iii) *The optimal* $x^{\mathrm{D}}_{\star ij}, x^{\mathrm{A}}_{\star ik}, \phi_{\star i}$ *in regions* $i \in [I]\backslash\mathcal{M}^{\mathrm{d}}$ *and the optimal* $\psi_{\star i}$ *in* $i \in [I]\backslash\mathcal{M}^{\mathrm{a}}$ *are not unique.*

The intuition for the non-uniqueness revealed in statement (iii) above is as follows. Letting $(\boldsymbol{d}^{(b^\star)}, \boldsymbol{p}^\star)$ be the worst-case scenario, it can be shown that the optimal value of (17) is $\sum_{i\in[I]} d_i^{(b^\star)}[p_i^\star(1-\bar{q}_i)\psi_{\star i} + \rho_1\phi_{\star i} + (1 - p_i^\star\bar{q}_i + \rho_2)\sum_{k\in[K]} x^{\mathrm{A}}_{\star ik} t^{\mathrm{A}}_{ik}]$. When $d_i^{(b^\star)} = 0$ for region i, the values of $\psi_{\star i}$, $\phi_{\star i}$, $x^{\mathrm{A}}_{\star ik}$ and $x^{\mathrm{D}}_{\star ij}$ do not affect the optimal value as long as they are feasible. Similarly, if $p_i^\star = 0$ for region i, then $\psi_{\star i}$ disappears from the optimal value expression. The possible non-uniqueness of the optimal solution becomes a problem when we implement a scenario generation algorithm, which will be overcome by introducing a rectification procedure in the next section.

6 The Row-and-Column Generation Algorithm

In this section, we develop a scenario generation algorithm that can efficiently solve the MILP (17) to optimality when the number of demand scenarios is large. The scenario generation algorithm iterates between a master problem and a subproblem that generates a potential worst-case demand scenario in the demand uncertainty set $\mathcal{D}$ until convergence. This algorithm is essentially a row-and-column generation algorithm because it simultaneously adds decision variables and constraints into the master problem in each iteration.

6.1 The Master Problem

The master problem is a relaxation of (17) where only a subset of the demand scenarios are considered. Let $\mathcal{S} \subset [B]$ be an index subset of all scenarios and $\mathcal{D}_\mathcal{S} \triangleq \{\boldsymbol{d}^{(s)}, s \in \mathcal{S}\}$ a subset of $\mathcal{D}$. The master problem with

scenario subset $\mathcal{S}$ is given by

$$
\begin{aligned}
\min_{\substack{\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}},\\ \boldsymbol{\nu},\boldsymbol{\phi},\boldsymbol{\psi},\underline{\boldsymbol{\lambda}}^{(s)},\bar{\boldsymbol{\lambda}}^{(s)},\boldsymbol{\mu}^{(s)},\zeta}} \quad & \zeta \\
\text{s.t.} \quad & \zeta \geq \sum_{i\in[I]} d_i^{(s)}\big(\rho_1\phi_i + (1+\rho_2)\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\big) + \sum_{v\in[V]}\sum_{i\in I_{(v)}} d_i^{(s)}(\bar{p}_{(v)}\bar{\lambda}_{(v)}^{(s)} - \underline{p}_{(v)}\underline{\lambda}_{(v)}^{(s)}) + \sum_{i\in[I]}\mu_i^{(s)},\ s\in\mathcal{S} \\
& d_i^{(s)}\sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}_{(v)}^{(s)} - \underline{\lambda}_{(v)}^{(s)}) + \mu_i^{(s)} \geq d_i^{(s)}(1-\bar{q}_i)(\psi_i - \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}),\ i\in I_{(v)}, v\in[V], s\in\mathcal{S} \\
& \underline{\boldsymbol{\lambda}}^{(s)} \in \mathbb{R}_+^V,\ \bar{\boldsymbol{\lambda}}^{(s)} \in \mathbb{R}_+^V,\ \boldsymbol{\mu}^{(s)} \in \mathbb{R}_+^I,\ s\in\mathcal{S}, \\
& \phi_i \geq \textstyle\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}} - M\nu_{i1},\ \phi_i \geq \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} - M\nu_{i2},\ i\in[I], \\
& \psi_i \geq \breve{u}_i - M\nu_{i3},\ \psi_i \geq \textstyle\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} - M\nu_{i4},\ i\in[I], \\
& \nu_{i1}+\nu_{i2}=1,\ \nu_{i3}+\nu_{i4}=1,\ i\in[I], \\
& (\nu_{i1},\nu_{i2},\nu_{i3},\nu_{i4}) \in \{0,1\}^4,\ i\in[I], \\
& \boldsymbol{y}^{\mathrm{D}} \in \mathcal{Y}^{\mathrm{D}},\ \boldsymbol{y}^{\mathrm{A}} \in \mathcal{Y}^{\mathrm{A}}, \\
& \boldsymbol{x}^{\mathrm{D}} \in \mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}),\ \boldsymbol{x}^{\mathrm{A}} \in \mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}).
\end{aligned}
\tag{MP}
$$

Let $Z_{\mathcal{S}}^{\mathrm{MP}}$ and $\boldsymbol{O}_{\mathcal{S}} \triangleq (\boldsymbol{y}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{y}_{\mathcal{S}}^{\mathrm{A}}, \boldsymbol{x}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{x}_{\mathcal{S}}^{\mathrm{A}}, \boldsymbol{\phi}_{\mathcal{S}}, \boldsymbol{\psi}_{\mathcal{S}})$ be the optimal value and solution of the above master problem. Since (MP) is a relaxation of (17), $Z_{\mathcal{S}}^{\mathrm{MP}}$ produces a lower bound for the optimal value of (17). This lower bound can be tightened by adding more scenario indices into $\mathcal{S}$. The idea is to identify a demand scenario in $\mathcal{D}$ that leads to the worst performance under the incumbent deployment $(\boldsymbol{y}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{y}_{\mathcal{S}}^{\mathrm{A}})$. This can be done by solving a series of subproblems, as detailed next.

6.2 The Subproblem

The subproblem takes the incumbent solution $\boldsymbol{O}_{\mathcal{S}}$ obtained from (MP) as input. The assignment/auxiliary variables $(\boldsymbol{x}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{x}_{\mathcal{S}}^{\mathrm{A}}, \boldsymbol{\phi}_{\mathcal{S}}, \boldsymbol{\psi}_{\mathcal{S}})$ are not unique by Proposition 4(iii). To fix this, we rectify $(\boldsymbol{x}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{x}_{\mathcal{S}}^{\mathrm{A}}, \boldsymbol{\phi}_{\mathcal{S}}, \boldsymbol{\psi}_{\mathcal{S}})$ by enforcing the equalities in (i) and (ii) of Proposition 4 for all $i\in[I]$. The main consideration is to minimize the response times under the incumbent deployment $(\boldsymbol{y}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{y}_{\mathcal{S}}^{\mathrm{A}})$ for each region $i\in[I]$. Denote the rectified incumbent solution as $\dot{\boldsymbol{O}}_{\mathcal{S}} \triangleq (\boldsymbol{y}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{y}_{\mathcal{S}}^{\mathrm{A}}, \dot{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{D}}, \dot{\boldsymbol{x}}_{\mathcal{S}}^{\mathrm{A}}, \dot{\boldsymbol{\phi}}_{\mathcal{S}}, \dot{\boldsymbol{\psi}}_{\mathcal{S}})$, which is still optimal to the master problem (MP). Then for each demand scenario $\boldsymbol{d}^{(b)} \in \mathcal{D}$, the performance of the incumbent deployment $(\boldsymbol{y}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{y}_{\mathcal{S}}^{\mathrm{A}})$ under this demand scenario can be obtained by solving the subproblem given by

$$
\begin{aligned}
\min_{\underline{\boldsymbol{\lambda}}^{(b)},\bar{\boldsymbol{\lambda}}^{(b)},\boldsymbol{\mu}^{(b)}} \quad & \sum_{i\in[I]} d_i^{(b)}\big(\rho_1\dot{\phi}_{\mathcal{S},i} + (1+\rho_2)\sum_{k\in[K]} \dot{x}_{\mathcal{S},ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\big) + \mu_i^{(b)} + \sum_{v\in[V]}\sum_{i\in I_{(v)}} d_i^{(b)}(\bar{p}_{(v)}\bar{\lambda}_{(v)}^{(b)} - \underline{p}_{(v)}\underline{\lambda}_{(v)}^{(b)}) \\
\text{s.t.} \quad & d_i^{(b)}\sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}_{(v)}^{(b)} - \underline{\lambda}_{(v)}^{(b)}) + \mu_i^{(b)} \geq d_i^{(b)}(1-\bar{q}_i)(\psi_{\mathcal{S},i} - \sum_{k\in[K]} \dot{x}_{\mathcal{S},ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}),\ i\in I_{(v)},\ v\in[V], \\
& \underline{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}_+^V,\ \bar{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}_+^V,\ \boldsymbol{\mu}^{(b)} \in \mathbb{R}_+^I.
\end{aligned}
\tag{SP}
$$

Let $Z_{\mathcal{S},b}^{\mathrm{SP}}$ be the optimal value of (SP). Then the worst-case demand scenario with respect to $(\boldsymbol{y}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{y}_{\mathcal{S}}^{\mathrm{A}})$ is identified as $b_{\mathcal{S}} \triangleq \arg\max_{b\in[B]} \{Z_{\mathcal{S},b}^{\mathrm{SP}}\}$. Since $(\boldsymbol{y}_{\mathcal{S}}^{\mathrm{D}}, \boldsymbol{y}_{\mathcal{S}}^{\mathrm{A}})$ is feasible, the worst-case objective value of (SP), $Z_{\mathcal{S},b_{\mathcal{S}}}^{\mathrm{SP}}$, provides an upper bound for the optimal value of (17).

If we directly solve (SP) using the incumbent solution $\boldsymbol{O}_{\mathcal{S}}$ without rectification, the identified scenario $b_{\mathcal{S}}$ may not be the worst with respect to the incumbent deployment $(\boldsymbol{y}^{\mathrm{D}}_{\mathcal{S}}, \boldsymbol{y}^{\mathrm{A}}_{\mathcal{S}})$. This is because the assignment/auxiliary variables $(\boldsymbol{x}^{\mathrm{D}}_{\mathcal{S}}, \boldsymbol{x}^{\mathrm{A}}_{\mathcal{S}}, \boldsymbol{\phi}_{\mathcal{S}}, \boldsymbol{\psi}_{\mathcal{S}})$ may not reflect the best achievable response time in some regions, as revealed in Proposition 4(iii). In practice, we find that including this rectification procedure leads to much faster convergence of the algorithm, which is also theoretically supported by our convergence analysis in Section 6.4.

6.3 The Iteration Procedure and Termination Criterion

We start the iteration procedure by initializing the scenario index subset $\mathcal{S}$ as a singleton randomly drawn from $[B]$. In each iteration, we solve the master problem (MP) to obtain the incumbent solution $\boldsymbol{O}_{\mathcal{S}}$ and the corresponding optimal value $Z^{\mathrm{MP}}_{\mathcal{S}}$. We then rectify the incumbent solution $\boldsymbol{O}_{\mathcal{S}}$ to obtain $\dot{\boldsymbol{O}}_{\mathcal{S}}$, and solve the series of subproblems (SP) indexed by $b \in [B]$ to obtain the worst-case scenario $b_{\mathcal{S}}$ and objective value $Z^{\mathrm{SP}}_{\mathcal{S},b_{\mathcal{S}}}$. If $Z^{\mathrm{MP}}_{\mathcal{S}} < Z^{\mathrm{SP}}_{\mathcal{S},b_{\mathcal{S}}}$, the incumbent solution $\boldsymbol{O}_{\mathcal{S}}$ is not yet optimal to the MILP (17), and we update the scenario index subset by $\mathcal{S} \leftarrow \mathcal{S} \cup \{b_{\mathcal{S}}\}$. This amounts to adding the dual variables $\underline{\boldsymbol{\lambda}}^{(b_{\mathcal{S}})}, \bar{\boldsymbol{\lambda}}^{(b_{\mathcal{S}})}, \boldsymbol{\mu}^{(b_{\mathcal{S}})}$ and the associated constraints in the form

$$\zeta \geq \sum_{i\in[I]} d_i^{(b_{\mathcal{S}})}\big(\rho_1\phi_i + (1+\rho_2)\sum_{k\in[K]} x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}\big) + \sum_{v\in[V]}\sum_{i\in I_{(v)}} d_i^{(b_{\mathcal{S}})}(\bar{p}_{(v)}\bar{\lambda}^{(b_{\mathcal{S}})}_{(v)} - \underline{p}_{(v)}\underline{\lambda}^{(b_{\mathcal{S}})}_{(v)}) + \sum_{i\in[I]}\mu_i^{(b_{\mathcal{S}})};$$

$$d_i^{(b_{\mathcal{S}})}\sum_{v\in[V]}\mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}^{(b_{\mathcal{S}})}_{(v)} - \underline{\lambda}^{(b_{\mathcal{S}})}_{(v)}) + \mu_i^{(b_{\mathcal{S}})} \geq d_i^{(b_{\mathcal{S}})}(1-\bar{q}_i)(\psi_i - \sum_{k\in[K]} x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}),\ i\in I_{(v)}, v\in[V]$$

into the master problem (MP). We then start the next round of iteration to solve the series of (MP) and (SP). In the case that $Z^{\mathrm{MP}}_{\mathcal{S}} = Z^{\mathrm{SP}}_{\mathcal{S},b_{\mathcal{S}}}$, we stop the iteration procedure, and the incumbent solution $\boldsymbol{O}_{\mathcal{S}}$ is optimal to the MILP (17). The above termination criterion is guaranteed to be met in finite steps, because the number of scenarios in $\mathcal{D}$ is finite. In practice, the scenario index subset $\mathcal{S}$ in the final round of iteration is usually much smaller than $[B]$. This is because the set of worst-case demand scenarios (with respect to all feasible deployment decisions) is usually a small subset of the demand uncertainty set $\mathcal{D}$, as is further illustrated in the next section.

6.4 Convergence Analysis

Denote $\boldsymbol{O} \triangleq (\boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}}, \boldsymbol{x}^{\mathrm{D}}, \boldsymbol{x}^{\mathrm{A}}, \boldsymbol{\phi}, \boldsymbol{\psi})$ as the collection of deployment, assignment and auxiliary decision variables in MILP (17). By Theorem 1, the feasible region for $\boldsymbol{O}$ is $O \triangleq \big\{(\boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}}, \boldsymbol{x}^{\mathrm{D}}, \boldsymbol{x}^{\mathrm{A}}, \boldsymbol{\phi}, \boldsymbol{\psi}) \in \mathcal{Y}^{\mathrm{D}} \times \mathcal{Y}^{\mathrm{A}} \times \mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}) \times \mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}) \times \mathbb{R}^I \times \mathbb{R}^I : \phi_i \geq \min\{\sum_{j\in[J]} x^{\mathrm{D}}_{ij}t^{\mathrm{D}}_{ij}, \sum_{k\in[K]} x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}\}, \psi_i \geq \min\{\check{u}_i, \sum_{k\in[K]} x^{\mathrm{A}}_{ik}t^{\mathrm{A}}_{ik}\}, i\in[I]\big\}$. Further, let $G(\boldsymbol{d}, \boldsymbol{O})$ be the optimal value of the subproblem (SP) under demand scenario $\boldsymbol{d} \in \mathcal{D}$ and incumbent solution $\boldsymbol{O} \in O$. Then we have

- The MILP (17) can be equivalently expressed as $\min_{\boldsymbol{O}\in O}\max_{\boldsymbol{d}\in\mathcal{D}} G(\boldsymbol{d}, \boldsymbol{O})$;
- Finding the optimal solution $\boldsymbol{O}_{\mathcal{S}}$ to the master problem (MP) is equivalent to the problem $\min_{\boldsymbol{O}\in O}\max_{\boldsymbol{d}\in\mathcal{D}_{\mathcal{S}}} G(\boldsymbol{d}, \boldsymbol{O})$ where recall that $\mathcal{D}_{\mathcal{S}} = \{\boldsymbol{d}^{(s)}, s\in\mathcal{S}\}$;

- Identification of the worst-case scenario with respect to the incumbent solution $\boldsymbol{O}_S$ is equivalent to the problem $\max_{\boldsymbol{d}\in\mathcal{D}} G(\boldsymbol{d},\boldsymbol{O}_S)$.

With the above representation, it is easy to see the following inequalities:

$$\max_{\boldsymbol{d}\in\mathcal{D}_S} G(\boldsymbol{d},\boldsymbol{O}_S) = \min_{\boldsymbol{O}\in O}\max_{\boldsymbol{d}\in\mathcal{D}_S} G(\boldsymbol{d},\boldsymbol{O}) \le \min_{\boldsymbol{O}\in O}\max_{\boldsymbol{d}\in\mathcal{D}} G(\boldsymbol{d},\boldsymbol{O}) \le \max_{\boldsymbol{d}\in\mathcal{D}} G(\boldsymbol{d},\boldsymbol{O}_S),$$

where the first equality is by the optimality of $\boldsymbol{O}_S$ in the master problem, and the solution to the rightmost term, denoted as $\boldsymbol{d}(\boldsymbol{O}_S) \triangleq \arg\max_{\boldsymbol{d}\in\mathcal{D}} G(\boldsymbol{d},\boldsymbol{O}_S)$, is the worst-case demand scenario identified from the subproblem. The leftmost and the rightmost terms serve as respective lower and upper bounds for the optimal value of the MILP (17). Our algorithm terminates when the two bounds coincide. Otherwise, we enlarge the scenario subset by adding $\boldsymbol{d}(\boldsymbol{O}_S)$ to $\mathcal{D}_S$ and then proceed to the next iteration. Let $\mathcal{D}^\star \triangleq \{\boldsymbol{d}(\boldsymbol{O}) : \boldsymbol{O}\in O\}$ be the set of (all possible) worst-case demand scenarios. The following proposition states that the size of this set turns out to be an upper bound on the number of iterations of our algorithm.

PROPOSITION 5. *The scenario generation algorithm terminates in not greater than $S+1$ iterations, where S is the cardinality of $\mathcal{D}^\star$.*

Through comprehensive simulations, we find that our algorithm normally converges in just a few iterations, which implies that $\mathcal{D}^\star$ is a small set. That is, the set of worst-case demand scenarios is sparsely supported on $\mathcal{D}$. Propositions 4 and 5 together reveal the benefit of rectifying the incumbent solution $\boldsymbol{O}_S$. The rectification procedure essentially reduces the feasible set O by adding more constraints on the decision variables $(\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}},\boldsymbol{\phi},\boldsymbol{\psi})$, but with no loss of optimality. With a smaller O, the size of $\mathcal{D}^\star$, defined as the collection of $\boldsymbol{d}(\boldsymbol{O})$, is reduced. From another perspective, $\mathcal{D}^\star$ defined above is the collection of all worse-case scenarios for every feasible $(\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}},\boldsymbol{\phi},\boldsymbol{\psi})$. Alternatively, we can define $\dot{\mathcal{D}}^\star$ as the collection of all worse-case scenarios for every feasible deployment $(\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}})$, which is obviously a subset of $\mathcal{D}^\star$. Our rectification ensures that $\mathcal{D}^\star = \dot{\mathcal{D}}^\star$.

7 Data Analysis for the Joint Deployment Study

We demonstrate the joint deployment model on the main island of Singapore, excluding the sparsely populated area in the west, the Changi Airport in the east, the reservoir and military area (as identified in the right panel of Figure 1). The planning area is discretized into $I = 131$ square demand regions each of size 2 km $\times$ 2 km, as shown in the left panel of Figure 1. This grid size is granular enough to capture the geographical nonhomogeneity of OHCA demand and responder response, and to accurately calculate ambulance/drone response times while keeping the deployment program size manageable. All public hospitals, polyclinics and fire stations in the above planning area are considered as candidate base locations for both ambulances and drones (Goverment 2020, SCDF 2019). Due to the extra portability of drone depots, we assume they can also be located at police centers (Goverment 2021). In total, we identify $J = 56$ and $K = 91$ candidate base locations for ambulances and drones, respectively. Their locations on map are plotted in Figure 1.

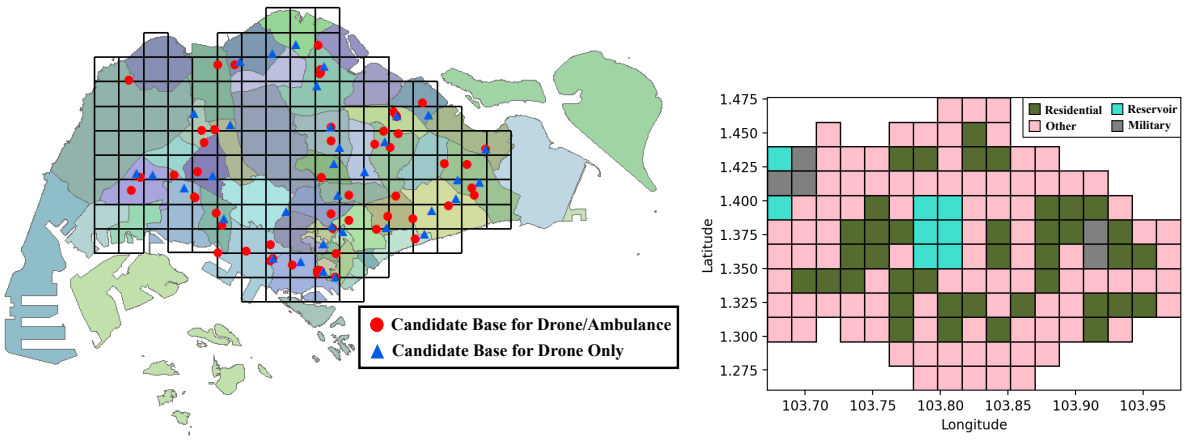


Figure 1 Discretization of the main island of Singapore using square regions (left) and the identification of region types based on Google Map information (right).

The ambulance travel time t_{ik}^{A} to each demand center is obtained from the Google Maps Distance Matrix API (Google Map 2022). On the other hand, the drone travel time t_{ij}^{D} is calculated based on the Euclidean distance to a region center divided by the drone speed 27.8 m/s, plus a fixed duration of 10 seconds to account for takeoff and landing (Boutilier and Chan 2022).

Our dataset consists of 292 OHCA incidents in Singapore in October 2017. For each recorded OHCA incident, the dataset includes the report time and the location of the incident, the ambulance arrival time, the number of responders responded to the alert, and the arrival time of the first responder if the arrival is earlier than the ambulance. The ambulance (resp. alert) response time in this incident can be calculated as the difference between the ambulance (resp. the fastest responder) arrival time and the report time.

Visualization of the data is provided in Figure 2. First, we count the total number of OHCA incidents in October 2017 within each demand region, as shown in the left panel. It can be seen that residential regions tend to have more OHCA incidents. Next, we count the total number of incidents with alert response in each region and normalize it by the incident number in that region to get the historical alert response probabilities, as shown in the middle panel. We see that the alert response probabilities exhibit large inter-region variations, which is mainly caused by the small sample size in each region. In some regions, these probabilities are not even computable due to zero OHCA occurrence in the available data. The above observation justifies our construction of the uncertainty set $\mathcal{P}$ that involves region aggregation. On average, the alert response probability in the whole planning area is circa 40%. Out of the 40% incidents, the right censoring proportion (by ambulance arrival) is about 49%. Finally, for each region we compute the proportion of incidents where responder(s) responded and arrived earlier than the ambulance, as shown in the right panel. The small numbers in this figure clearly shows that the current first responder program is not very effective. In the planning area, this proportion is about 21%.

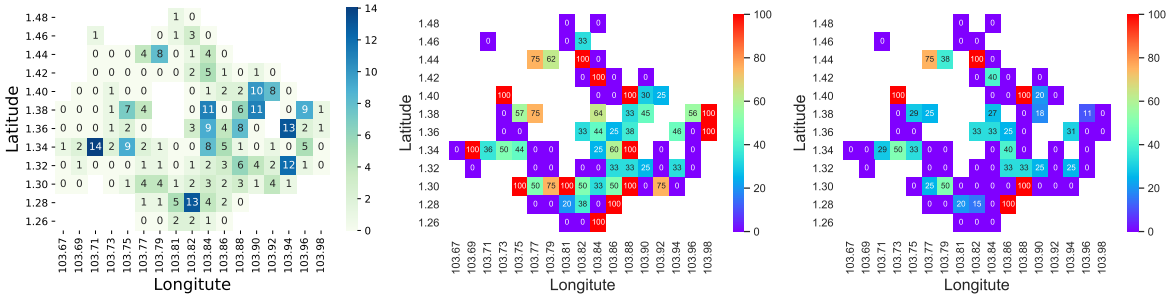


Figure 2 Visualization of the October 2017 data: Heatmaps for the historical OHCA occurrences (left), alert response probability (middle), and proportion of incidents where responder(s) arrived first (right).

Based on this dataset, we can estimate the alert response time distribution ambiguity set Q, alert response probability uncertainty set $\mathcal{P}$, and demand uncertainty set $\mathcal{D}$. The detailed estimation procedure and results are provided in Appendices EC.4 and EC.5.

8 Numerical Experiments for the Joint Deployment Study

With results from the previous section, the robust deployment model (R–JD) can be solved when the numbers of ambulances and drones are given. To evaluate the performance of (R–JD), we generate simulated OHCA incidents under different settings as test data and introduce two benchmark models in Section 8.1. Model comparison results are presented in Section 8.2 to demonstrate the values of our proposed model. In Section 8.3, we conduct sensitivity analysis on various factors to explore their impacts on survival outcome.

8.1 Test Data Generation and Benchmarking Models

In this subsection, we generate test data that would be used to evaluate the performance of a deployment decision produced from either our proposed model (R–JD) or the benchmarks. We consider five settings in test data generation to cover a wide range of possible future events from the one that is similar to the historical data to the one that deviates much from the historical data. To this end, for a fixed $h \in \{1, 1.5, 2, 2.5, 3\}$, we generate a bootstrap dataset from historical data, and use it to: (i) estimate the density for OHCA locations using KDE with bandwidth $h \times \omega^*$, (ii) estimate the alert response probabilities in the residential area and the other area, and (iii) fit two Weibull distributions to the alert response times in these two types of area. With this fitted model, we generate a set of test data of the same size as the historical data ($n = 292$). The above procedure is repeated 100 times to generate 100 sets of test data for each h, which we shall call the bw-h test data in the sequel. The bw-1 setting aims to mimic the mechanism that generates the October 2017 data by accounting for estimation uncertainty, while the bw-h settings with $h > 1$ simulate deviations from this mechanism, where OHCA locations tend to be more widespread and deviate from the historical pattern.

To examine the performance of (R–JD), we consider two deployment models for benchmarking. The first is a SAA of the nominal deployment model (N–JD) that finds a deployment to maximize the average

survival rate over the October 2017 data. The second is an ex post model that takes each test set as input and finds a deployment to maximize the average test survival rate. Hence, the ex post model is guaranteed to yield the highest average survival rate on test data, and it can be interpreted as the oracle model having perfect foresight in future events. Their exact MILP formulations are given in Appendix EC.6.

8.2 Values of Accounting for Uncertainty

We first set $\tau = 12$ minutes based on the existing ambulance response time target in Singapore (Minister for Home Affairs 2019), and set the number of available drones/ambulances as $P^{\mathrm{D}} = P^{\mathrm{A}} = 15$. A sensitivity analysis that varies these quantities is provided in the next subsection. Based on the October 2017 data, optimal deployments produced by the robust model (R–JD) and the SAA model are visualized in Figure 3. As expected, SAA tends to deploy drones and ambulances in areas with high density of historical OHCAs.

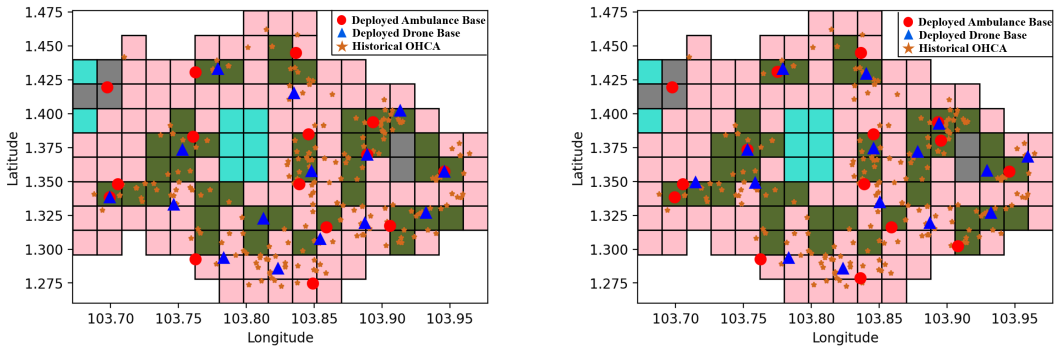


Figure 3 Deployment solutions of 15 ambulances/drones on the main island of Singapore produced by the proposed robust model (left) and the SAA model (right) based on the October 2017 data (cf. Section 7). The grid map in the background is adopted from Figure 1.

Next, we evaluate the performance of the deployments from the robust model, SAA and the ex post model using the bw-h test data generated above. For each h, we compute the overall survival rate in each of the 100 test sets by (2), and compute means and standard deviations (STDs) of the three intervention times by (1), as shown in Figure 4 and Table 2. In this computation, we set the ambulance's travel time to the nearest hospital used in (1) as $o_i = 12$ minutes for all regions $i \in [I]$ (Erkut et al. 2008), and we set the regression coefficients in (2) as $\beta_0 = 67\%$, $\beta_1 = 2.3\%$, $\beta_2 = 1.1\%$, $\beta_3 = 2.1\%$ according to Larsen et al. (1993).

Figure 4 reveals that the proposed robust model performs similar to the SAA under bw-1 that mimics the October 2017 data, and it yields substantially higher mean survival rate and smaller standard deviation compared with the SAA in all other settings that deviate from the October 2017 pattern. The ex post model, serving as the oracle, always yields the highest test survival rate. In Table 1, we report the survival rate gaps of the robust model and the SAA with respect to the ex post model for each setting. It can be seen that the

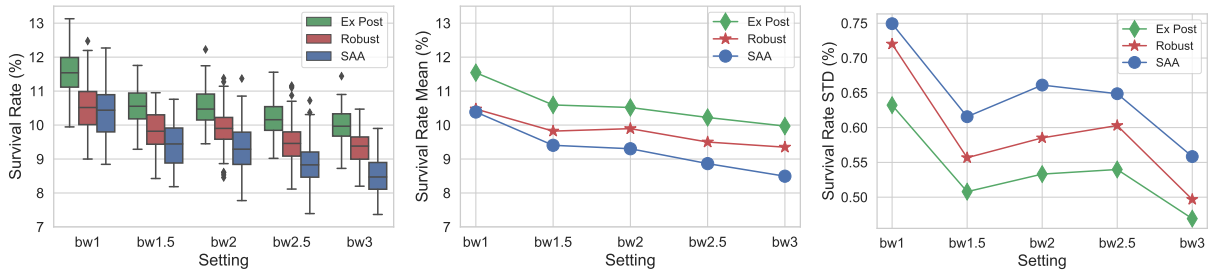


Figure 4 The survival rates (left), means (middle) and standard deviations (right) of the survival rates under each test data setting for the robust, SAA and ex post deployments. The bw-1 setting mimics the October 2017 data, while the other settings simulate deviations from the October 2017 data.

Table 1 Survival rate gaps of the robust and SAA deployments with respect to the ex post deployment under each test data setting. The gap between two mean survival rate $s_1 < s_2$ is calculated by $(s_2 - s_1)/s_2 \times 100\%$.

Setting	bw-1	bw-1.5	bw-2	bw-2.5	bw-3	Total Gap
Robust (in %)	9.261	7.265	5.919	7.033	6.202	35.680
SAA (in %)	10.031	11.231	11.567	13.275	14.803	60.907

robust model cuts the total gap by circa $(61-36)/61 \times 100\% = 41\%$. The above observations demonstrate the values of our proposed model in accounting for uncertainty.

As for the intervention time, Table 2 shows that compared with SAA, the robust model achieves smaller means and standard deviations for all the three interventions (i.e., CPR, defibrillation, and advanced care) in general. This is because SAA tries to exploit the most out of the historical data, making the resulting deployment unreliable to future that deviates from the history. From Table 2, we can see that the defibrillation time is much shorter than the advanced care time. This demonstrates the value of the AED-loaded drone, which is able to arrive at site much earlier than an ambulance to help sustain the chain of survival. By comparing the CPR time and the ambulance's response time (i.e., the advanced care time subtracted by $o_i = 12$ minutes), we see that the first responder initiative shortens the CPR time by nearly one minute. Such improvement explains the high survival rates under the robust deployment (above 9.3% for all test data settings) observed in the middle panel of Figure 4.

8.3 Sensitivity Analysis on Factors Contributing to Survival Outcome

We conduct sensitivity analysis to examine the impacts of various factors, including ambulance's response time target, resource level and responders' behavior, on model performance and survival outcome.

8.3.1 Ambulance response time target.

To understand the impact of ambulance's response time target on the deployment decision, we vary its value by examining $\tau \in \{12, 15, 18\}$ (in minutes). For each τ, we repeat the data analysis in Appendix EC.4 to construct the ambiguity set for the alert response time distribution using the October 2017 data. After

Table 2 Means and standard deviations (in seconds) of the intervention times to CPR, defibrillation, advanced care for the robust, SAA and ex post deployments under each test data setting.

Setting	Intervention	Mean			STD		
		Robust	SAA	Ex Post	Robust	SAA	Ex Post
bw-1	CPR	388.2	389.0	374.7	148.8	162.3	152.7
	Defibrillation	78.7	78.4	72.1	32.9	35.3	34.3
	Advanced Care	548.6	550.5	533.8	130.8	147.9	139.9
bw-1.5	CPR	401.7	406.7	392.3	141.3	154.6	144.1
	Defibrillation	83.4	83.3	76.6	33.3	36.3	35.1
	Advanced Care	550.1	556.5	540.2	126.3	141.7	132.1
bw-2	CPR	393.7	400.2	385.6	148.4	159.7	149
	Defibrillation	87.2	88.9	81.1	35.7	38.1	34.5
	Advanced Care	554.7	563.5	546.2	129.4	142.5	133.2
bw-2.5	CPR	399.7	406.6	390.4	147.2	157.4	147.7
	Defibrillation	87.8	90.3	81.8	35.8	38.9	35.7
	Advanced Care	559	568.3	549.1	128.9	140.4	132.7
bw-3	CPR	402.9	412.5	395.1	147.2	158.4	147.4
	Defibrillation	89.1	92.3	83.2	36.5	39.8	35.8
	Advanced Care	559.3	571.5	550.8	127.5	139.4	130.5

that, we repeat the optimization in Section 8.2 and evaluate the performance of the proposed model, SAA and the ex post model using test data generated in Section 8.1. Similar to Section 8.2, the ex post model outperforms the proposed model and the proposed model outperforms SAA under each τ. As such, we only present results from the proposed model, as shown in Figure 5. It can be seen that a less stringent response time target (i.e., a larger τ) increases the mean survival rate and standard deviation simultaneously. This is because a larger τ enlarges the feasible set of ambulance deployment in (6), giving the deployment program more flexibility in maximizing the total survivals. Meanwhile, as the response time target becomes less stringent, the program tends to deploy ambulances farther to regions with historically low OHCA demands. This can lead to prolonged interventions for OHCA incidents in the test data that realize in these regions, thereby increasing the variation of survival outcomes. In this regard, the response time target τ captures the equity-efficiency tradeoff in this smart response system, where a larger τ leads to a higher efficiency of the system in saving more lives, but a more imbalanced distribution of rescue resources among regions.

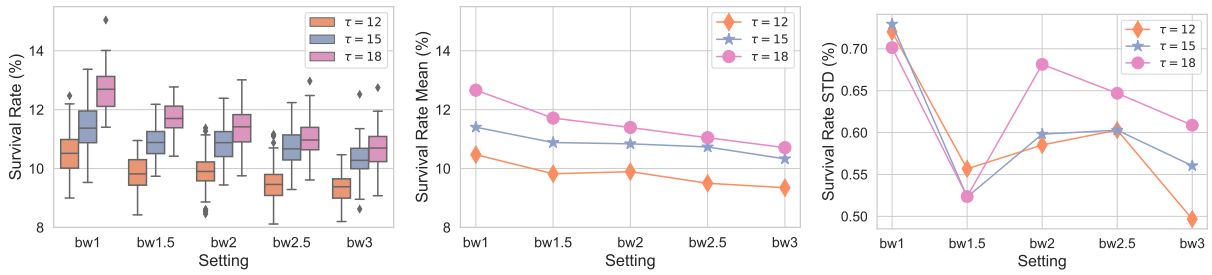


Figure 5 Sensitivity analysis of the ambulance's response time target τ: The left panel shows the survival rates for the optimal deployment of 15 drones/ambulances produced by the robust model with different τ under each test data setting. The middle and right panels show the mean and standard deviation of the associated survival rates under each test data setting, respectively.

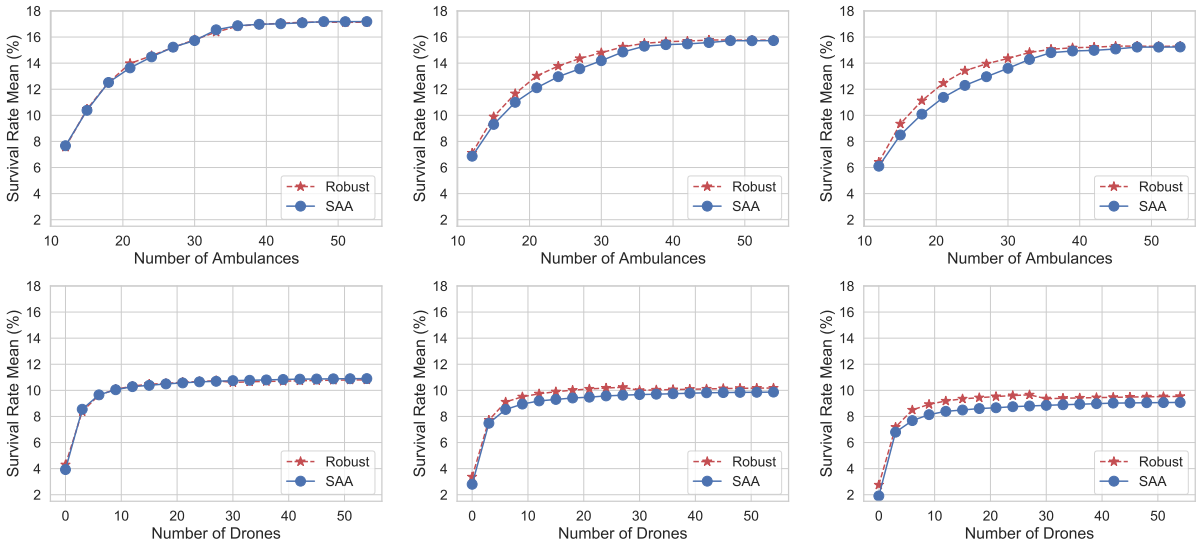
8.3.2 Resource level.

Next, we examine how the survival rate changes with the numbers of ambulances/drones. We vary the number of ambulances P^{A} (resp. drones P^{D}) from 12 to 54 (resp. from 0 to 54) with a step size of 3 while keeping the number of drones P^{D} (resp. ambulances P^{A}) fixed at 15. Continue with the setting in Section 8.2. For each $(P^{\mathrm{D}}, P^{\mathrm{A}})$, we repeat the optimization in Section 8.2 to obtain optimal deployments using SAA and the proposed model. We then evaluate the mean survival rate under each of the deployments using the bw-h test data for $h \in \{1,2,3\}$, as shown in Figure 6. Some quick observations are made: (i) The robust model either performs similar to or better than the SAA for all resource levels; (ii) without drones, the mean survival rate under bw-1 (mimicking the October 2017 data) is circa 4%, comparable to the prevailing 3%–6% OHCA survival rate in Singapore; (iii) the mean survival rate generally increases as the number of drones/ambulances increases, but the marginal benefits diminish quickly.

A closer look into Figure 6 reveals that the survival curve for drone has a steep slope at the beginning. This is because drones have high mobility and velocity, so a few drones are already sufficient to serve well the whole planning area and increase the survival rate significantly. On the other hand, as the numbers of ambulances/drones increase, the ambulance survival rate curve stabilizes at a level higher than that for drone. This is due to the fact that drones affect defibrillation time only, while ambulances affect all the three interventions. These observations shed light on the distinct roles of drones and ambulances in the response system: As a specialized rescue force with high mobility, even a small number of drones are able to dramatically improve survival rate. On the other hand, as the core of the response system, a large number of ambulances are needed to achieve the maximum survival rate of the system. From an economical standpoint, the cost of deploying an extra ambulance, including the purchase and salary to the paramedics, is often much higher than that for a drone. Our finding suggests that it is cost-effective for the system designer to spend budget on a moderate number of drones.

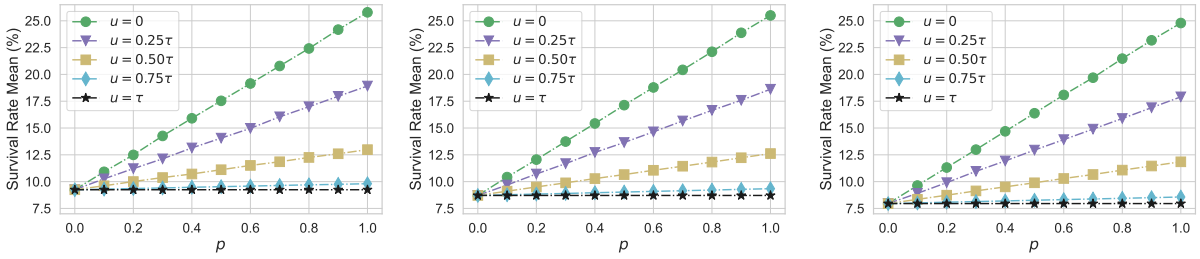
8.3.3 Community responders' behavior.

We continue with the robust deployment of 15 drones/ambulances derived in Section 8.2, and investigate the impacts of community responders' behavior on survival outcome. To this end, for fixed $p \in \{0, 0.1, \ldots, 1\}$ and $u \in \{0, 0.25\tau, \ldots, \tau\}$, we regenerate the responders' behavior records in the bw-h test data (Section 8.1) for $h \in \{1,2,3\}$ as follows: an incident is responded by a responder with probability p, and the responder's response time (if there is a response) is exactly equal to u. Next, we evaluate the mean survival rate under the robust deployment in Figure 3 using the test data indexed by (h, p, u), as shown in Figure 7. The figure clearly shows that responders' behavior significantly affects the survival rate. In the ideal situation where every collapsed patient receives immediate CPR intervention, i.e., $(p,u) = (1,0)$, the survival rate can be even higher than 25%, contrasted with the circa 10% rate in Section 8.2 with 15 ambulances, and the circa 16% rate in Figure 6 with 54 ambulances. This observation supports the need to conduct educational campaign to enhance public awareness and crowd's life-saving skill level.



(a) bw-1 (b) bw-2 (c) bw-3

Figure 6 Mean survival rates for the robust and SAA deployments of different numbers of ambulances/drones under the test data settings bw-1 (a), bw-2 (b) and bw-3 (c). The number of ambulances (resp. drones) is fixed at 15 when varying the number of drones (resp. ambulances).



(a) $h = 1$ (b) $h = 2$ (c) $h = 3$

Figure 7 Mean survival rates for the robust deployment of 15 ambulances/drones derived in Section 8.2 under the test data indexed by (h, p, u): When generating the test data, each OHCA incident is responded with probability p by a responder whose response time is exactly u. Larger h indicates higher dissimilarity between the test data and the October 2017 data in terms of OHCA spatial distribution.

9 Concluding Remarks

In this paper, we have developed a data-driven robust deployment model for a smart OHCA emergency response system involving drones for AED delivery and responders crowdsourced via a mobile App. To the best of our knowledge, this is the first deployment model for EMS that incorporates responders' behavior in a robust data-driven manner. We constructed the uncertainty/ambiguity sets for the robust model so that they are adapted to the structural missing pattern in the historical data, including right-censoring in the alert response time. This is different from the existing uncertainty/ambiguity sets that considers complete data, and our construction sheds lights on the applicability of robust models in complex real-life problems. The way that the uncertainty/ambiguity sets are constructed leads to a large-scale MILP, which is solvable by

the proposed scenario generation algorithm. Our convergence analysis supports the good performance of similar scenario generation algorithms observed in some recent studies, e.g., Chan et al. (2018).

From our extensive numerical experiments based on real data, we find that the proposed robust model leads to a higher survival rate compared with the SAA model even when the data generation mechanism slightly deviates from the training data. We also find that while a less stringent ambulance's response time target τ improves the survival rate of the response system, it aggravates the imbalance in the distribution of rescue resources among regions. This observation resonates with the heated debate on the equity-efficiency tradeoff in healthcare (Asamani et al. 2021, Guindo et al. 2012). In addition to these hardware investments, an important insight from the experiments is the significant impact of responders' behavior on survival outcomes, which outweighs that of simply adding drones/ambulances. This observation highlights the importance of incorporating human responses when making deployment decision.

Our work leaves several avenues for future research. First, while this paper assumes that defibrillation can be conducted as soon as the drone-delivered AED arrives – since it is a highly automatic electronic device that can be used by the general public, we can take a more risk-averse attitude to assume that the delivered AED has to be used by a responder. Under this setting, responders' behavior will have a more dominant role in the deployment of drones. Second, it will be interesting to integrate other smart initiatives with the response system, e.g., AED-loaded taxis or priviate cars.

References

Agatz, Niels, Paul Bouman, Marie Schmidt. 2018. Optimization approaches for the traveling salesman problem with drone. *Transportation Science*, 52 (4), 965-981.

Asamani, James Avoka, Samuel Anongiba Alugsi, Hamza Ismaila, Juliet Nabyonga-Orem. 2021. Balancing equity and efficiency in the allocation of health resources—where is the middle ground? *Healthcare*, vol. 9. MDPI, 1257.

Atamtürk, Alper, Muhong Zhang. 2007. Two-stage robust network flow and design under demand uncertainty. *Operations Research*, 55 (4), 662-673.

Baron, Opher, Joseph Milner, Hussein Naseraldin. 2011. Facility location: A robust optimization approach. *Production and Operations Management*, 20 (5), 772-785.

Boutilier, Justin J, Steven C Brooks, Alyf Janmohamed, Adam Byers, Jason E Buick, Cathy Zhan, Angela P Schoellig, Sheldon Cheskes, Laurie J Morrison, Timothy CY Chan. 2017. Optimizing a drone network to deliver automated external defibrillators. *Circulation*, 135 (25), 2454-2465.

Boutilier, Justin J, Timothy CY Chan. 2022. Drone network design for cardiac arrest response. *Manufacturing & Service Operations Management*, to appear.

Brooks, Steven C, Graydon Simmons, Heather Worthington, Bentley J Bobrow, Laurie 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.

Carlsson, John Gunnar, Siyuan Song. 2018. Coordinated logistics with a truck and a drone. *Management Science*, 64 (9), 4052-4069.

Chan, Timothy CY, Derya Demirtas, Roy H Kwon. 2016. Optimizing the deployment of public access defibrillators. *Management Science*, 62 (12), 3617-3635.

Chan, Timothy CY, Heyse Li, Gerald Lebovic, Sabrina K Tang, Joyce YT Chan, Horace CK Cheng, Laurie J Morrison, Steven C Brooks. 2013. Identifying locations for public access defibrillators using mathematical optimization. *Circulation*, 127 (17), 1801-1809.

Chan, Timothy CY, Zuo-Jun Max Shen, Auyon Siddiq. 2018. Robust defibrillator deployment under cardiac arrest location uncertainty via row-and-column generation. *Operations Research*, 66 (2), 358-379.

Cho, Soo-Haeng, Hoon Jang, Taesik Lee, John Turner. 2014. Simultaneous location of trauma centers and helicopters for emergency medical service planning. *Operations Research*, 62 (4), 751-771.

Claesson, Andreas, D Fredman, Leif Svensson, Mattias Ringh, Jacob Hollenberg, Per Nordberg, M Rosenqvist, Therese Djarv, S Österberg, Josefin Lennartsson, et al. 2016. Unmanned aerial vehicles (drones) in out-of-hospital-cardiac-arrest. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 24 (1), 1-9.

Coute, Ryan A, Brian H Nathanson, Michael C Kurz, Stephanie DeMasi, Bryan McNally, Timothy J Mader. 2021. Annual and lifetime economic productivity loss due to adult out-of-hospital cardiac arrest in the united states: A study for the cares surveillance group. *Resuscitation*, 167 111-117.

Delhomme, Clémence, Mario Njeim, Emilie Varlet, Louis Pechmajou, Nordine Benameur, Pascal Cassan, Clément Derkenne, Daniel Jost, Lionel Lamhaut, Eloi Marijon, et al. 2019. Automated external defibrillator use in out-of-hospital cardiac arrest: Current limitations and solutions. *Archives of Cardiovascular Diseases*, 112 (3), 217-222.

Derevitskii, Ivan, Nikita Kogtikov, Michael H Lees, Wentong Cai, Marcus EH Ong. 2020. Risk-based aed placement-singapore case. *International Conference on Computational Science*. Springer, 577-590.

Erkut, Erhan, Armann Ingolfsson, Güneş Erdoğan. 2008. Ambulance location for maximum survival. *Naval Research Logistics*, 55 (1), 42-58.

Google Map. 2022. Google map distance api. https://developers.google.com/maps/documentation/distance-matrix. Accessed: 2022-07-08.

Goverment. 2020. Singapore hospitals and polyclinics. https://www.gov.sg/article/quiz-how-many-hospitals-and-polyclinics-do-we-have. Accessed: 2022-04-07.

Goverment. 2021. Singapore police station. https://www.sgdi.gov.sg/other-organisations/police-station. Accessed: 2022-04-07.

Guindo, Lalla Aïda, Monika Wagner, Rob Baltussen, Donna Rindress, Janine van Til, Paul Kind, Mireille M Goetghebeur. 2012. From efficacy to equity: Literature review of decision criteria for resource allocation and healthcare decisionmaking. *Cost Effectiveness and Resource Allocation*, 10 (1), 1-13.

Gundry, John W, Keith A Comess, Frances A DeRook, Dawn Jorgenson, Gust H Bardy. 1999. Comparison of naive sixth-grade children with trained professionals in the use of an automated external defibrillator. *Circulation*,

100 (16), 1703-1707.

Huang, Ling Hsuan, Yu-Ni Ho, Ming-Ta Tsai, Wei-Ting Wu, Fu-Jen Cheng. 2021. Response time threshold for predicting outcomes of patients with out-of-hospital cardiac arrest. *Emergency Medicine International*, 2021.

Javanmard, Adel, Andrea Montanari. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15 (1), 2869-2909.

Kaplan, Edward L, Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53 (282), 457-481.

Kim, Seon Jin, Gino J Lim, Jaeyoung Cho, Murray J Côté. 2017. Drone-aided healthcare services for patients with chronic diseases in rural areas. *Journal of Intelligent & Robotic Systems*, 88 (1), 163-180.

Lai, Hsuan, Caroline V Choong, Stephanie Fook-Chong, Yih Yng Ng, Eric A Finkelstein, Benjamin Haaland, E Shaun Goh, Benjamin Sieu-Hon Leong, Han Nee Gan, David Foo, et al. 2015. Interventional strategies associated with improvements in survival for out-of-hospital cardiac arrests in singapore over 10 years. *Resuscitation*, 89 155-161.

Larsen, Mary P, Mickey S Eisenberg, Richard O Cummins, Alfred P Hallstrom. 1993. Predicting survival from out-of-hospital cardiac arrest: a graphic model. *Annals of Emergency Medicine*, 22 (11), 1652-1658.

Leeson, George W. 2018. The growth, ageing and urbanisation of our world. *Journal of Population Ageing*, 11 (2), 107-115.

Minister for Home Affairs. 2019. Written reply to parliamentary question on ambulance response times. https://www.mha.gov.sg/mediaroom/parliamentary/written-reply-to-parliamentary-question-on-ambulance-response-times-by-mr-k-shanmugam- Accessed: 2022-07-08.

Mohajerin Esfahani, Peyman, Daniel Kuhn. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171 (1), 115-166.

Murray, Chase C, Amanda G Chu. 2015. The flying sidekick traveling salesman problem: Optimization of drone-assisted parcel delivery. *Transportation Research Part C: Emerging Technologies*, 54 86-109.

Ng, Wei Ming, Carl Ross De Souza, Pin Pin Pek, Nur Shahidah, Yih Yng Ng, Shalini Arulanandam, Alexander Elgin White, Benjamin Sieu-Hon Leong, Marcus Eng Hock Ong. 2021. myResponder smartphone application to crowdsource basic life support for out-of-hospital cardiac arrest: the singapore experience. *Prehospital Emergency Care*, 25 (3), 388-396.

Pulver, Aaron, Ran Wei, Clay Mann. 2016. Locating aed enabled medical drones to enhance cardiac arrest response times. *Prehospital Emergency Care*, 20 (3), 378-389.

Rajagopalan, Hari K, Cem Saydam. 2009. A minimum expected response model: Formulation, heuristic solution, and application. *Socio-Economic Planning Sciences*, 43 (4), 253-262.

SCDF. 2019. Scdf division hq. https://www.scdf.gov.sg/home/about-us/organisation-structure/scdf-division-hq. Accessed: 2022-04-07.

Schierbeck, Sofia, Jacob Hollenberg, Anette Nord, Leif Svensson, Per Nordberg, Mattias Ringh, Sune Forsberg, Peter Lundgren, Christer Axelsson, Andreas Claesson. 2022. Automated external defibrillators delivered by drones to patients with suspected out-of-hospital cardiac arrest. *European Heart Journal*, 43 (15), 1478-1487.

Scott, Judy E, Carlton H Scott. 2019. Models for drone delivery of medications and other healthcare items. *International Journal of Healthcare Information Systems and Informatics*, 376-392.

Scquizzato, Tommaso, Ottavia Pallanch, Alessandro Belletti, Antonio Frontera, Luca Cabrini, Alberto Zangrillo, Giovanni 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 16-25.

Siddiq, Auyon A, Steven C Brooks, Timothy CY Chan. 2013. Modeling the impact of public access defibrillator range on public location cardiac arrest coverage. *Resuscitation*, 84 (7), 904-909.

SingHealth. 2022. New program to improve standard of emergency care outside of hospital. https://www.straitstimes.com/singapore/new-programme-to-improve-standard-of-emergency-care-outside-of-hospital-launched-by-te Accessed: 2022-09-20.

Smith, Christopher M, Mark H Wilson, Ali Ghorbangholi, Christopher Hartley-Sharpe, Carl Gwinnutt, Bridget Dicker, Gavin D Perkins. 2017. The use of trained volunteers in the response to out-of-hospital cardiac arrest–the goodsam experience. *Resuscitation*, 121 123-126.

Sun, Christopher LF, Derya Demirtas, Steven C Brooks, Laurie J Morrison, Timothy CY Chan. 2016. Overcoming spatial and temporal barriers to public access defibrillators via optimization. *Journal of the American College of Cardiology*, 68 (8), 836-845.

Toro-Díaz, Hector, Maria E Mayorga, Sunarin Chanta, Laura A McLay. 2013. Joint location and dispatching decisions for emergency medical services. *Computers & Industrial Engineering*, 64 (4), 917-928.

van den Berg, Pieter L, Karen Aardal. 2015. Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operational Research*, 242 (2), 383-389.

Virani, Salim S, Alvaro Alonso, Hugo J Aparicio, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Susan Cheng, Francesca N Delling, et al. 2021. Heart disease and stroke statistics—2021 update: a report from the american heart association. *Circulation*, 143 (8), e254-e743.

Wang, Zheng, Jiuh-Biing Sheu. 2019. Vehicle routing problem with drones. *Transportation Research Part B: Methodological*, 122 350-364.

Warden, Craig, Michael T Cudnik, Comilla Sasson, Greg Schwartz, Hugh Semple. 2012. Poisson cluster analysis of cardiac arrest incidence in columbus, ohio. *Prehospital Emergency Care*, 16 (3), 338-346.

White, AE, JS Poh, PP Pek, et al. 2021. Singapore out-of-hospital cardiac arrest registry report 2011-2018. *Unit for Pre-hospital Emergency Care, Singapore. Available at: https://www.myheart.org.sg/wp-content/uploads/2021/01/Singapore-OHCA-Data-Report-2011-2018.pdf.*, 3.

Electronice Companion to "Robust Data-Driven Design of a Smart Cardiac Arrest Response System"

EC.1 Proofs of Main Results

In this section we provide proofs for all Propositions and Theorem in the main article.

Proof of Proposition 1. Note that the ambiguity set Q defined in (11) can be written equivalently as

$$Q = \left\{ \mathbb{Q} \in \mathcal{P}_0(\mathbb{R}_+^I) \middle| \begin{array}{l} \Pi_{\tilde{u}_i}\mathbb{Q} = \mathbb{P}_i, i \in [I]] \\ \mathbb{P}_i \in \mathcal{F}_i, i \in [I] \end{array} \right\},$$

where the individual ambiguity set $\mathcal{F}_i \subset \mathcal{P}_0(\mathbb{R}_+)$ is given by

$$\mathcal{F}_i \triangleq \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}_+) \middle| \begin{array}{l} \tilde{u}_i \sim \mathbb{P} \\ \mathbb{P}\left[\tilde{u}_i > \tau\right] \in [\underline{q}_i, \bar{q}_i] \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{u}_i \middle| \tilde{u}_i \le \tau\right] = \check{u}_i \\ \mathbb{E}_{\mathbb{P}}\left[|\tilde{u}_i - \check{u}_i| \middle| \tilde{u}_i \le \tau\right] \le \varepsilon \end{array} \right\}, \ i \in [I].$$

Hence, we have

$$\begin{aligned} & \max_{\mathbb{Q}\in Q} \ \mathbb{E}_{\mathbb{Q}} \sum_{i\in[I]} d_i p_i \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \\ = & \max_{\mathbb{Q}\in Q} \ \sum_{i\in[I]} d_i p_i \mathbb{E}_{\mathbb{Q}} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \\ = & \max_{\mathbb{P}_i\in\mathcal{F}_i, i\in[I]} \ \sum_{i\in[I]} d_i p_i \mathbb{E}_{\mathbb{P}_i} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \\ = & \sum_{i\in[I]} d_i p_i \max_{\mathbb{P}_i\in\mathcal{F}_i} \ \mathbb{E}_{\mathbb{P}_i} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \end{aligned} \tag{EC.1}$$

Because (EC.1) is the summation of I separable maximization problems, it suffices to focus on the evaluation of one of them, i.e., the evaluation of $\max_{\mathbb{P}_i\in\mathcal{F}_i} \ \mathbb{E}_{\mathbb{P}_i} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$. To this end, We first introduce a closely related ambiguity set $\hat{\mathcal{F}}_i(q_i)$ that replaces the first constraint in $\mathcal{F}_i$ by $\mathbb{P}\left[\tilde{u}_i > \tau\right] = q_i$, i.e., the set

$$\hat{\mathcal{F}}_i(q_i) \triangleq \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}_+) \middle| \begin{array}{l} \tilde{u}_i \sim \mathbb{P} \\ \mathbb{P}\left[\tilde{u}_i > \tau\right] = q_i \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{u}_i \middle| \tilde{u}_i \le \tau\right] = \check{u}_i \\ \mathbb{E}_{\mathbb{P}}\left[|\tilde{u}_i - \check{u}_i| \middle| \tilde{u} \le \tau\right] \le \varepsilon \end{array} \right\}.$$

With the above ambiguity set, we have

$$\begin{aligned} & \max_{\mathbb{P}_i\in\mathcal{F}_i} \ \mathbb{E}_{\mathbb{P}_i} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \\ = & \max_{q_i\in[\underline{q}_i,\bar{q}_i]} \max_{\mathbb{P}_i\in\hat{\mathcal{F}}_i(q_i)} \ \mathbb{E}_{\mathbb{P}_i} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} && \text{(EC.2)} \\ = & \max_{q_i\in[\underline{q}_i,\bar{q}_i]} \max_{\mathbb{P}_i\in\hat{\mathcal{F}}_i(q_i)} \ \mathbb{E}_{\mathbb{P}_i}[\min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}|\tilde{u}_i \le \tau]\mathbb{P}_i[\tilde{u}_i \le \tau] + \mathbb{E}_{\mathbb{P}_i}[\min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}|\tilde{u}_i > \tau]\mathbb{P}_i[\tilde{u}_i > \tau] && \text{(EC.3)} \\ = & \max_{q_i\in[\underline{q}_i,\bar{q}_i]} \max_{\mathbb{P}_i\in\hat{\mathcal{F}}_i(q_i)} \ (1-q_i)\mathbb{E}_{\mathbb{P}_i}[\min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}|\tilde{u}_i \le \tau] + q_i \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}. && \text{(EC.4)} \end{aligned}$$

In the above, (EC.2) is due to the definition of $\hat{\mathcal{F}}_i(q_i)$ and the separability of maximization operator; (EC.3) is due to the rule of total probability; (EC.4) is due to the constraint $\mathbb{P}_i[\tilde{u}_i > \tau] = 1 - \mathbb{P}_i[\tilde{u}_i \le \tau] = q_i$, and because $\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \le \tau$.

To evaluate (EC.4), note that the inner maximization problem $\max_{\mathbb{P}_i \in \hat{\mathcal{F}}_i(q_i)} \mathbb{E}_{\mathbb{P}_i}[\min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} | \tilde{u}_i \le \tau]$ can be written as $\max_{\mathbb{P}\in\hat{\mathcal{F}}_i^{\mathrm{c}}} \mathbb{E}_{\mathbb{P}}[\min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}]$, where

$$\hat{\mathcal{F}}_i^{\mathrm{c}} \triangleq \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}_+) \middle| \begin{array}{l} \tilde{u}_i \sim \mathbb{P} \\ \mathbb{P}\left[\tilde{u}_i \in [0,\tau]\right] = 1 \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{u}_i\right] = \check{u}_i \\ \mathbb{E}_{\mathbb{P}}\left[|\tilde{u}_i - \check{u}_i|\right] \le \varepsilon \end{array} \right\}$$

is the conditional ambiguity set of $\hat{\mathcal{F}}_i(q_i)$ conditioning on $\tilde{u}_i \le \tau$. Hence, the inner maximization problem is equivalent to the following moment problem:

$$\begin{array}{ll} \max\limits_{f(u)\ge 0} & \int_0^\tau \min\{u, \sum\limits_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} f(u)\mathrm{d}u \\ \text{s.t.} & \int_0^\tau f(u)\mathrm{d}u = 1, \\ & \int_0^\tau u f(u)\mathrm{d}u = \check{u}_i, \\ & \int_0^\tau |u - \check{u}_i| f(u)\mathrm{d}u \le \varepsilon. \end{array} \tag{EC.5}$$

Introducing dual variables $\alpha \in \mathbb{R}, \beta \in \mathbb{R}, \gamma \in \mathbb{R}_+$ to the normalization, expectation, and dispersion constraints, respectively, the dual of (EC.5) can be formulated as the following semi-infinite program

$$\begin{array}{ll} \max\limits_{\alpha,\beta,\gamma} & \alpha + \check{u}_i\beta + \varepsilon\gamma \\ \text{s.t.} & \alpha \ge \min\{u, \sum\limits_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} - \beta u - \gamma|u - \check{u}_i|, \quad \forall u \in [0,\tau], \\ & \gamma \ge 0. \end{array} \tag{EC.6}$$

Strong duality between (EC.5) and (EC.6) is guaranteed by Proposition 3.4 in Shapiro (2001), because the Dirac measure placing unit mass on $\check{u}_i$ is a Slater distribution with respect to the conditional ambiguity set $\hat{\mathcal{F}}_i^{\mathrm{c}}$. Since $\gamma \ge 0$, the maximum of the right-hand side (RHS) of the above semi-infinite constraint amounts to the maximization problem

$$\begin{array}{ll} \max\limits_{\lambda,u,\mu} & \lambda - \beta u - \gamma\mu \\ \text{s.t.} & \lambda \le u,\ \lambda \le \sum\limits_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}, \\ & \mu \ge u - \check{u}_i,\ \mu \ge -u + \check{u}_i, \\ & 0 \le u \le \tau. \end{array} \tag{EC.7}$$

In (EC.7), λ is an auxiliary variable to linearize the term $\min\{u, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$, and μ is an auxiliary variable to linearize the term $|u - \check{u}_i|$. Introducing dual variables κ_1, κ_2 to the first row of constraints, θ_1, θ_2 to the second row, and η_1, η_2 to the third row, the dual of (EC.7) becomes

$$\begin{aligned} \min_{\kappa_1,\kappa_2,\theta_1,\theta_2,\eta_1,\eta_2} \quad & \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \kappa_2 + \check{u}_i \theta_1 - \check{u}_i \theta_2 + \tau \eta_2 \\ \text{s.t.} \quad & \kappa_1 + \kappa_2 = 1, \\ & \beta = \kappa_1 - \theta_1 + \theta_2 + \eta_1 - \eta_2, \\ & \gamma = \theta_1 + \theta_2, \\ & \kappa_1, \kappa_2, \theta_1, \theta_2, \eta_1, \eta_2 \geq 0. \end{aligned} \tag{EC.8}$$

Strong duality between (EC.7) and (EC.8) is guaranteed by the duality theory of linear programming. Replacing the semi-infinite constraint in (EC.6) by (EC.8), the dual of (EC.5) becomes

$$\begin{aligned} \min_{\alpha,\beta,\gamma,\boldsymbol{\kappa},\boldsymbol{\theta},\boldsymbol{\eta}} \quad & \alpha + \check{u}_i \beta + \varepsilon \gamma \\ \text{s.t.} \quad & \alpha \geq \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \kappa_2 + \check{u}_i \theta_1 - \check{u}_i \theta_2 + \tau \eta_2, \\ & \kappa_1 + \kappa_2 = 1, \\ & \beta = \kappa_1 - \theta_1 + \theta_2 + \eta_1 - \eta_2, \\ & \gamma = \theta_1 + \theta_2, \\ & \kappa_1, \kappa_2, \theta_1, \theta_2, \eta_1, \eta_2, \gamma \geq 0. \end{aligned}$$

Since this is a minimization problem, the first constraint must be tight at optimality. Expressing α, β, γ in the objective function by $\boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{\eta}$ using the first, third and fourth constraints, we arrive at the equivalent program

$$\begin{aligned} \min_{\boldsymbol{\kappa},\boldsymbol{\theta},\boldsymbol{\eta}} \quad & \check{u}_i \kappa_1 + \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \kappa_2 + \check{u}_i \eta_1 + (\tau - \check{u}_i) \eta_2 + \varepsilon(\theta_1 + \theta_2) \\ \text{s.t.} \quad & \kappa_1 + \kappa_2 = 1, \\ & \kappa_1, \kappa_2, \theta_1, \theta_2, \eta_1, \eta_2 \geq 0. \end{aligned}$$

Since $\check{u}_i \geq 0$, $\tau - \check{u}_i \geq 0$, and $\varepsilon \geq 0$ by the regularity condition, we must have $\boldsymbol{\theta} = \boldsymbol{\eta} = \mathbf{0}$ at optimality. It follows that the above program simplifies to

$$\begin{aligned} \min_{\boldsymbol{\kappa}} \quad & \check{u}_i \kappa_1 + \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \kappa_2 \\ \text{s.t.} \quad & \kappa_1 + \kappa_2 = 1, \\ & \kappa_1, \kappa_2 \geq 0, \end{aligned}$$

whose optimal value is essentially $\min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$. Putting all the pieces together, we have shown that the dual of (EC.5) has the optimal value $\min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$, which implies that

$$\max_{\mathbb{P}_i \in \hat{\mathcal{F}}_i(q_i)} \mathbb{E}_{\mathbb{P}_i}[\min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} | \tilde{u}_i \leq \tau] = \min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \tag{EC.9}$$

due to the strong duality established above.

Finally, we obtain

$$\begin{aligned}
&\max_{\mathbb{P}_i\in\mathcal{F}_i} \mathbb{E}_{\mathbb{P}_i} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \\
&= \max_{q_i\in[\underline{q}_i,\bar{q}_i]} (1-q_i)\min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} + q_i \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \qquad \text{(EC.10)}\\
&= (1-\bar{q}_i)\min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} + \bar{q}_i \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}. \qquad \text{(EC.11)}
\end{aligned}$$

In the above, (EC.10) is obtained by substituting (EC.9) into (EC.4); and (EC.11) is because $\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} - \min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} > 0$, which implies that the maximum of (EC.10) is achieved when $q_i = \bar{q}_i$. Therefore, we have

$$\begin{aligned}
&\max_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}_{\mathbb{Q}} \left[\sum_{i\in[I]} d_i p_i \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}\right] \\
&= \sum_{i\in[I]} d_i p_i \max_{\mathbb{P}_i\in\mathcal{F}_i} \mathbb{E}_{\mathbb{P}_i} \min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \\
&= \sum_{i\in[I]} d_i p_i \left((1-\bar{q}_i)\min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} + \bar{q}_i \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\right).
\end{aligned}$$

The proof is complete. □

Proof of Proposition 2. By the definition of $\mathcal{P}(\boldsymbol{d})$ defined in (13), the maximization problem can be written as

$$\begin{aligned}
\max_{\boldsymbol{p}\geq\mathbf{0}} \quad & \sum_{i\in[I]} d_i w_i p_i \\
\text{s.t.} \quad & \underline{p}_{(v)} \leq \frac{\sum_{i\in I_{(v)}} d_i p_i}{\sum_{i\in I_{(v)}} d_i} \leq \bar{p}_{(v)},\ v\in[V], \qquad \text{(EC.12)}\\
& p_i \leq 1,\ i\in[I].
\end{aligned}$$

Let $\underline{\lambda}_{(v)} \geq 0$, $\bar{\lambda}_{(v)} \geq 0$, $v\in[V]$ and $\mu_i \geq 0$, $i\in[I]$ be the dual variables associated with the first and second line of constraints in the above program, respectively. Then the Lagrangian function associated with the above program can be written as

$$\begin{aligned}
f^{\mathrm{L}}(\underline{\boldsymbol{\lambda}}, \bar{\boldsymbol{\lambda}}, \boldsymbol{\mu}, \boldsymbol{p}) &\triangleq \sum_{i\in[I]} d_i w_i p_i + \sum_{v\in[V]} \left(\sum_{i\in I_{(v)}} d_i p_i - \sum_{i\in I_{(v)}} d_i \underline{p}_{(v)}\right) \underline{\lambda}_{(v)} + \sum_{v\in[V]} \left(\sum_{i\in I_{(v)}} d_i \bar{p}_{(v)} - \sum_{i\in I_{(v)}} d_i p_i\right) \bar{\lambda}_{(v)} + \sum_{i\in[I]} (1-p_i)\mu_i \\
&= \sum_{i\in[I]} \left(d_i w_i + \sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}} d_i \left(\underline{\lambda}_{(v)} - \bar{\lambda}_{(v)}\right) - \mu_i\right) p_i + \sum_{v\in[V]} d_i \sum_{i\in I_{(v)}} \left(\bar{p}_{(v)}\bar{\lambda}_{(v)} - \underline{p}_{(v)}\underline{\lambda}_{(v)}\right) + \sum_{i\in[I]} \mu_i.
\end{aligned}$$

Hence, the dual of (EC.12), given by

$$\min_{\underline{\boldsymbol{\lambda}}\geq\mathbf{0}, \bar{\boldsymbol{\lambda}}\geq\mathbf{0}, \boldsymbol{\mu}\geq\mathbf{0}} \max_{\boldsymbol{p}\geq\mathbf{0}} f^{\mathrm{L}}(\underline{\boldsymbol{\lambda}}, \bar{\boldsymbol{\lambda}}, \boldsymbol{\mu}, \boldsymbol{p}),$$

can be written as

$$\begin{aligned}\min_{\underline{\lambda},\bar{\lambda},\mu} \quad & \sum_{v\in[V]}\sum_{i\in I_{(v)}} d_i(\bar{p}_{(v)}\bar{\lambda}_{(v)} - \underline{p}_{(v)}\underline{\lambda}_{(v)}) + \sum_{i\in[I]}\mu_i \\ \text{s.t.} \quad & d_i \sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}_{(v)} - \underline{\lambda}_{(v)}) + \mu_i \le d_i w_i,\ i\in[I], \\ & \underline{\lambda}_{(v)} \ge 0, \bar{\lambda}_{(v)} \ge 0,\ v\in[V], \\ & \mu_i \ge 0,\ i\in[I].\end{aligned}$$

Strong duality is guaranteed by the duality theory of linear programming, since the primal problem (EC.12) is feasible and bounded due to the non-emptiness condition of $\boldsymbol{p}(\boldsymbol{d})$. The proof is complete. □

Proof of Proposition 3. First, since the V clusters are non-overlapping, we can construct a feasible solution $\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d})$ by letting $p_i = p_v \in [\underline{p}_{(v)}, \bar{p}_{(v)}]$ for all $i\in I_{(v)}$, $v\in[V]$. Therefore, the set $\mathcal{P}(\boldsymbol{d})$ is bounded and non-empty, and thus the program (EC.12) has a finite optimal value which is attained by at least one optimal solution in the set $\mathcal{P}(\boldsymbol{d})$. We first show that there exists an optimal solution $\boldsymbol{p}^\star$ to (EC.12) such that $\sum_{i\in I_{(v)}} d_i p_i^\star / \sum_{i\in I_{(v)}} d_i = \underline{p}_{(v)}$. Assume for the sake of contradiction that such $\boldsymbol{p}^\star$ does not exist. Let $\boldsymbol{p}'$ be an optimal solution to (EC.12). By our hypothesis, $\sum_{i\in I_{(v')}} d_i p_i' / \sum_{i\in I_{(v')}} d_i = \underline{p}_{(v')} + \delta$ for some $\delta > 0$ and $v'\in[V]$. We then construct $\boldsymbol{p}^\star$ by letting $p_i^\star = p_i'$ for all $i\in I_{(v)}$, $v\in[V]\backslash\{v'\}$, and $p_i^\star = p_i' - \delta_i$ for all $i\in I_{(v')}$, where the value of δ_i satisfy $0\le\delta_i\le p_i'$ for all $i\in I_{(v')}$, and $\sum_{i\in I_{(v')}} d_i\delta_i / \sum_{i\in I_{(v')}} d_i = \delta$. By our construction, the solution $p^\star$ remains feasible in the set $\mathcal{P}(\boldsymbol{d})$ and satisfies $\sum_{i\in I_{(v')}} d_i p_i^\star / \sum_{i\in I_{(v')}} d_i = \underline{p}_{(v')}$. Further, we have $\sum_{i\in[I]} d_i w_i p_i' - \sum_{i\in[I]} d_i w_i p_i^\star = \sum_{i\in I_{(v')}} d_i w_i \delta_i \le 0$ because of the condition $d_i w_i \le 0$ and $\delta_i \ge 0$. Since the solution $p^\star$ yields an objective value of (EC.12) that is no less than the optimal solution p', the solution $p^\star$ must be optimal as well. This leads to a contraction. As such, the program (EC.12) is equivalent to:

$$\begin{aligned}\max_{\boldsymbol{p}\ge\boldsymbol{0}} \quad & \sum_{i\in[I]} d_i w_i p_i \\ \text{s.t.} \quad & \frac{\sum_{i\in I_{(v)}} d_i p_i}{\sum_{i\in I_{(v)}} d_i} = \underline{p}_{(v)},\ v\in[V], \\ & p_i \le 1,\ i\in[I].\end{aligned}$$

Let $\lambda_{(v)}\ge 0$, $\bar{\lambda}_{(v)}\ge 0$, $v\in[V]$ and $\mu_i\ge 0$, $i\in[I]$ be the dual variables associated with the first and second line of constraints in the above program, respectively. We can then obtain the desired dual program by following the same procedure as that in the proof of Proposition 2. Finally, the relation that $p_{i_1^v}^\star \le p_{i_2^v}^\star \le \cdots \le p_{i_{|I_{(v)}|}^v}^\star$ follows from the nature of the maximization. The proof is complete. □

Proof of Theorem 1. To start, let us define $\phi_i \triangleq \min\{\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$, $\psi_i \triangleq \min\{\breve{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$ and $w_i \triangleq (1-\bar{q}_i)(\psi_i - \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}})$, $i\in[I]$.

By Proposition 1 (applicable because $\sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik} \le \tau$, $i \in [I]$, by the set of constraints on ambulance's maximum allowable response time) and Proposition 2, the three-layer maximization problem in (R–JD), i.e.,

$$\max_{\boldsymbol{d}\in\mathcal{D}} \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d})} \max_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}_{\mathbb{Q}}\left[\sum_{i\in[I]} d_i \left(p_i \min\{\tilde{u}_i, \sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}\} + \rho_1 \min\{\sum_{j\in[J]} x^{\mathrm{D}}_{ij} t^{\mathrm{D}}_{ij}, \sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}\} + (1 - p_i + \rho_2) \sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}\right)\right],$$

is equivalent to

$$\begin{aligned}
\max_{\boldsymbol{d}\in\mathcal{D}} \min_{\underline{\boldsymbol{\lambda}},\bar{\boldsymbol{\lambda}},\boldsymbol{\mu}} \ & \sum_{i\in[I]} d_i\big(\rho_1\phi_i + (1+\rho_2)\sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}\big) + \sum_{v\in[V]}\sum_{i\in I_{(v)}} d_i(\bar{p}_{(v)}\bar{\lambda}_{(v)} - \underline{p}_{(v)}\underline{\lambda}_{(v)}) + \sum_{i\in[I]}\mu_i \\
\text{s.t.} \quad & d_i \sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}_{(v)} - \underline{\lambda}_{(v)}) + \mu_i \ge d_i(1-\bar{q}_i)(\psi_i - \sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}),\ i \in I_{(v)}, v \in [V], \\
& \underline{\boldsymbol{\lambda}} \in \mathbb{R}^V_+, \bar{\boldsymbol{\lambda}} \in \mathbb{R}^V_+, \boldsymbol{\mu} \in \mathbb{R}^I_+.
\end{aligned}$$

Because the number of scenarios in $\mathcal{D}$ is finite and equal to B, the above program can be written as

$$\begin{aligned}
\max_{b\in[B]} \min_{\underline{\boldsymbol{\lambda}}^{(b)},\bar{\boldsymbol{\lambda}}^{(b)},\boldsymbol{\mu}^{(b)}} \ & \sum_{i\in[I]} d^{(b)}_i\big(\rho_1\phi_i + (1+\rho_2)\sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}\big) + \sum_{v\in[V]}\sum_{i\in I_{(v)}} d^{(b)}_i(\bar{p}_{(v)}\bar{\lambda}^{(b)}_{(v)} - \underline{p}_{(v)}\underline{\lambda}^{(b)}_{(v)}) + \sum_{i\in[I]}\mu^{(b)}_i \\
\text{s.t.} \quad & d^{(b)}_i \sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}^{(b)}_{(v)} - \underline{\lambda}^{(b)}_{(v)}) + \mu^{(b)}_i \ge d^{(b)}_i(1-\bar{q}_i)(\psi_i - \sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}),\ i \in I_{(v)}, v \in [V], \\
& \underline{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}^V_+, \bar{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}^V_+, \boldsymbol{\mu}^{(b)} \in \mathbb{R}^I_+.
\end{aligned} \tag{EC.13}$$

The above program is essentially the maximum of B separable minimization programs, and the decision variables in the B-th minimization program is $\left(\underline{\boldsymbol{\lambda}}^{(b)}, \bar{\boldsymbol{\lambda}}^{(b)}, \boldsymbol{\mu}^{(b)}\right)$. To obtain a single-layer minimization problem, we introduce an epigraph decision variable $\zeta \in \mathbb{R}$, and add the constraint that ζ is no less than the optimal value of each of the B minimization programs. Then the optimal value of (EC.13) coincides with the smallest ζ that satisfies such constraint. Therefore, (EC.13) is equivalent to

$$\begin{aligned}
\min_{\underline{\boldsymbol{\lambda}}^{(b)},\bar{\boldsymbol{\lambda}}^{(b)},\boldsymbol{\mu}^{(b)},\zeta} \ & \zeta \\
\text{s.t.} \quad & \zeta \ge \sum_{i\in[I]} d^{(b)}_i\big(\rho_1\phi_i + (1+\rho_2)\sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}\big) + \sum_{v\in[V]}\sum_{i\in I_{(v)}} d^{(b)}_i(\bar{p}_{(v)}\bar{\lambda}^{(b)}_{(v)} - \underline{p}_{(v)}\underline{\lambda}^{(b)}_{(v)}) + \sum_{i\in[I]}\mu^{(b)}_i,\ b \in [B], \\
& d^{(b)}_i \sum_{v\in[V]} \mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}^{(b)}_{(v)} - \underline{\lambda}^{(b)}_{(v)}) + \mu^{(b)}_i \ge d^{(b)}_i(1-\bar{q}_i)(\psi_i - \sum_{k\in[K]} x^{\mathrm{A}}_{ik} t^{\mathrm{A}}_{ik}),\ i \in I_{(v)},\ v \in [V],\ b \in [B], \\
& \underline{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}^V_+,\ \bar{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}^V_+,\ \boldsymbol{\mu}^{(b)} \in \mathbb{R}^I_+,\ b \in [B].
\end{aligned}$$

Further, considering the minimization over $(\boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}}, \boldsymbol{x}^{\mathrm{D}}, \boldsymbol{x}^{\mathrm{A}})$ in the above program, (R–JD) is equivalent to

$$
\begin{aligned}
\min_{\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}},\boldsymbol{\phi},\boldsymbol{\psi},\underline{\boldsymbol{\lambda}}^{(b)},\bar{\boldsymbol{\lambda}}^{(b)},\boldsymbol{\mu}^{(b)},\zeta} \quad & \zeta \\
\text{s.t.} \quad & \zeta \geq \sum_{i\in[I]} d_i^{(b)}\big(\rho_1\phi_i + (1+\rho_2)\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\big) + \sum_{v\in[V]}\sum_{i\in I_{(v)}} d_i^{(b)}(\bar{p}_{(v)}\bar{\lambda}_{(v)}^{(b)} - \underline{p}_{(v)}\underline{\lambda}_{(v)}^{(b)}) + \sum_{i\in[I]}\mu_i^{(b)}, b\in[B] \\
& d_i^{(b)}\sum_{v\in[V]}\mathbb{1}_{\{i\in I_{(v)}\}}(\bar{\lambda}_{(v)}^{(b)} - \underline{\lambda}_{(v)}^{(b)}) + \mu_i^{(b)} \geq d_i^{(b)}(1-\bar{q}_i)(\psi_i - \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}),\ i\in I_{(v)}, v\in[V], b\in[B] \\
& \underline{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}_+^V, \bar{\boldsymbol{\lambda}}^{(b)} \in \mathbb{R}_+^V, \boldsymbol{\mu}^{(b)} \in \mathbb{R}_+^I, b\in[B] \\
& \phi_i = \min\{\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\},\ i\in[I] \\
& \psi_i = \min\{\breve{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\},\ i\in[I] \\
& \boldsymbol{y}^{\mathrm{D}} \in \mathcal{Y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}} \in \mathcal{Y}^{\mathrm{A}} \\
& \boldsymbol{x}^{\mathrm{D}} \in \mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}), \boldsymbol{x}^{\mathrm{A}} \in \mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}).
\end{aligned}
\tag{EC.14}
$$

Note that the above program is nonlinear due to the nonlinear constraints for ϕ_i and ψ_i, $i\in[I]$. For linearization purpose, we introduce a large enough positive number M and a set of auxiliary binary decision variables $(\nu_{i1},\nu_{i2},\nu_{i3},\nu_{i4})\in\{0,1\}^4$, $i\in[I]$. In view of the minimization program, the equality constraint for ϕ_i is equivalent to the set of constraints

$$
\begin{cases}
\phi_i \geq \sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}} - M\nu_{i1}, \\
\phi_i \geq \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} - M\nu_{i2}, \\
\nu_{i1} + \nu_{i2} = 1.
\end{cases}
\tag{EC.15}
$$

This is because when $\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}} < \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}$, we have $\nu_{i1}=0$ and $\nu_{i2}=1$, so the only binding constraint is $\phi_i \geq \sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}$. Since ϕ_i strives to achieve the smallest possible value, we have $\phi_i = \sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}$. Likewise, we can show that $\phi_i = \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}$ when $\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} < \sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}$, which concludes that the equality $\phi_i = \min\{\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$ holds. By a similar argument, the equality constraint for ψ_i is equivalent to the set of constraints

$$
\begin{cases}
\psi_i \geq \breve{u}_i - M\nu_{i3}, \\
\psi_i \geq \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} - M\nu_{i4}, \\
\nu_{i3} + \nu_{i4} = 1.
\end{cases}
\tag{EC.16}
$$

From the above linearization procedure, it can be readily seen that the constant M can be any value no less than $\max\{\max_{i\in[I],j\in[J]} t_{ij}^{\mathrm{D}}, \max_{i\in[I],k\in[K]} t_{ik}^{\mathrm{A}}\}$. By substituting (EC.15) and (EC.16) into (EC.14), we obtain the MILP reformulation (17) for (R–JD). The proof for Theorem 1 is thus complete. □

Proof of Proposition 4. By Theorem 1, $\boldsymbol{O}_\star \triangleq (\boldsymbol{y}^{\mathrm{D}}_\star, \boldsymbol{y}^{\mathrm{A}}_\star, \boldsymbol{x}^{\mathrm{D}}_\star, \boldsymbol{x}^{\mathrm{A}}_\star, \boldsymbol{\psi}_\star, \boldsymbol{\phi}_\star)$ is the optimal solution to the following program

$$
\begin{aligned}
\min_{\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{A}},\boldsymbol{\psi},\boldsymbol{\phi}} \quad & \sum_{i\in[I]} d_i^{(b^\star)} \left(p_i^\star(1-\bar{q}_i)\psi_i + \rho_1\phi_i + (1-p_i^\star\bar{q}_i+\rho_2) \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \right) \\
\text{s.t.} \quad & \phi_i \geq \textstyle\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}} - M\nu_{i1},\ \phi_i \geq \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} - M\nu_{i2},\ i\in[I], \\
& \psi_i \geq \check{u}_i - M\nu_{i3},\ \psi_i \geq \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} - M\nu_{i4},\ i\in[I], \\
& \nu_{i1}+\nu_{i2}=1,\ \nu_{i3}+\nu_{i4}=1,\ i\in[I], \\
& (\nu_{i1},\nu_{i2},\nu_{i3},\nu_{i4}) \in \{0,1\}^4,\ i\in[I], \\
& \boldsymbol{y}^{\mathrm{D}} \in \mathcal{Y}^{\mathrm{D}},\ \boldsymbol{y}^{\mathrm{A}} \in \mathcal{Y}^{\mathrm{A}}, \\
& \boldsymbol{x}^{\mathrm{D}} \in \mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}),\ \boldsymbol{x}^{\mathrm{A}} \in \mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}).
\end{aligned}
$$

By the first four lines of constraints, we have $\phi_i \geq \min\{\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$, and $\psi_i \geq \min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$. Since the above program is a minimization program, we must have $\phi_i = \min\{\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$ for $i\in\mathcal{M}$, and $\psi_i = \min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$ for $i\in\mathcal{M}'$ at optimality. For the same reason, under a given deployment $(\boldsymbol{y}^{\mathrm{D}}, \boldsymbol{y}^{\mathrm{A}})$, the response times $\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}$ and $\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}$ must attain their respective minimal values in regions $i\in\mathcal{M}$ at optimality. This is achieved by assigning the nearest drone base and ambulance base to those regions, i.e., $x_{i,j(i)}^{\mathrm{D}} = x_{i,k(i)}^{\mathrm{A}} = 1$, $i\in\mathcal{M}'$. However, because the values of $\sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}$, $\sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}$ and ϕ_i do not contribute to the optimal value for regions with $d_i^{(b^\star)} = 0$, and the values of ψ_i do not contribute to the optimal value for regions with $d_i^{(b^\star)} p_i^\star = 0$, they can violate the above equalities without affecting the optimality of $\boldsymbol{O}_\star$. □

Proof of Proposition 5. In an iteration step, if the scenario subset $\mathcal{D}_S$ already contains the set $\{\boldsymbol{d}(\boldsymbol{O}) : \boldsymbol{O}\in O\}$, then the lower and upper bounds must coincide in that iteration, because we have

$$
\max_{\boldsymbol{d}\in\mathcal{D}_S} G(\boldsymbol{d}, \boldsymbol{O}_S) \leq \max_{\boldsymbol{d}\in\mathcal{D}} G(\boldsymbol{d}, \boldsymbol{O}_S) = G(\boldsymbol{d}(\boldsymbol{O}_S), \boldsymbol{O}_S) \leq \max_{\boldsymbol{d}\in\mathcal{D}_S} G(\boldsymbol{d}, \boldsymbol{O}_S).
$$

In the above, the first inequality is because $\mathcal{D}_S \subseteq \mathcal{D}$; the second equality is by the definition of the worst-case scenario $\boldsymbol{d}(\boldsymbol{O}_S)$; the last inequality is because $\boldsymbol{d}(\boldsymbol{O}_S) \in \{\boldsymbol{d}(\boldsymbol{O}) : \boldsymbol{O}\in O\} \subseteq \mathcal{D}_S$ by our assumption. Since in each iteration step (excluding the first step where we initialize $\mathcal{D}_S$ as a singleton randomly drawn from $\mathcal{D}$), a different worst-case demand scenario is added into $\mathcal{D}_S$, it takes at most $S+1$ iterations for $\mathcal{D}_S$ to contain the set $\{\boldsymbol{d}(\boldsymbol{O}) : \boldsymbol{O}\in O\}$. Therefore, the algorithm must converge in no greater than $S+1$ iterations. □

EC.2 On the Construction of Uncertainty Set for Alert Response Probability

In this section we give supporting examples and Propositions for our construction of the uncertainty set (13). Proofs of propositions that appear in this section are relegated to Appendix EC.2.5.

EC.2.1 The Benefit of Region Aggregation and Tradeoff in the Choice of Cluster Number

To illustrate the benefit and tradeoff in capturing the uncertainty of $\boldsymbol{p}$ from an aggregate level (i.e., cluster of multiple demand regions), we consider a simplified ambulance deployment problem involving two regions. Distance unit is scaled so that region 1 is located at $x = x_1 \triangleq 0$ and region 2 is located at $x = x_2 \triangleq 1$. We are interested in deploying a single ambulance on the location $x = \ell \in [0,1]$ to minimize the demand-weighted survival churn rate (4). For simplicity, we assume that (i) the two regions have the same demand scaled to 1, i.e., $d_1^* = d_2^* = 1$, (ii) conditional on the scenario that at least one responder responds to the alert, the responder arrives immediately, i.e., $\mathbb{Q}^*[\tilde{u}_1 = 0] = \mathbb{Q}^*[\tilde{u}_2 = 0] = 1$, and (iii) time unit is scaled so that ambulance's travel time is proportional to the travel distance, i.e., $t^{\mathrm{A}}_{1,\ell} = 1 - t^{\mathrm{A}}_{2,\ell} = \ell$. Then the optimization problem of minimizing (4) becomes

$$\min_{\ell\in[0,1]} \quad [p_1^* \min\{0,\ell\} + \rho_1\ell + (1 - p_1^* + \rho_2)\ell] + [p_2^* \min\{0,1-\ell\} + \rho_1(1-\ell) + (1 - p_2^* + \rho_2)(1-\ell)]\,. \tag{EC.17}$$

Because $\boldsymbol{p}^*$ is generally unknown, we adopt the robust deployment approach to solve

$$\min_{\ell\in[0,1]} \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d}^*)} \quad [p_1 \min\{0,\ell\} + \rho_1\ell + (1 - p_1 + \rho_2)\ell] + [p_2 \min\{0,1-\ell\} + \rho_1(1-\ell) + (1 - p_2 + \rho_2)(1-\ell)]\,,$$

which is equivalent to

$$\min_{\ell\in[0,1]} \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d}^*)} \quad (1-p_1)\ell + (1-p_2)(1-\ell). \tag{EC.18}$$

In the above, $1 - p_i$ is the alert neglect probability in region $i \in \{1,2\}$, which is also the probability that the ambulance's travel time becomes relevant as no responder arrives earlier.

Depending on the choice of the cluster number V, there are two possible formulations for the set $\mathcal{P}$:

$$\mathcal{P}(\boldsymbol{d}^*; V=1) = \left\{(p_1,p_2)\middle|\frac{p_1+p_2}{2} \in [\underline{p},\bar{p}]\right\} \quad \text{and} \quad \mathcal{P}(\boldsymbol{d}^*; V=2) = \left\{(p_1,p_2)\middle|p_1 \in [\underline{p}_1,\bar{p}_1], p_2 \in [\underline{p}_2,\bar{p}_2]\right\},$$

where $0 \le \underline{p} \le \bar{p} \le 1$ and $0 \le \underline{p}_v \le \bar{p}_v \le 1$ for $v = 1,2$. The set $\mathcal{P}(\boldsymbol{d}^*; V=2)$ corresponds to the scenario where we estimate a confidence interval for each p_i without aggregation, while the set $\mathcal{P}(\boldsymbol{d}^*; V=1)$ corresponds to the scenario where regions 1 and 2 are clustered into a group, and we estimate a confidence interval for the average response probability in the cluster level. We are interested in the deployments produced by the robust program (EC.18) under these two uncertainty sets, and how they compare to the true optimal deployment that solves the nominal program (EC.17) assuming perfect knowledge of $\boldsymbol{p}^*$.

Let ℓ^* be the true optimal deployment that solves the nominal program (EC.17), and $Z^* \triangleq Z(\ell^*)$ be the corresponding optimal value. Let $\ell_1^\star$ and $\ell_2^\star$ be the respective optimal solutions for the robust program (EC.18) under uncertainty sets $\mathcal{P}(\boldsymbol{d}^*; V=1)$ and $\mathcal{P}(\boldsymbol{d}^*; V=2)$, and let $Z(\ell_i^\star)$ be the objective value of the nominal program (EC.17) when the deployment $\ell_i^\star$, $i \in \{1,2\}$, is used. The following Proposition characterizes the relationship between the deployments and also their resulting objective values.

PROPOSITION EC.1. *We have (i)* $\ell^* = 0$ *if* $p_1^* < p_2^*$ *and* $\ell^* = 1$ *if* $p_2^* < p_1^*$, *(ii)* $\ell_1^\star = 1/2$ *as long as* $\underline{p} < \bar{p}$, *and (iii)* $\ell_2^\star = 0$ *if* $\underline{p}_1 < \underline{p}_2$ *and* $\ell_2^\star = 1$ *if* $\underline{p}_2 < \underline{p}_1$. *Furthermore, when* $(\underline{p}_1 - \underline{p}_2)(p_1^* - p_2^*) > 0$, *we have* $Z^* = Z(\ell_2^\star) < Z(\ell_1^\star)$. *When* $(\underline{p}_1 - \underline{p}_2)(p_1^* - p_2^*) < 0$, *we have* $Z^* < Z(\ell_1^\star) < Z(\ell_2^\star)$.

It can be seen from Proposition EC.1 that when the confidence intervals are accurate enough to faithfully reflect the true values, i.e., $(\underline{p}_1 - \underline{p}_2)(p_1^* - p_2^*) > 0$, the robust deployment $\ell_2^\star$ under the uncertainty set $\mathcal{P}(\boldsymbol{d}^*; V = 2)$ is also optimal in the nominal program (EC.17). However, when the confidence intervals are not accurate enough, i.e., $(\underline{p}_1 - \underline{p}_2)(p_1^* - p_2^*) > 0$, the robust deployment $\ell_1^\star$ under the uncertainty set $\mathcal{P}(\boldsymbol{d}^*; V = 1)$ would have better performance than $\ell_2^\star$, i.e., $Z(\ell_1^\star) < Z(\ell_2^\star)$. Hence, there is a tradeoff between the quality and quantity of information when choosing between the uncertainty sets $\mathcal{P}(\boldsymbol{d}^*; V = 1)$ and $\mathcal{P}(\boldsymbol{d}^*; V = 2)$. When the data is abundant for an accurate estimation of the confidence intervals, the set $\mathcal{P}(\boldsymbol{d}^*; V = 2)$ is preferred because it incorporates more side information on the spatial distribution of the alert response probability. This side information can lead to an optimal solution when it is also accurate. However, when the data is limited and the estimated confidence intervals are not accurate, the set $\mathcal{P}(\boldsymbol{d}^*; V = 1)$ is preferred because at that time the side information in $\mathcal{P}(\boldsymbol{d}^*; V = 2)$ would be misleading and can lead to an inferior deployment. In conclusion, we should only increase the cluster number when each of the resulting confidence interval is accurate enough.

EC.2.2 The Benefit of Aggregation Based on Spatial Proximity

For a given cluster number V (its determination is discussed in Appendix EC.2.1), we argue that the aggregation scheme $\{I_{(v)}, v \in [V]\}$ should be designed based on spatial proximity of the demand regions. To better support this argument, we consider a simplified ambulance deployment problem, similar to the one in Appendix EC.2.1, but involving three regions. Specifically, region 1, 2 and 3 is located at $x = x_1 \triangleq 0$, $x = x_2 \triangleq 1/2$ and $x = x_3 \triangleq 1$, respectively. We are interested in deploying a single ambulance on the location $x = \ell \in [0, 1]$. Similar to the problem setting in Appendix EC.2.1, we assume that (i) the three regions have the same demand scaled to 1, i.e., $d_1^* = d_2^* = d_3^* = 1$, (ii) conditional on the scenario that at least one responder responds to the alert, the responder arrives immediately, i.e., $\mathbb{Q}^*[\tilde{u}_1 = 0] = \mathbb{Q}^*[\tilde{u}_2 = 0] = \mathbb{Q}^*[\tilde{u}_3 = 0] = 1$, and (iii) ambulance's travel time is proportional to the travel distance, i.e., $t_{1,\ell}^{\mathrm{A}} = \ell$, $t_{2,\ell}^{\mathrm{A}} = |1/2 - \ell|$, and $t_{3,\ell}^{\mathrm{A}} = 1 - \ell$.

To fully expose the effect of the uncertainty set (13) on the deployment decision ℓ, we focus on minimizing the demand-weighted intervention time to CPR, which is directly affected by the alert response probability. It is equivalent to setting the weight parameters $\rho_1 = \rho_2 = 0$ in the expression of the survival churn rate (4). Hence, we are interested in the optimization problem

$$\min_{\ell \in [0,1]} \quad [p_1^* \min\{0, \ell\} + (1 - p_1^*)\ell] + [p_2^* \min\{0, |1/2 - \ell|\} + (1 - p_2^*)|1/2 - \ell|] + [p_3^* \min\{0, 1 - \ell\} + (1 - p_3^*)(1 - \ell)].$$

Similar to Appendix EC.2.1, the robust version of the above program can be shown to be equivalent to

$$\min_{\ell\in[0,1]} \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d}^*)} \quad (1-p_1)\ell + (1-p_2)\Big|\frac{1}{2}-\ell\Big| + (1-p_3)(1-\ell). \tag{EC.19}$$

For a pre-specified cluster number $V=2$, if we cluster demand regions based on their spatial proximity, regions 1 and 2 will be clustered together, while region 3 forms a cluster itself, i.e., $I_{(1)}=\{1,2\}$, $I_{(2)}=\{3\}$. The case where $I_{(1)}=\{1\}$, $I_{(2)}=\{2,3\}$ is purely symmetric, and it suffices to discuss one of them. On the other hand, if we cluster demand regions not based on their spatial proximity, we will have regions 1 and 3 in a cluster, while region 2 forms a cluster itself, i.e., $I_{(1)}=\{1,3\}$, $I_{(2)}=\{2\}$. We denote the corresponding uncertainty sets in the two scenarios by $\mathcal{P}^{\mathrm{S}}(\boldsymbol{d}^*)$ and $\mathcal{P}^{\mathrm{NS}}(\boldsymbol{d}^*)$, respectively, which is given by

$$\mathcal{P}^{\mathrm{S}}(\boldsymbol{d}^*) = \left\{(p_1,p_2,p_3)\Big|\frac{p_1+p_2}{2}\in[\underline{p}_{(1)},\bar{p}_{(1)}], p_3\in[\underline{p}_{(2)},\bar{p}_{(2)}]\right\}$$

and

$$\mathcal{P}^{\mathrm{NS}}(\boldsymbol{d}^*) = \left\{(p_1,p_2,p_3)\Big|\frac{p_1+p_3}{2}\in[\underline{p}_{(1)},\bar{p}_{(1)}], p_2\in[\underline{p}_{(2)},\bar{p}_{(2)}]\right\},$$

where $0\le \underline{p}_{(v)}\le \bar{p}_{(v)}\le 1$ for $v=1,2$. Let $\ell^\star_{\mathrm{S}}$ and $\ell^\star_{\mathrm{NS}}$ be the respective optimal solutions for the robust program (EC.19) under uncertainty sets $\mathcal{P}^{\mathrm{S}}(\boldsymbol{d}^*)$ and $\mathcal{P}^{\mathrm{NS}}(\boldsymbol{d}^*)$, respectively. The following proposition characterizes the behavior of $\ell^\star_{\mathrm{S}}$ and $\ell^\star_{\mathrm{NS}}$.

PROPOSITION EC.2. *We have (i-I)* $\ell^\star_{\mathrm{S}}=1/2$ *if* $\underline{p}_{(1)}\in\left[0,1/2-\underline{p}_{(2)}/2\right]$, *(i-II)* $\ell^\star_{\mathrm{S}}=1/4$ *if* $\underline{p}_{(1)}\in\left(1/2-\underline{p}_{(2)}/2,1/2+\underline{p}_{(2)}/2\right)$, *(i-III)* $\ell^\star_{\mathrm{S}}=1$ *if* $\underline{p}_{(1)}\in\left[1/2+\underline{p}_{(2)}/2,1\right]$, *and (ii)* $\ell^\star_{\mathrm{NS}}=1/2$ *in all scenarios.*

It can be seen that the uncertainty set $\mathcal{P}^{\mathrm{NS}}(\boldsymbol{d}^*)$ leads to a static deployment $\ell^\star_{\mathrm{NS}}=1/2$ regardless of the estimated confidence intervals. In contrast, under the uncertainty set $\mathcal{P}^{\mathrm{S}}(\boldsymbol{d}^*)$, the robust program is able to produce different deployments according to the estimated confidence intervals. This is because the uncertainty of $\boldsymbol{p}$ is restricted to a set of geographically concentrated regions, which makes the confidence interval more informative on the spatial distribution of $\boldsymbol{p}$. In conclusion, the aggregation scheme $\{I_{(v)}, v\in[V]\}$ should ideally be designed based on the spatial proximity of demand regions, which allows the confidence intervals to better reflect the spatial distribution of $\boldsymbol{p}$, and reduces the conservatism of the robust program.

EC.2.3 Over-Conservatism Caused by Using the Estimated Demand Rate

An alternative way to construct the set $\mathcal{P}$ is to depend it on the estimated demand rate $\hat{\boldsymbol{d}}$ instead of each demand scenario $\boldsymbol{d}^{(b)}$. In this way, the set $\mathcal{P}$ becomes

$$\mathcal{P}(\hat{\boldsymbol{d}}) \triangleq \left\{\boldsymbol{p}\in[0,1]^I\Big|\frac{\sum_{i\in I_{(v)}}\hat{d}_i p_i}{\sum_{i\in I_{(v)}}\hat{d}_i}\in[\underline{p}_{(v)},\bar{p}_{(v)}], v\in[V]\right\}. \tag{EC.20}$$

Although the quantity $\sum_{i\in I_{(v)}}\hat{d}_i p_i/\sum_{i\in I_{(v)}}\hat{d}_i$ has a natural interpretation being the proportion of OHCA incidents with alert response under the historical demand, it does not take into consideration the outer layer of

robustness over $\boldsymbol{d} \in \mathcal{D}$, which can lead to the over-conservatism of the robust program (R–JD). To illustrate, suppose the V clusters are non-overlapping and consider a demand uncertainty set $\mathcal{D}'$, where the only information is the total demand rate being $D_{(v)}$ in each cluster $v \in [V]$, i.e., $\mathcal{D}' \triangleq \{\boldsymbol{d} \in \mathbb{R}_+^I \,|\, \sum_{i \in I_{(v)}} d_i = D_{(v)}, v \in [V]\}$. The set $\mathcal{D}'$ is in line with the fact that each demand scenario in the scenario-based uncertainty set $\mathcal{D}$ has the same total demand rate $\sum_{v \in [V]} D_{(v)}$. The following Proposition shows that when the set $\mathcal{P}(\hat{\boldsymbol{d}})$ is used, the number of OHCA incidents with alert response in the worst-case scenario, i.e., $\min_{\boldsymbol{d} \in \mathcal{D}', \boldsymbol{p} \in \mathcal{P}(\hat{\boldsymbol{d}})} \sum_{i \in [I]} d_i p_i$, can be zero even if each of the confidence lower bound $\underline{p}_{(v)}$ is strictly positive.

PROPOSITION EC.3. *When the V clusters are non-overlapping and each of the confidence lower bound $\underline{p}_{(v)}$ is not too large, i.e., satisfying $0 \le \underline{p}_{(v)} \le 1 - \min_{i \in I_{(v)}} \{\hat{d}_i\} / \sum_{i \in I_{(v)}} \hat{d}_i$ for $v \in [V]$, we have $\min_{\boldsymbol{d} \in \mathcal{D}', \boldsymbol{p} \in \mathcal{P}(\hat{\boldsymbol{d}})} \sum_{i \in [I]} d_i p_i = 0$ even if $\underline{p}_{(v)} > 0$ for $v \in [V]$.*

In contrast, if we construct the set $\mathcal{P}(\boldsymbol{d})$ based on each demand scenario in $\mathcal{D}'$, i.e., the set

$$\mathcal{P}(\boldsymbol{d}) \triangleq \left\{ \boldsymbol{p} \in [0,1]^I \,\middle|\, \frac{\sum_{i \in I_{(v)}} d_i p_i}{\sum_{i \in I_{(v)}} d_i} \in [\underline{p}_{(v)}, \bar{p}_{(v)}], v \in [V] \right\} \quad \text{for} \quad \boldsymbol{d} \in \mathcal{D}',$$

the worst-case value, i.e., $\min_{\boldsymbol{d} \in \mathcal{D}', \boldsymbol{p} \in \mathcal{P}(\boldsymbol{d})} \sum_{i \in [I]} d_i p_i$, must be positive unless $\underline{p}_{(v)} = 0$ for all $v \in [V]$.

EC.2.4 On the Nonemptiness of the Uncertainty set

We take for granted that $\sum_{i \in I_{(v)}} d_i^{(b)} > 0$ for all $v \in [V]$ and $b \in [B]$, since it is very unlikely that all regions in a cluster have zero demand rate when the cluster number are properly chosen and when the demand scenarios are properly generated. When the V clusters are non-overlapping, each p_i belongs to exactly one of the clusters $I_{(v)}, v \in [V]$. Therefore, a feasible $\boldsymbol{p}$ can be constructed by letting $p_i = p_v \in [\underline{p}_{(v)}, \bar{p}_{(v)}]$ for all $i \in I_{(v)}$, $v \in [V]$. Under this construction, we have $\sum_{i \in I_{(v)}} d_i p_i / \sum_{i \in I_{(v)}} d_i = p_v \in [\underline{p}_{(v)}, \bar{p}_{(v)}]$ for all $v \in [V]$. When some clusters are overlapping, certain p_i appears simultaneously in the constraint $\sum_{i \in I_{(v)}} d_i p_i / \sum_{i \in I_{(v)}} d_i = p_v \in [\underline{p}_{(v)}, \bar{p}_{(v)}]$ for multiple clusters. Emptiness of $\boldsymbol{p}(\boldsymbol{d})$ can happen only if the uncertainty intervals $[\underline{p}_{(v)}, \bar{p}_{(v)}]$ contradict with each other significantly. Such contradiction can be avoided by estimating the uncertainty intervals through a consistent estimator applied to the historical data in each cluster, e.g., the Wilson score method.

EC.2.5 Proofs of Propositions in Appendix EC.2

In this section we provide proofs for Propositions EC.1, EC.2 and EC.3 that appear in Appendix EC.2.

Proof of Proposition EC.1. For notation convenience, we prove the Proposition under the transformation that $p_i \leftarrow 1 - p_i$ and $p_i^* \leftarrow 1 - p_i^*, i = 1, 2$. Correspondingly, the confidence intervals are transformed by $\underline{p} \leftarrow 1 - \bar{p}$, $\bar{p} \leftarrow 1 - \underline{p}$ and $\underline{p}_i \leftarrow 1 - \bar{p}_i$, $\bar{p}_i \leftarrow 1 - \underline{p}_i, i = 1, 2$. Under this transformation, the nominal program becomes

$$\min_{\ell \in [0,1]} \quad p_1^* \ell + p_2^* (1 - \ell),$$

and the robust program (EC.18) becomes

$$\min_{\ell\in[0,1]} \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d}^*)} \quad p_1\ell + p_2(1-\ell).$$

For the nominal program, it can be readily seen that $\ell^* = 0$ if $p_1^* > p_2^*$ and $\ell^* = 1$ if $p_2^* > p_1^*$, and also $Z^* = \min\{p_1^*, p_2^*\}$. On the other hand, the robust program with $\mathcal{P}(\boldsymbol{d}^*) = \mathcal{P}(\boldsymbol{d}^*; V = 2)$ is equivalent to

$$\min_{\ell\in[0,1]} \quad \bar{p}_1\ell + \bar{p}_2(1-\ell),$$

because there exists a worst-case $\boldsymbol{p}^\star$ solving the inner maximization program that satisfies $p_i^\star = \bar{p}_i$ for $i = 1, 2$. Therefore, we have $\ell_2^\star = 0$ if $\bar{p}_1 > \bar{p}_2$ and $\ell_2^\star = 1$ if $\bar{p}_2 > \bar{p}_1$. Consequently, we have $Z(\ell_2^\star) = p_2^*$ if $\bar{p}_1 > \bar{p}_2$ and $Z(\ell_2^\star) = p_1^*$ if $\bar{p}_2 > \bar{p}_1$.

When $\mathcal{P}(\boldsymbol{d}^*) = \mathcal{P}(\boldsymbol{d}^*; V = 1)$, the robust program is equivalent to

$$\begin{aligned} \min_{\ell\in[0,1]} \max_{p_2\in[0,1]} \quad & (2\bar{p} - p_2)\ell + p_2(1-\ell) \\ \text{s.t.} \quad & 2\bar{p} - 1 \le p_2 \le 2\bar{p}, \end{aligned}$$

because there exists a worst-case $\boldsymbol{p}^\star$ solving the inner maximization program that satisfies $(p_1^\star + p_2^\star)/2 = \bar{p}$. Hence, we can substitute p_1 by $2\bar{p} - p_2$, and add the constraint that $2\bar{p} - p_2 \in [0, 1]$ to obtain the above robust program. Since the objective function can be rewritten as $2\bar{p}\ell + (1-2\ell)p_2$, we have $p_2^\star = \min\{1, 2\bar{p}\}$ when $\ell \in [0, 1/2]$, and the objective function becomes $(2\bar{p} - 2\min\{1, 2\bar{p}\})\ell + \min\{1, 2\bar{p}\}$, and hence $\ell_1^\star = 1/2$ (since $2\bar{p} - 2\min\{1, 2\bar{p}\} \le 0$). When $\ell \in [1/2, 1]$, we have $p_2^\star = \max\{0, 2\bar{p} - 1\}$, and thus the objective function becomes $(2\bar{p} - 2\max\{0, 2\bar{p} - 1\} + 2)\ell + \max\{0, 2\bar{p} - 1\}$, and hence $\ell_1^\star = 1/2$ (since $2\bar{p} - 2\max\{0, 2\bar{p} - 1\} + 2 \ge 0$). Combining the above, we have $\ell_1^\star = 1/2$ and $Z(\ell_1^\star) = (p_1^* + p_2^*)/2$.

Therefore, when $(\bar{p}_1 - \bar{p}_2)(p_1^* - p_2^*) > 0$, we have $Z^* = Z(\ell_2^\star) = \min\{p_1^*, p_2^*\} < Z(\ell_1^\star) = (p_1^* + p_2^*)/2$. When $(\bar{p}_1 - \bar{p}_2)(p_1^* - p_2^*) < 0$, we have $Z^* = \min\{p_1^*, p_2^*\} < Z(\ell_1^\star) = (p_1^* + p_2^*)/2 < Z(\ell_2^\star) = \max\{p_1^*, p_2^*\}$. The proof is complete when we reverse the transformation that $\underline{p} \leftarrow 1 - \bar{p}$, $\bar{p} \leftarrow 1 - \underline{p}$ and $p_i^* \leftarrow 1 - p_i^*$, $\underline{p}_i \leftarrow 1 - \bar{p}_i$, $\bar{p}_i \leftarrow 1 - \underline{p}_i$, $i = 1, 2$. □

Proof of Proposition EC.2. For notation convenience, we prove the proposition under the transformation that $p_i \leftarrow 1 - p_i$ for $i = 1, 2, 3$. Correspondingly, the confidence intervals are transformed by $\underline{p}_{(v)} \leftarrow 1 - \bar{p}_{(v)}$ for $v = 1, 2$. Under this transformation, the robust program (EC.19) becomes

$$\min_{\ell\in[0,1]} \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{d}^*)} \quad p_1\ell + p_2\Big|\frac{1}{2} - \ell\Big| + p_3(1-\ell). \tag{EC.21}$$

We first prove part (ii) of the proposition. When $\mathcal{P}(\boldsymbol{d}^*) = \mathcal{P}^{\mathrm{NS}}(\boldsymbol{d}^*)$, the above program can be shown to be equivalent to

$$\begin{aligned} \min_{\ell\in[0,1]} \max_{p_3} \quad & (2\bar{p}_{(1)} - p_3)\ell + \bar{p}_{(2)}\Big|\frac{1}{2} - \ell\Big| + p_3(1-\ell) \\ \text{s.t.} \quad & \max\{0, 2\bar{p}_{(1)} - 1\} \le p_3 \le \min\{1, 2\bar{p}_{(1)}\}, \end{aligned} \tag{EC.22}$$

because there exists a worst-case $\boldsymbol{p}^\star$ solving the inner maximization program that satisfies $(p_1^\star+p_3^\star)/2=\bar{p}_{(1)}$ and $p_2^\star=\bar{p}_{(2)}$, so we can substitute p_1 by $2\bar{p}_{(1)}-p_3$, p_2 by $\bar{p}_{(2)}$, and add the constraint that $2\bar{p}_{(1)}-p_3\in[0,1]$ to obtain the above robust program. We next discuss the case $\ell\in[0,1/2]$ and $\ell\in[1/2,1]$ separately.

- When $\ell\in[0,1/2]$, the objective function of (EC.22) becomes $(1-2\ell)p_3+2\bar{p}_{(1)}\ell+\bar{p}_{(2)}(1/2-\ell)$. Because $1-2\ell\geq 0$, we have $p_3^\star=\min\{1,2\bar{p}_{(1)}\}$, and thus the objective function becomes $(2\bar{p}_{(1)}-2\min\{1,2\bar{p}_{(1)}\}-\bar{p}_{(2)})\ell+\min\{1,2\bar{p}_{(1)}\}+\bar{p}_{(2)}/2$. Because $2\bar{p}_{(1)}-2\min\{1,2\bar{p}_{(1)}\}-\bar{p}_{(2)}\leq 0$, we have $\ell_{\mathrm{NS}}^\star=1/2$.
- When $\ell\in[1/2,1]$, the objective function of (EC.22) becomes $(1-2\ell)p_3+2\bar{p}_{(1)}\ell+\bar{p}_{(2)}(\ell-1/2)$. Because $1-2\ell\leq 0$, we have $p_3^\star=\max\{0,2\bar{p}_{(1)}-1\}$, and thus the objective function becomes $(2\bar{p}_{(1)}-2\max\{0,2\bar{p}_{(1)}-1\}+\bar{p}_{(2)})\ell+\max\{0,2\bar{p}_{(1)}-1\}-\bar{p}_{(2)}/2$. Because $2\bar{p}_{(1)}-2\max\{0,2\bar{p}_{(1)}-1\}+\bar{p}_{(2)}\geq 0$ (by the fact that $\bar{p}_{(v)}\in[0,1], v=1,2$), we have $\ell_{\mathrm{NS}}^\star=1/2$.

Combing both cases, we conclude that $\ell_{\mathrm{NS}}^\star=1/2$ regardless of how the confidence intervals take value.

We next prove part (i-I) to (i-III) of the Proposition. When $\mathcal{P}(\boldsymbol{d}^*)=\mathcal{P}^{\mathrm{S}}(\boldsymbol{d}^*)$, the program (EC.21) can be shown to be equivalent to

$$\begin{aligned}\min_{\ell\in[0,1]}\max_{p_2}\quad & (2\bar{p}_{(1)}-p_2)\ell+p_2\Big|\frac{1}{2}-\ell\Big|+\bar{p}_{(2)}(1-\ell)\\ \text{s.t.}\quad & \max\{0,2\bar{p}_{(1)}-1\}\leq p_2\leq\min\{1,2\bar{p}_{(1)}\},\end{aligned}\tag{EC.23}$$

because there exists a worst-case $\boldsymbol{p}^\star$ solving the inner maximization program that satisfies $(p_1^\star+p_2^\star)/2=\bar{p}_{(1)}$ and $p_3^\star=\bar{p}_{(2)}$, so we can substitute p_1 by $2\bar{p}_{(1)}-p_2$, p_3 by $\bar{p}_{(2)}$, and add the constraint that $2\bar{p}_{(1)}-p_2\in[0,1]$ to obtain the above robust program. We next discuss the case $\ell\in[0,1/4],\ell\in[1/4,1/2]$ and $\ell\in[1/2,1]$ separately. When $\ell\leq 1/2$, the objective function of (EC.23) becomes $(1/2-2\ell)p_2+2\bar{p}_{(1)}\ell+\bar{p}_{(2)}(1-\ell)$.

- When $\ell\in[0,1/4]$, we have $p_2^\star=\min\{1,2\bar{p}_{(1)}\}$ (since $1/2-2\ell\geq 0$), and thus the objective function becomes $(2\bar{p}_{(1)}-2\min\{1,2\bar{p}_{(1)}\}-\bar{p}_{(2)})\ell+\min\{1,2\bar{p}_{(1)}\}/2+\bar{p}_{(2)}$. Because $2\bar{p}_{(1)}-2\min\{1,2\bar{p}_{(1)}\}-\bar{p}_{(2)}\leq 0$, we have $\ell_{\mathrm{S}}^\star=1/4$ and the corresponding optimal value of (EC.23) is $Z^\star=\bar{p}_{(1)}/2+3\bar{p}_{(2)}/4$.
- When $\ell\in[1/4,1/2]$, we have $p_2^\star=\max\{0,2\bar{p}_{(1)}-1\}$ (since $1/2-2\ell\leq 0$), and thus the objective function becomes $(2\bar{p}_{(1)}-2\max\{0,2\bar{p}_{(1)}-1\}-\bar{p}_{(2)})\ell+\max\{0,2\bar{p}_{(1)}-1\}/2+\bar{p}_{(2)}$. The sign of $\Delta\triangleq 2\bar{p}_{(1)}-2\max\{0,2\bar{p}_{(1)}-1\}-\bar{p}_{(2)}$ can be either positive or negative, so we further consider the following two cases:

  —When $2\bar{p}_{(1)}\leq\bar{p}_{(2)}$ or $2\bar{p}_{(1)}\geq 2-\bar{p}_{(2)}$, we have $\Delta\leq 0$ and thus $\ell_{\mathrm{S}}^\star=1/2$. The corresponding optimal value is $Z^\star=\bar{p}_{(1)}-\max\{0,2\bar{p}_{(1)}-1\}/2+\bar{p}_{(2)}/2=\min\{1/2,\bar{p}_{(1)}\}+\bar{p}_{(2)}/2$.

  —When $\bar{p}_{(2)}\leq 2\bar{p}_{(1)}\leq 2-\bar{p}_{(2)}$, we have $\Delta\leq 0$ and thus $\ell_{\mathrm{S}}^\star=1/4$. The corresponding optimal value is $\bar{p}_{(1)}/2+3\bar{p}_{(2)}/4$.
- When $\ell\in[1/2,1]$, the objective function of (EC.23) becomes $-p_2/2+2\bar{p}_{(1)}\ell+\bar{p}_{(2)}(1-\ell)$. Hence, we have $p_2^\star=\max\{0,2\bar{p}_{(1)}-1\}$, and thus the objective function becomes $(2\bar{p}_{(1)}-\bar{p}_{(2)})\ell-\max\{0,2\bar{p}_{(1)}-1\}/2+\bar{p}_{(2)}$. We further consider the following two cases:

— When $2\bar{p}_{(1)} \le \bar{p}_{(2)}$, we have $\ell_S^\star = 1$ and the corresponding optimal value is $Z^\star = 2\bar{p}_{(1)}$, since $2\bar{p}_{(1)} \le \bar{p}_{(2)} \le 1$.

— When $2\bar{p}_{(1)} \ge \bar{p}_{(2)}$, we have $\ell_S^\star = 1/2$ and the corresponding optimal value is $\min\{1/2, \bar{p}_{(1)}\} + \bar{p}_{(2)}/2$.

In what follows, we combine the above results on the optimal deployment $\ell_S^\star$ and objective value $Z^\star$ over the three intervals $\ell \in [0, 1/4], \ell \in [1/4, 1/2]$ and $\ell \in [1/2, 1]$, which allows us to prove the part (i-I) to part (i-III) of the Proposition.

- When $2\bar{p}_{(1)} \le \bar{p}_{(2)}$, we have $\ell_S^\star([0,1/4]) = 1/4$, $\ell_S^\star([1/4,1/2]) = 1/2$ and $\ell_S^\star([1/2,1]) = 1$. Correspondingly, we have $Z^\star([0,1/4]) = \bar{p}_{(1)}/2 + 3\bar{p}_{(2)}/4$, $Z^\star([1/4,1/2]) = \min\{1/2, \bar{p}_{(1)}\} + \bar{p}_{(2)}/2$ and $Z^\star([1/2,1]) = 2\bar{p}_{(1)}$. Since when $2\bar{p}_{(1)} \le \bar{p}_{(2)}$, it can be shown that $Z^\star([1/2,1]) \le Z^\star([1/4,1/2])$ and $Z^\star([1/2,1]) \le Z^\star([0,1/4])$, we conclude that the optimal deployment under this parameter condition is $\ell_S^\star = 1$.
- When $\bar{p}_{(2)} \le 2\bar{p}_{(1)} \le 2 - \bar{p}_{(2)}$, we have $\ell_S^\star([0,1/4]) = \ell_S^\star([1/4,1/2]) = 1/4$ and $\ell_S^\star([1,1/2]) = 1/2$. Correspondingly, we have $Z^\star([0,1/4]) = Z^\star([1/4,1/2]) = \bar{p}_{(1)}/2 + 3\bar{p}_{(2)}/4$, and $Z^\star([1/2,1]) = \min\{1/2, \bar{p}_{(1)}\} + \bar{p}_{(2)}/2$. Since when $\bar{p}_{(2)} \le 2\bar{p}_{(1)} \le 2 - \bar{p}_{(2)}$, it can be shown that $Z^\star([0,1/4]) = Z^\star([1/2,1/4]) \le Z^\star([1/2,1])$, we conclude that the optimal deployment under this parameter condition is $\ell_S^\star = 1/4$.
- When $2\bar{p}_{(1)} \ge 2 - \bar{p}_{(2)}$, we have $\ell_S^\star([0,1/4]) = 1/4$ and $\ell_S^\star([1/4,1/2]) = \ell_S^\star([1/2,1]) = 1/2$. Correspondingly, we have $Z^\star([0,1/4]) = \bar{p}_{(1)}/2 + 3\bar{p}_{(2)}/4$ and $Z^\star([1/2,1]) = Z^\star([1/4,1/2]) = \min\{1/2, \bar{p}_{(1)}\} + \bar{p}_{(2)}/2$. Since when $2\bar{p}_{(1)} \ge 2 - \bar{p}_{(2)}$, it can be shown that $Z^\star([1/2,1/4]) = Z^\star([1/2,1]) \le Z^\star([0,1/4])$, we conclude that the optimal deployment under this parameter condition is $\ell_S^\star = 1/2$.

The proof is complete when we reverse the transformation that $\underline{p}_{(v)} \leftarrow 1 - \bar{p}_{(v)}$ for $v = 1, 2$. □

Proof of Proposition EC.3. Since $0 \le \underline{p}_{(v)} \le 1 - \min_{i \in I_{(v)}}\{\hat{d}_i\} / \sum_{i \in I_{(v)}} \hat{d}_i$, we can always construct a feasible $\boldsymbol{p}^* \in \mathcal{P}(\hat{\boldsymbol{d}})$ such that for each cluster $v \in [V]$, there exists at least one $i^v \in I_{(v)}$ satisfying $p^*_{i^v} = 0$. Based on such construction, we further construct a feasible $\boldsymbol{d}^* \in \mathcal{D}'$ by letting $d_i^* = D_{(v)}$ for $i = i^v, v \in [V]$ and zero otherwise. It follows that

$$0 \le \min_{\boldsymbol{d} \in \mathcal{D}', \boldsymbol{p} \in \mathcal{P}(\hat{\boldsymbol{d}})} \sum_{i \in [I]} d_i p_i \le \sum_{i \in [I]} d_i^* p_i^* = \sum_{v \in [V]} d^*_{i^v} p^*_{i^v} = 0,$$

where the first inequality is because both $\boldsymbol{d}$ and $\boldsymbol{p}$ must be nonnegative for them to be feasible; the second inequality is due to the feasibility of $\boldsymbol{d}^*$ and $\boldsymbol{p}^*$; the third equality is because $d_i^* = 0$ for all $i \notin \{i^1, i^2, \ldots, i^V\}$; the last equality is because $p_i^* = 0$ for all $i \in \{i^1, i^2, \ldots, i^V\}$. The proof is complete. □

EC.3 On the Independence of Dispersion Upper Bounds

The phenomenon that the optimal solution of a distributionally robust optimization program finally becomes independent of certain parameters in the ambiguity set has also been observed in Mohajerin Esfahani and Kuhn (2018) (Remark 6.7 therein). They observe this phenomenon when the objective function is convex in

the decision variables and the ambiguity set is Wasserstein-distance-based. In contrast, our objective function is concave in the decision variables, and our ambiguity set (11) is condition-wise and mean-dispersion-based.

From a theoretical perspective, the closed form in Proposition 1 will continue to be independent of ε_i for a wide range of dispersion constraints that impose upper bound on deviation from the conditional mean value, e.g., the variance dispersion constraint $\mathbb{E}_{\mathbb{P}_i}\left[\Gamma_{\mathrm{v}}(\tilde{u}_i;\check{u}_i)\middle|\tilde{u}_i\le\tau\right]\le\varepsilon_i$ with $\Gamma_{\mathrm{v}}(\tilde{u}_i;\check{u}_i)\triangleq(\tilde{u}_i-\check{u}_i)^2$. This is formalized in the following Proposition.

PROPOSITION EC.4. *For any dispersion function $\Gamma(\cdot;\check{u}_i):\mathbb{R}\mapsto\mathbb{R}_+$ that satisfies $\Gamma(\check{u}_i;\check{u}_i)=0$, the maximization problem over $\mathbb{Q}\in\mathcal{Q}$ in Proposition 1 admits the same closed-form expression if the absolute deviation constraint in the ambiguity set $\mathcal{Q}$ defined by (11), i.e., the constraint $\mathbb{E}_{\mathbb{P}_i}\left[|\tilde{u}_i-\check{u}_i|\middle|\tilde{u}_i\le\tau\right]\le\varepsilon_i$, is replaced by $\mathbb{E}_{\mathbb{P}_i}\left[\Gamma(\tilde{u}_i;\check{u}_i)\middle|\tilde{u}_i\le\tau\right]\le\varepsilon_i$.*

Proof of Proposition EC.4. To prove Proposition EC.4, it suffices to show that Proposition 1 still holds when the ambiguity set $\mathcal{Q}$ is not specified by (11), but by

$$\mathcal{Q}\triangleq\left\{\mathbb{Q}\in\mathcal{P}_0(\mathbb{R}_+^I)\middle|\begin{array}{l}\Pi_{\tilde{u}_i}\mathbb{Q}=\mathbb{P}_i,i\in[I]\\ \mathbb{P}_i[\tilde{u}_i>\tau]\in[\underline{q}_i,\bar{q}_i]\\ \mathbb{E}_{\mathbb{P}_i}\left[\tilde{u}_i\middle|\tilde{u}_i\le\tau\right]=\check{u}_i\\ \mathbb{E}_{\mathbb{P}_i}\left[\Gamma(\tilde{u}_i;\check{u}_i)\middle|\tilde{u}_i\le\tau\right]\le\varepsilon_i\end{array}\right\},$$

where $\Gamma(\cdot;\check{u}_i):\mathbb{R}\mapsto\mathbb{R}^+$ is a mean-value-centered dispersion function that satisfies $\Gamma(\check{u}_i;\check{u}_i)=0$. By carefully examining the proof of Proposition 1, we know that the above task boils down to proving that the equality

$$\max_{\mathbb{P}\in\hat{\mathcal{F}}_i^{\mathrm{c}}}\mathbb{E}_{\mathbb{P}}[\min\{\tilde{u}_i,\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}]=\min\{\check{u}_i,\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}\tag{EC.24}$$

still holds, where

$$\hat{\mathcal{F}}_i^{\mathrm{c}}\triangleq\left\{\mathbb{P}\in\mathcal{P}(\mathbb{R}_+)\middle|\begin{array}{l}\tilde{u}_i\sim\mathbb{P}\\ \mathbb{P}[\tilde{u}_i\in[0,\tau]]=1\\ \mathbb{E}_{\mathbb{P}}[\tilde{u}_i]=\check{u}_i\\ \mathbb{E}_{\mathbb{P}}[\Gamma(\tilde{u}_i;\check{u}_i)]\le\varepsilon\end{array}\right\}.$$

is the new conditional ambiguity set. To prove this, note that the function $g(u_i)\triangleq\min\{u_i,\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}$ is concave. Hence by the probabilistic version of Jensen's equality, we have

$$\mathbb{E}_{\mathbb{P}}[\min\{\tilde{u}_i,\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}]\le\min\{\mathbb{E}_{\mathbb{P}}[\tilde{u}_i],\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}$$

for every $\mathbb{P}\in\hat{\mathcal{F}}_i^{\mathrm{c}}$. Since every $\mathbb{P}\in\hat{\mathcal{F}}_i^{\mathrm{c}}$ satisfies $\mathbb{E}_{\mathbb{P}}[\tilde{u}_i]=\check{u}_i$, RHS of the above inequality is a fixed value, i.e., $\min\{\check{u}_i,\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}$. Therefore, we obtain

$$\max_{\mathbb{P}\in\hat{\mathcal{F}}_i^{\mathrm{c}}}\mathbb{E}_{\mathbb{P}}[\min\{\tilde{u}_i,\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}]\le\min\{\check{u}_i,\sum_{k\in[K]}x_{ik}^{\mathrm{A}}t_{ik}^{\mathrm{A}}\}.\tag{EC.25}$$

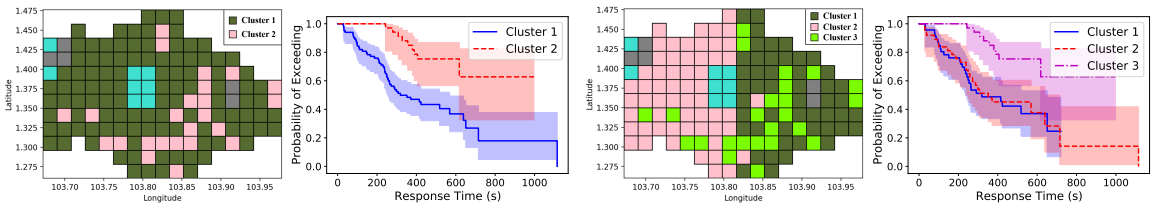
To prove (EC.24), we next show that the other direction of inequality (EC.25) also holds. The idea is to show that there exists a feasible probability distribution $\mathbb{P}^\star \in \hat{\mathcal{F}}_i^{\mathrm{c}}$ that satisfies $\mathbb{E}_{\mathbb{P}^\star}[\min\{\tilde{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}] = \min\{\check{u}_i, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\}$. This is shown by letting $\mathbb{P}^\star$ to be the Dirac measure placing unit mass on $\check{u}_i$. Since Γ is mean-value-centered, we have

$$\mathbb{E}_{\mathbb{P}^\star}[\Gamma(\tilde{u}_i;\check{u}_i)] = \Gamma(\check{u}_i;\check{u}_i) = 0 \le \varepsilon_i,$$

which implies that $\mathbb{P}^\star$ is feasible in $\hat{\mathcal{F}}_i^{\mathrm{c}}$. Because we have proved (EC.24), the proof is complete. □

EC.4 Ambiguity Set Estimation for Alert Response Time Distribution

From the left and middle panels of Figure 2, we can see that the total number of incidents with alert responses in each region is small. In conjunction with the 49% censoring proportion on average, Kaplan-Meier estimation for the alert response time distribution would be highly variable. To decrease statistical error, we use spectral clustering (Von Luxburg 2007) to cluster neighboring regions with similar $\Pi_{\tilde{u}_i}\mathbb{Q}$, and then aggregate data to estimate a common Kaplan-Meier curve for each cluster. The key to the clustering is a distance matrix that characterizes the dissimilarity between each pair of regions. In this study, the distance between two regions takes into account both the geographical distance and the dissimilarity between the Kaplan-Meier curves of the two regions. The first distance $D_{i,i'}^{\mathrm{g}}$ is the geographical distance between the centers of region i and i', while the second distance $D_{i,i'}^{\mathrm{c}}$ is an integration of the absolute difference between the two Kaplan-Meier curves over their common support normalized by the length of the common support, where we set $D_{i,i'}^{\mathrm{c}} = 0$ if the number of observed/censored alert response times in either region i or i' is zero. We normalize the two distances by their respective means. The final distance is a sum of the two normalized distances. With the final distance matrix and a pre-specified cluster number V^{RT}, we use the SpectralClustering algorithm under default settings provided by the scikit-learn library to do the clustering (Pedregosa et al. 2011). Based on the historical data, we set the cluster number as $V^{\mathrm{RT}} = 2$ so that the estimated Kaplan-Meier curves are most distinct and their supports cover the interval $[0,\tau]$, as shown in Figure EC.1. For cluster $v \in [V^{\mathrm{RT}}]$, let $\hat{S}_{(v)}(t)$, $\underline{S}_{(v)}(t)$ and $\bar{S}_{(v)}(t)$ denote the Kaplan-Meier curve, its 95%



(a) $V^{\mathrm{RT}} = 2$ (b) $V^{\mathrm{RT}} = 3$

Figure EC.1 The clustering results and estimated Kaplan-Meier curves from the proposed ambiguity set estimation procedure under different cluster numbers.

lower confidence estimate and upper confidence estimate, calculated using the exponential Greenwood formula (Kaplan and Meier 1958), respectively. Letting $\tilde{u}_{(v)}$ be a generic alert response time in cluster v, a point

estimate for $\Pr(\tilde{u}_{(v)} > \tau)$ is $\hat{q}_{(v)} \triangleq \hat{S}_{(v)}(\tau)$, and a 95% confidence interval is $[\underline{q}_{(v)}, \bar{q}_{(v)}]$, where $\underline{q}_{(v)} \triangleq \underline{S}_{(v)}(\tau)$ and $\bar{q}_{(v)} \triangleq \bar{S}_{(v)}(\tau)$. We then truncate and normalize the Kaplan-Meier curve for each cluster at τ as an estimate of the distribution of $[\tilde{u}_{(v)} | \tilde{u}_{(v)} \le \tau]$, based on which we can estimate the conditional mean of $[\tilde{u}_{(v)} | \tilde{u}_{(v)} \le \tau]$, denoted as $\check{u}_{(v)}$. Estimation results for the parameters $\hat{\boldsymbol{q}} \triangleq (\hat{q}_{(1)}, \ldots, \hat{q}_{(V^{\mathrm{RT}})})$, $\bar{\boldsymbol{q}} \triangleq (\bar{q}_{(1)}, \ldots, \bar{q}_{(V^{\mathrm{RT}})})$ and $\check{\boldsymbol{u}} \triangleq (\check{u}_{(1)}, \ldots, \check{u}_{(V^{\mathrm{RT}})})$ based on our alert response data are reported in the first row of Table EC.1. Finally, we set the ambiguity set parameters of Q as $\check{u}_i = \check{u}_{(v)}$ and $\bar{q}_i = \bar{q}_{(v)}$ for $i \in I^{\mathrm{RT}}_{(v)}$ and $v \in [V^{\mathrm{RT}}]$.

Table EC.1 Estimation result for the parameters used in the alert response time distribution ambiguity set Q and the alert response probability uncertainty set $\mathcal{P}$ based on the October 2017 data.

Ambiguity/Uncertainty Set	$\hat{\boldsymbol{p}}$	$\underline{\boldsymbol{p}}$	$\hat{\boldsymbol{q}}$	$\bar{\boldsymbol{q}}$	$\check{\boldsymbol{u}}$ (in seconds)
Q	–	–	(.179, .628)	(.377, .824)	(339, 432)
$\mathcal{P}$	(.383, .516, .348)	(.307, .394, .257)	–	–	–

EC.5 Uncertainty Sets Construction for d and p

To generate demand scenarios in the uncertainty set $\mathcal{D}$, we use KDE with a Gaussian kernel to fit a two-dimensional density for the spatial distribution of OHCA. Following the convention (Silverman 2018), we determine the bandwidth through a 10-fold cross-validation by searching over the candidate set $\{0.1\omega, 0.2\omega, \ldots, 2\omega\}$, where ω is the Silverman's rule-of-thumb bandwidth. The fitted density under the optimal bandwidth $\omega^* \triangleq 0.8\omega$ is shown in Figure EC.2. To generate a demand scenario $\boldsymbol{d}^{(b)}$, we resample $n = 292$ realizations (same size as the October 2017 data) from the fitted density and count the number of realizations in each region. The above procedure is repeated $B = 100$ times to generate 100 demand scenarios $\boldsymbol{d}^{(b)}$, $b \in [100]$.

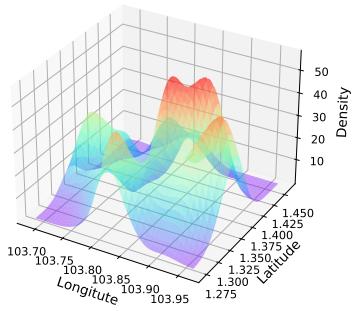


Figure EC.2 Fitted spatial density of OHCA using KDE with the optimal cross-validation bandwidth ω^* based on the October 2017 data, where the goodness of fit is measured by the log-likelihood value.

Next we look at the alert response probability. As argued in Section 4.2, we aggregate the demand regions based on spatial proximity when constructing the uncertainty set $\mathcal{P}$. To achieve this, we do a k-means clustering based on the coordinate of each region center to aggregate the regions into V non-overlapping

clusters. We then estimate $\hat{p}_{(v)}$ as the proportion of OHCA incidents with alert response in cluster $v \in [V]$, and estimate the corresponding 95% confidence interval $[\underline{p}_{(v)}, \bar{p}_{(v)}]$ by the Wilson score method (Wilson 1927). We set the cluster number as $V = 3$ to ensure that the confidence intervals are accurate enough to satisfy $\min_{v\in[V]}\{\underline{p}_{(v)}/\hat{p}_{(v)}\} \geq 0.7$, as discussed in Section 4.2. Estimation results for the parameters $\hat{\boldsymbol{p}} \triangleq (\hat{p}_{(1)}, \ldots, \hat{p}_{(V)})$ and $\underline{\boldsymbol{p}} \triangleq (\underline{p}_{(1)}, \ldots, \underline{p}_{(V)})$ are reported in the second row of Table EC.1. We do not report $\bar{p}_{(v)}$ because it is not used in the robust program (R–JD).

EC.6 Benchmarks: The SAA and Ex Post Models

In this section we give detailed formulations for the SAA model and the ex post model. In essence, these two models share the same formulation except that the SAA model (resp. the ex post model) is constructed based on the training samples (resp. test samples). To present a general formulation, we consider a given set of historical data on OHCA and responder response that contains N samples, which can represent either the set of training data or the set of test data.

From the set of historical data, we extract the information set $\mathcal{F} \triangleq \{(i_n, \Lambda_n^{\text{Accept}}, \lambda_n^{\text{First}}, \hat{u}_n), n \in [N]\}$. In the above, $i_n \in [I]$ is the region that incident $n \in [N]$ happened in; $\Lambda_n^{\text{Accept}} \in \{0,1\}$ is an indicator of alert response, where $\Lambda_n^{\text{Accept}} = 1$ indicates that at least one responder responded to the alert in incident n; $\Lambda_n^{\text{First}} \in \{0,1\}$ is an indicator of first arrival, where $\Lambda_n^{\text{First}} = 1$ indicates that the responder arrived earlier that the ambulance in incident n (and thus implies $\Lambda_n^{\text{Accept}} = 1$); $\hat{u}_n$ is the alert response time in incident n, which is observed if and only if $\Lambda_n^{\text{First}} = 1$. If $\hat{u}_n$ is censored/unobserved, we treat its value as τ when constructing the model, which ensures that ambulance arrives earlier under any feasible deployment $(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}})$. Then for a given deployment $(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}})$, the three intervention times in incident $n \in [N]$ can be calculated by (1) as

$$\begin{aligned}
\hat{T}_n^{\text{CPR}}(\boldsymbol{y}^{\text{A}}) &\triangleq \lambda_n^{\text{First}} \min\{\hat{u}_n, \sum_{k\in[K]} x^{\text{A}}_{i_n,k} t^{\text{A}}_{i_n,k}\} + (1-\Lambda_n^{\text{First}}) \sum_{k\in[K]} x^{\text{A}}_{i_n,k} t^{\text{A}}_{i_n,k};\\
\hat{T}_n^{\text{DEF}}(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}}) &\triangleq \min\{\sum_{j\in[J]} x^{\text{D}}_{i_n,j} t^{\text{D}}_{ij}, \sum_{k\in[K]} x^{\text{A}}_{ik} t^{\text{A}}_{i_n,k}\};\\
\hat{T}_n^{\text{AC}}(\boldsymbol{y}^{\text{A}}) &\triangleq \sum_{k\in[K]} x^{\text{A}}_{i_n,k} t^{\text{A}}_{i_n,k} + o_i,
\end{aligned}$$

where the assignment $(\boldsymbol{x}^{\text{D}}, \boldsymbol{x}^{\text{A}})$ assign each incident to its nearest drone/ambulance base under the deployment $(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}})$, and in the above expression, we have abbreviated the dependence of $(\boldsymbol{x}^{\text{D}}, \boldsymbol{x}^{\text{A}})$ on $(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}})$ for clarity (see Proposition 4(i) for details). The induced survival rate, according to (2), is given by

$$\hat{s}_n(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}}) = \beta_0 - \beta_1 \hat{T}_n^{\text{CPR}}(\boldsymbol{y}^{\text{A}}) - \beta_2 \hat{T}_n^{\text{DEF}}(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}}) - \beta_3 \hat{T}_n^{\text{AC}}(\boldsymbol{y}^{\text{A}}), \quad n \in [N].$$

Therefore, the optimization model which maximizes the average survival rate can be written as

$$\min_{\boldsymbol{y}^{\text{D}}\in\mathcal{Y}^{\text{D}}, \boldsymbol{y}^{\text{A}}\in\mathcal{Y}^{\text{A}}} \quad \sum_{n\in[N]} \hat{s}_n(\boldsymbol{y}^{\text{D}}, \boldsymbol{y}^{\text{A}}).$$

Using similar techniques that allow us to obtain (3), the above program can be shown to be equivalent to

$$\begin{aligned}\min_{\boldsymbol{y}^{\mathrm{D}},\boldsymbol{y}^{\mathrm{A}},\boldsymbol{x}^{\mathrm{D}},\boldsymbol{x}^{\mathrm{D}}} \quad & \sum_{i\in[I]} \hat{d}_i \left(\rho_1 \min\{ \sum_{j\in[J]} x_{ij}^{\mathrm{D}} t_{ij}^{\mathrm{D}}, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \} + (1 - \hat{p}_i^{\mathrm{First}} + \rho_2) \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}} \right) \\ & + \sum_{i\in[I]} \sum_{n\in\hat{\mathcal{N}}_i} \min\{\hat{u}_n, \sum_{k\in[K]} x_{ik}^{\mathrm{A}} t_{ik}^{\mathrm{A}}\} \\ \text{s.t.} \quad & \boldsymbol{y}^{\mathrm{D}} \in \mathcal{Y}^{\mathrm{D}},\ \boldsymbol{y}^{\mathrm{A}} \in \mathcal{Y}^{\mathrm{A}} \\ & \boldsymbol{x}^{\mathrm{D}} \in \mathcal{X}^{\mathrm{D}}(\boldsymbol{y}^{\mathrm{D}}),\ \boldsymbol{x}^{\mathrm{A}} \in \mathcal{X}^{\mathrm{A}}(\boldsymbol{y}^{\mathrm{A}}),\end{aligned} \tag{EC.26}$$

which is a MILP that can readily be solved using off-the-shelf solvers, e.g., Gurobi. In the above MILP, $\hat{d}_i \triangleq \sum_{n\in[N]} \mathbb{I}_{\{i_n=i\}}$ is the historical demand in region i; $\hat{p}_i^{\mathrm{First}} \triangleq \sum_{n\in[N]} \mathbb{I}_{\left\{i_n=i,\Lambda_n^{\mathrm{First}}=1\right\}}/\hat{d}_i$ is the historical fraction of incidents in region i in which the responder arrived first; $\hat{\mathcal{N}}_i \triangleq \{n\in[N]|\ i_n = i, \Lambda_n^{\mathrm{First}} = 1\}$ is the index set for incidents in which the responder arrived first and happened in region i. When the quantities $\{(\hat{d}_i, \hat{p}_i^{\mathrm{First}}, \hat{\mathcal{N}}_i), i\in[I]\}$ are estimated from the set of training data (resp. test data), formulation (EC.26) essentially becomes the SAA model (resp. ex post model).

References

Shapiro, Alexander. 2001. On duality theory of conic linear problems. *Semi-infinite programming.* Springer, 135-165.

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and Computing,* 17 (4), 395-416.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12 2825-2830.

Kaplan, Edward L, Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* 53 (282), 457-481.

Silverman, Bernard W. 2018. *Density Estimation for Statistics and Data Analysis.* Routledge.

Wilson, Edwin B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association,* 22 (158), 209-212.