The Value of Time- and Location-Commitment for Decentralized Emergency Medical Services

Pieter L. van den Berg
Rotterdam School of Management, Erasmus University, vandenberg@rsm.nl

Andre P. Calmon
Scheller College of Business, Georgia Institute of Technology, andre.calmon@gatech.edu

Andreas K. Gernert
Department of Logistics, Kühne Logistics University, andreas.gernert@klu.org

Stef Lemmens
Rotterdam School of Management, Erasmus University, s.lemmens@rsm.nl

Maria Rabinovich
Flare, maria@rescue.co

Gonzalo Romero
Rotman School of Management, University of Toronto, gonzalo.romero@rotman.utoronto.ca

Problem definition: Emergency medical services (EMS) in many low- and middle-income countries utilize decentralized platforms coordinating independent ambulance providers. However, significant operational challenges arise from uncertainty in provider time availability and unpredictable idle locations. These uncertainties hinder reliable service coverage and negatively impact patient outcomes. Using data from our partner Flare in Nairobi, Kenya, we investigate the relative effectiveness of enhancing provider temporal commitment (time availability) versus spatial commitment (strategic location) to improve system coverage.

Methodology/results: We employ optimization models adapted for ambulance commitment uncertainty, a detailed case study analysis, data-driven simulations, and a game-theoretic model. Our findings quantify a stark "cost of decentralization": the coverage provided by Flare's ~ 340 loosely committed ambulances could potentially be matched by fewer than 15 optimally deployed fully committed units. We find that enhancing spatial commitment generally yields higher marginal returns for improving coverage than solely increasing time availability. Adding just five optimized, location-flexible ambulances increased coverage substantially in simulation (e.g., by $\sim 5\%$ over the baseline fleet) and reduced service variability. Simulations confirm the practical impact of interventions and validate model assumptions, while a game-theoretic model offers generalizable insights; both approaches align in highlighting the significant value of spatial coordination.

Managerial implications: For managers and decision-makers overseeing decentralized EMS platforms, prioritizing strategies that improve spatial coordination offers an efficient path to enhancing service reliability and performance. Actionable strategies include targeted incentives that encourage providers to relocate strategically or deploy a small fleet of location-flexible, platform-controlled units to fill critical coverage gaps. Our framework offers practical tools for managers to identify coverage gaps and assess the potential impact of such interventions in resource-constrained settings, ultimately aiming to enhance emergency response.

Key words: Emergency medical services; platforms; Integer optimization models; low- and middle-income countries; coordination

1. Introduction

Emergency medical services (EMS) in many low- and middle-income countries (LMICs) are often fragmented and under-resourced (Macintyre and Hotchkiss 1999, Thomson 2005, Das and Desai 2017). Innovate platforms such as Flare in Kenya, Red Health and Dial4242 (Khanna 2017) in India have emerged to reshape these systems. These platforms act like an "Uber for ambulances" (Kuo 2016), and play a crucial coordination role, connecting large, decentralized fleets of ambulances operated by numerous independent providers with patients in need of emergency care.

This platform model contrasts sharply with traditional centralized EMS systems in high-income countries, as platforms typically lack direct operational control over their fleet. Furthermore, unlike ride-hailing services, EMS platforms often operate under different objectives, such as ensuring broad service coverage for subscribers (Gernert et al. 2024), which limits the use of mechanisms like dynamic pricing to balance supply and demand.

Many of the operational challenges these platforms face stem from managing the uncertainty of ambulance fleet availability, which is driven by heterogeneity in provider commitment along two critical dimensions. The first is **temporal commitment**: providers vary substantially in their availability for platform-dispatched calls, often balancing platform duties with their own direct calls or other activities. The second is **spatial commitment**: providers exhibit diverse locating behaviors when idle, ranging from returning to fixed bases (e.g., hospitals) to roaming in search of demand, and possess varying willingness or ability to strategically reposition at the platform's request. Adding to these supply-side complexities, platforms also face the inherent demand uncertainty typical of all emergency services with respect to call timing and location.

This heterogeneity, combined with the platform's lack of central control, results in substantial system-wide inefficiencies. Indeed, our analysis starkly quantifies this "cost of decentralization": the level of service coverage provided by the approximately 340 largely uncommitted ambulances coordinated by Flare in Nairobi could *potentially* be matched by fewer than 15 optimally deployed, fully committed ambulances, highlighting the immense potential value locked up in improving coordination and commitment.

This situation presents a key strategic dilemma for platform managers: given limited resources and influence, which dimension of commitment offers a more effective lever for improving system performance? Should platforms prioritize interventions aimed at securing higher *temporal commitment* (e.g., incentivizing providers to be available more often or exclusively) or enhancing *spatial commitment* (e.g., incentivizing providers to locate strategically or adding controllable, location-flexible units)?

We frame our main research question accordingly: *What is the relative effectiveness of improving temporal versus spatial commitment in enhancing coverage for decentralized emergency response platforms?* While the operations of centralized EMS systems (where supply uncertainty is negligible) are well-studied (Bélanger et al. 2019, Brotcorne et al. 2003), and the platform operations literature addresses supply uncertainty mainly via dynamic pricing (which is largely inapplicable in the EMS context) (Cachon et al. 2017, Taylor 2018, Bimpikis et al. 2019), the trade-offs inherent in managing *heterogeneous spatio-temporal commitment* in decentralized EMS, and particularly the *marginal value* of improving commitment along these distinct dimensions, remains largely unexplored.

To address this question comprehensively, we employ a multi-method research design. We ground our analysis in the empirical context of our partner, Flare, utilizing their operational data from Nairobi. Building on established EMS literature (Daskin 1983), we develop quantitative models for coverage evaluation and optimization specifically adapted to incorporate the dual uncertainties of time and location commitment inherent in decentralized platforms. We further validate these models and explore system dynamics using trace-driven simulations based on out-of-sample demand data. Finally, we complement these numerical findings with a stylized game-theoretic model designed to derive generalizable structural insights into provider behavior and commitment trade-offs in equilibrium.

Our primary contribution is the finding that enhancing **spatial commitment**—ensuring ambulances achieve effective strategic positioning to fill coverage gaps—generally yields a significantly higher marginal return for improving system coverage compared to interventions focused solely on increasing **temporal commitment**, particularly within fleets exhibiting mixed commitment levels. Adding even a small number of location-flexible units, directed by optimization models to fill identified geographical coverage gaps, proves highly effective. This approach is especially potent in countering the suboptimal clustering of resources often observed when independent providers select locations based on individual objectives rather than system-wide needs (Marla et al. 2021b). The core actionable insight for managers of these platforms is that prioritizing strategies that improve spatial control and strategic positioning offers an efficient path to enhancing service reliability and coverage across the network.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 details our modeling framework for coverage evaluation and optimization under commitment uncertainty. Section 4 describes the case study context and data analysis for Nairobi, Kenya. Section 5 presents numerical experiments based on the case study, including simulation results for model validation, quantifying the impact of time and location commitment. Section 6 develops and analyzes a stylized game-theoretic model. Finally, Section 7 concludes the paper.

2. Literature

Our work draws from and contributes to the literature on centralized vehicle management, their time and location commitment, and the ride-hailing literature. We position our paper with respect to each one of these streams next.

2.1. Centralized vehicle management

EMS platforms in LMICs try to centralize a system of independent ambulance providers. Gernert et al. (2024) study strategic decisions by such a platform, for example, the acquisition of own ambulances and the reimbursement of provider ambulances. There is extensive literature on the operational decisions of centralized EMS systems (see Bélanger et al. (2019) and Keskinocak and Savva (2020) for recent reviews). Marla et al. (2017) produce efficient algorithms for ambulance location and dynamic redeployment and Marla et al. (2021a) research the phenomenon of ambulance abandonment requests by callers, particularly seen in LMICs, and the resulting management of the fleet. Boutilier and Chan (2020) present an approach to optimize both the location and routing of emergency response vehicles, accounting for travel time and demand uncertainty in LMICs. Complementing Boutilier and Chan (2020), our paper focuses on supply-side uncertainty.

Our paper also relates to the vast literature on centrally planned vehicle fleet management, e.g., He et al. (2017) and Benjaafar et al. (2021). While He et al. (2020), Balseiro et al. (2021), and Hosseini et al. (2024) focus on "real-time" relocation strategies, Taylor (2018) and Bai et al. (2019) focus on optimal pricing strategies. Complementary to these studies, Shu et al. (2013) and Benjaafar et al. (2021) study the optimal day-ahead location of bikes on a ride-sharing platform to maximize the number of services per day. Benjaafar et al. (2021) studies the fleet size's effect on service-probability goals.

While these papers mostly assume that the central planner of the EMS system or fleet has almost perfect control over servers or drivers—and thus has substantial operational flexibility—our paper studies EMS platforms in LMICs and thus examines platforms with limited control and information on the fleet they coordinate. Additionally, within the context of a subscription-based EMS platform we study, demand and supply matching cannot be achieved by demand-side pricing.

2.2. Location commitment

For several decades, ambulance fleet (re)positioning has been one of the main areas of study in EMS research (Brotcorne et al. 2003, McLay and Mayorga 2013, Nasrollahzadeh et al. 2018). Real-time repositioning has been studied in Alanis et al. (2013), Sudtachat et al. (2016), van Barneveld (2016), and van Barneveld et al. (2017). These papers make use of compliance tables to inform ambulance providers where ambulances should be located depending on specific times of day and

the number of ambulances available. In these papers, the decision maker has full control over the ambulances. In contrast, we focus on settings where the decision lacks this control either entirely or at least partially, which is the most common situation in LMICs.

Daskin (1983)'s MEXCLP model provided a seminal framework to study the optimal location of ambulances that have a fixed probability of being able to serve a call, which has been extensively built upon, e.g., Brotcorne et al. (2003), van den Berg and Aardal (2015), and El Itani et al. (2019). We extend this literature stream by considering decentralized EMS in which some ambulances may roam around searching for demand, some unwilling to change their locations, while others could be repositioned.

2.3. Time commitment

Several studies integrate the time (un)availability of vehicles in a platform or centralized EMS setting. For example, Enayati et al. (2018) incorporate workload restrictions for EMS providers for real-time ambulance redeployment. Their workload constraint dynamically controls the total active time of ambulances. McLay and Mayorga (2013) incorporates equity constraints—specifically, allocating an ambulance to one patient makes that ambulance unavailable for other patients—in Markov decision processes. For bike-sharing systems, Kabra et al. (2020) determine the likelihood of finding a bike nearby. They consider six time-windows in one day and model the average bike availability for a given location and time-window. In contrast to these studies, our work focuses on decentralized EMS platforms where time availability is not centrally controlled but driven by heterogeneous and often unpredictable ambulance provider behavior. We model this uncertainty probabilistically and incorporate it into our optimization and simulation model to evaluate its impact on coverage.

2.4. Ride hailing platforms

The challenge of balancing supply and demand through driver relocations and pricing strategies is a central theme in the ride-hailing literature (Cachon et al. 2017, Bimpikis et al. 2019, Afèche et al. 2023, Jin et al. 2023, Chen et al. 2024). Both Afèche et al. (2023) and Bimpikis et al. (2019) examine equilibrium outcomes in a platform-driver-rider game. Bimpikis et al. (2019) focuses on spatial pricing—how platforms set fares and driver compensation across different origins and destinations—while Afèche et al. (2023) compares equilibrium outcomes when the platform has full versus no control over driver repositioning under homogeneous pricing. Chen et al. (2024) develop a stochastic model to highlight the limitations of static pricing policies when the platform can enforce real-time relocations. A common assumption in this literature is that drivers are homogeneous. Complementing this, we study systems where drivers differ along two key dimensions—time and spatial commitment—and analyze the value that different driver types bring to the system.

3. Model

This section develops the modeling framework that we use to evaluate and improve ambulance coverage for EMS platforms operating with partially committed decentralized fleets, as commonly found in LMICs. Coverage is defined as the fraction of requests for which Flare can respond within their target of 15 minutes. First, in Section 3.1, we introduce a model to estimate the expected coverage provided by a given fleet, considering uncertainties in the availability and location of ambulances. Second, in Section 3.2, we extend this framework to formulate an optimization problem for the deployment of additional location-flexible ambulances to maximize the increase in overall coverage. As EMS platforms in LMICs typically lack real-time data about the location and availability of the ambulances, making operational relocactions ineffective, our focus is on tactical planning.

3.1. Modeling Coverage Evaluation

A primary challenge for EMS platforms coordinating independent providers in LMICs is the inherent uncertainty surrounding fleet resources. Unlike traditional centralized systems, platforms often lack full control over when ambulances are available (time commitment) and where they are located when idle (spatial commitment). We first model the expected coverage provided by a given set of ambulances that are coordinated by the platform to quantify system performance under these conditions.

We categorize the existing fleet based on these two key dimensions of commitment. Let K be the set of all ambulances coordinated by the platform. We partition K into the following four types:

- K_1: Time-committed and location-committed ambulances.
- K_2: Time-uncommitted and location-committed ambulances.
- K_3: Time-committed and location-uncommitted ambulances.
- K_4: Time-uncommitted and location-uncommitted ambulances.

Following standard practice in EMS modeling, we discretize the service region into a set of demand locations I. Each demand location $i \in I$ represents a zone with a demand weight d_i. We also define a set of potential discrete locations J where ambulances might be stationed or found when idle. Note that I and J do not necessarily coincide. To determine if an ambulance at location $j \in J$ can serve demand at location $i \in I$ within the target response time, we define the set $J_i \subseteq J$. This set contains all ambulance locations $j \in J$ from which the response time to the demand point i (including a pre-trip delay and the travel time) is less than or equal to a predefined threshold (e.g., 15 minutes, as discussed in Section 4).

A key feature of platforms such as Flare, particularly in their early stages or in markets with many providers, is a relatively low call volume channeled *through the platform* compared to the

total number of ambulances. Flare's operational data confirm that simultaneous calls requiring dispatch from the platform are scarce (further validated by simulation in Section 5.3). Therefore, for this baseline coverage evaluation, we assume that an ambulance will not be busy serving another *platform* call if available and located appropriately. This implies that a single available ambulance within the response time threshold suffices to cover a demand location in this model.

Although platform-induced congestion is low, uncertainty arises from ambulance external activities and location preferences. We model this uncertainty using two parameters: the **availability probability** q_k and the **location probability** p_{jk}. For each ambulance $k \in K$, q_k denotes the probability that it is available for platform dispatch when an incident occurs, capturing *time-commitment* (for $k \in K_1 \cup K_3$, $q_k = 1$; for $k \in K_2 \cup K_4$, $0 < q_k < 1$). For each ambulance $k \in K$ and location $j \in J$, p_{jk} denotes the probability that k is situated at j when idle and available, capturing *spatial commitment* (for $k \in K_1 \cup K_2$, $p_{jk} = 1$ for a single j; for $k \in K_3 \cup K_4$, p_{jk} follows a distribution that models roaming behavior, detailed in Section 4; we assume $\sum_{j \in J} p_{jk} = 1$ for all k).

Given these parameters, we calculate the probability, $s_i(K)$, that a demand point $i \in I$ is covered by the fleet K. Coverage occurs if at least one ambulance $k \in K$ is available (q_k) and is located within the response time threshold ($\sum_{j \in J_i} p_{jk}$). Assuming the independence of the availability and location decisions between ambulances, the probability that a specific ambulance k can cover the demand point i is $q_k \times (\sum_{j \in J_i} p_{jk})$. The probability that demand point i is *not* covered by ambulance k is $1 - (q_k \times \sum_{j \in J_i} p_{jk})$. Therefore, the probability that *no* ambulance in the fleet K covers point i is $\prod_{k \in K} \left(1 - \left(q_k \times \sum_{j \in J_i} p_{jk}\right)\right)$. The coverage probability $s_i(K)$ is the complement:

$$s_i(K) = 1 - \prod_{k \in K} \left(1 - \left(q_k \times \sum_{j \in J_i} p_{jk}\right)\right). \tag{1}$$

This probabilistic approach provides a tractable way to estimate baseline coverage potential, explicitly incorporating the primary sources of uncertainty (time and location commitment) inherent in decentralized EMS platforms. The overall expected coverage for the entire region, $s(K)$, is then calculated as the demand-weighted average of the individual location coverage probabilities:

$$s(K) = \sum_{i \in I} d_i \times s_i(K) = \sum_{i \in I} d_i \times \left(1 - \prod_{k \in K} \left(1 - \left(q_k \times \sum_{j \in J_i} p_{jk}\right)\right)\right). \tag{2}$$

Many of the underlying assumptions here, such as independence between ambulances, the use of static probabilities for tactical planning, and initially ignoring system congestion for baseline evaluation, are common starting points in the EMS location modeling literature (e.g., Daskin 1983,

Brotcorne et al. 2003, Bélanger et al. 2019) and are often needed for developing tractable models used in practice.

Assumptions and Limitations of Coverage Evaluation Model: This model provides a valuable estimate for tactical planning but relies on several assumptions, including the aforementioned independence, static probabilities (q_k, p_{jk}), and low platform-induced congestion. It also treats coverage as binary based on the threshold and simplifies complex ambulance behaviors. Despite these limitations, which mean it does not capture real-time fluctuations or detailed dynamics, it effectively quantifies the impact of time and spatial commitment uncertainty on potential coverage, highlighting areas of systemic weakness.

3.2. Modeling Coverage Optimization

While Section 3.1 evaluates the coverage provided by an existing, potentially uncoordinated fleet K, platforms may seek interventions to improve performance. One strategy is to add *location-flexible* ambulances, which the platform can direct to specific idle locations to fill coverage gaps. This section formulates an optimization model to determine the optimal static locations for a predetermined number of such location-flexible ambulances to maximize the *increase* in overall expected coverage.

The target for these newly deployed flexible ambulances is the demand currently *uncovered* by the existing fleet K. We define the **residual demand** at location i as $(1 - s_i(K))d_i$, representing the demand at i that the existing fleet is expected to miss. The optimization aims to cover as much of this residual demand as possible.

To model potential platform interventions for improving coverage, we focus on adding location-flexible ambulances, as strategically directing ambulance placement is a key lever available to platforms. We consider two types of flexible additions: time-committed units (type K_5), representing perhaps newly acquired dedicated resources or fully incentivized partners, and time-uncommitted units (type K_6), representing partners willing to relocate strategically while maintaining some external operational commitments. These represent realistic options for platform expansion. Let C be the number of type K_5 ambulances ($q_k = 1$) and U be the number of type K_6 ambulances ($q_k = q$, where $0 < q < 1$ is an assumed common availability probability) to be optimally located.

Our model adapts the structure of the classic Maximum Expected Covering Location Problem (MEXCLP) (Daskin 1983). However, a key distinction lies in the interpretation of the probability parameter q. In traditional MEXCLP applied to centralized EMS, unavailability often arises from ambulances being busy with other calls (an endogenous system state). Here, the unavailability $1 - q$ for type K_6 ambulances stems primarily from external commitments, making q an *exogenous*

parameter representing the probability of being available *to the platform*. This context aligns well, perhaps even better than the traditional one, with the MEXCLP's mathematical treatment of availability probability as an exogenous input parameter.

The optimization model determines the optimal placement for these new ambulances. We define $x_j^c \in \mathbb{N}$ as the number of new time-committed, location-flexible ambulances (type K_5) assigned to location $j \in J$, and $x_j^u \in \mathbb{N}$ as the number of new time-uncommitted, location-flexible ambulances (type K_6) assigned to location j. To track the coverage contribution from the uncommitted additions, we use binary variables $y_{ik} \in \{0,1\}$, where $y_{ik} = 1$ signifies that demand point $i \in I$ is within the response threshold of at least k of these newly added time-uncommitted ambulances. Finally, the variable $y_i \in [0,1]$ captures the resulting fraction of the residual demand at point i that is covered by the combined deployment of these C committed and U uncommitted new ambulances.

$$\text{maximize} \quad \sum_{i \in I} (1 - s_i(K)) d_i y_i \tag{3a}$$

$$\text{subject to} \quad \sum_{j \in J} x_j^c = C \tag{3b}$$

$$\sum_{j \in J} x_j^u = U \tag{3c}$$

$$\sum_{j \in J_i} x_j^u \geq \sum_{k=1}^{U} y_{ik}, \qquad \forall i \in I \tag{3d}$$

$$y_i \leq \sum_{j \in J_i} x_j^c + \sum_{k=1}^{U} q(1-q)^{k-1} y_{ik}, \qquad \forall i \in I \tag{3e}$$

$$x_j^c, x_j^u \in \mathbb{N}, \qquad \forall j \in J \tag{3f}$$

$$y_{ik} \in \{0,1\}, \qquad \forall i \in I, k \in \{1,2,...,U\} \tag{3g}$$

$$0 \leq y_i \leq 1, \qquad \forall i \in I. \tag{3h}$$

The objective function (3a) maximizes the total additional coverage gained, expressed as the sum across all demand points i of the residual demand $(1 - s_i(K))d_i$ multiplied by the fraction y_i covered by the new deployment. The constraints (3b) and (3c) are resource constraints that ensure that precisely C committed and U uncommitted new ambulances are assigned to potential locations J. The relationship between the number of uncommitted ambulances located within the range of a demand point i (i.e. $\sum_{j \in J_i} x_j^u$) and the coverage indicators y_{ik} is enforced by constraint (3d); this structure ensures that if m such ambulances cover point i, then $y_{i1}, ..., y_{im}$ are correctly set to 1.

The core calculation occurs in constraint (3e), which determines the fraction of residual demand covered, y_i. If any new *committed* ambulance covers point i ($\sum_{j \in J_i} x_j^c \geq 1$), y_i can reach its maximum value of 1. Otherwise, y_i is limited by the probability that at least one of the covering

uncommitted ambulances is actually available. This probability is calculated using the standard marginal contribution approach from MEXCLP literature: the term $q(1-q)^{k-1}$ represents the probability that the k-th covering uncommitted ambulance (indicated by $y_{ik}=1$) is the first one found to be available. Summing these marginal probabilities yields the total probability of finding at least one available uncommitted ambulance among those covering i. Finally, constraints (3f)-(3h) define the domains for the decision variables. The formulation results in a mixed-integer linear program.

Assumptions and Limitations of Coverage Optimization Model: This optimization finds the best *static* locations based on *expected* residual demand. It does not account for the costs of establishing or incentivizing ambulances at these locations, nor the operational complexities of managing flexible units. It assumes the platform can perfectly assign and enforce these locations for the flexible units. The objective is solely focused on maximizing the expected coverage increase, neglecting other potential objectives like equity or response time variance.

4. Case Description and Data Analysis

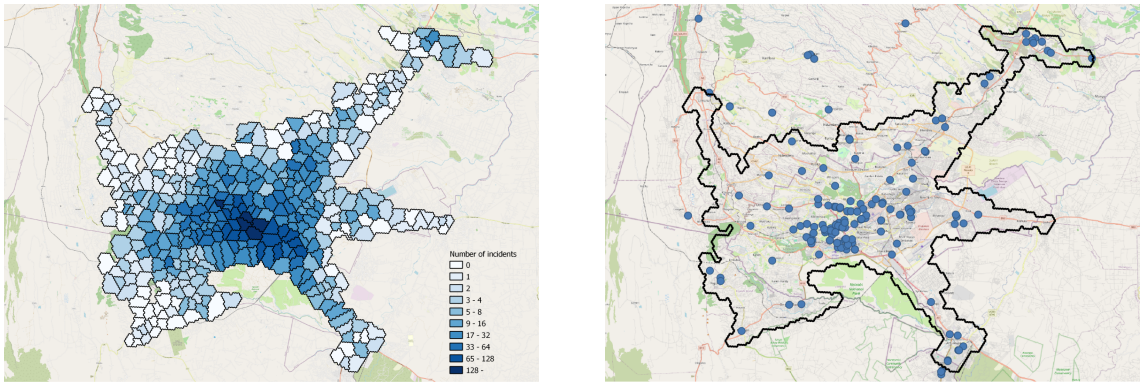
We now detail the application of our modeling framework (Section 3) to the operations of our industry partner, Flare, in Nairobi, Kenya. Nairobi features a fragmented EMS landscape with numerous independent ambulance providers displaying varying degrees of commitment. This context presents a compelling case study for evaluating the impact of fleet coordination challenges prevalent in many LMICs. Flare operates as an EMS platform, aggregating demand and coordinating a large, decentralized fleet it does not own. A key operational challenge identified by Flare is the low and unpredictable time and location commitment of many providers, which significantly complicates efforts to ensure reliable service coverage for their subscribers across the region.

This context makes Nairobi an ideal setting for applying our models to evaluate coverage under commitment uncertainty (Section 3.1) and optimize potential interventions using location-flexible resources (Section 3.2). Furthermore, the subscription-based model increases the need for broad geographical coverage, unlike purely pay-per-ride systems, which may focus only on high-density areas. Although the sheer number of available ambulances (~ 340) might seem sufficient, Flare recognized that low levels of commitment create significant coverage gaps, motivating this study on the value of coordination and commitment. Overcoming data scarcity, common in LMICs, required combining Flare's operational data with publicly available sources, as detailed below.

Service Region and Discretization. Our analysis focuses on Nairobi, the capital and largest city of Kenya. To apply our spatial models, we discretized the service region based on the H3 hexagonal hierarchical spatial index used by Uber Movement. We adopted this discretization because it allows us to leverage publicly available, well-calibrated travel time estimates from Uber Movement

for Nairobi (`movement.uber.com`). This divides the region into $I = 400$ discrete demand locations (hexagonal clusters, or hexaclusters), with an average area of approximately 2.94 km^2. The set of potential ambulance locations J is assumed to coincide with the centroids of these hexaclusters.

Demand Estimation. We estimated the relative demand weight d_i for each demand location $i \in I$ using historical incident data provided by Flare. The data covers the period from August 2017 to August 2021, during which Flare responded to 6,840 incidents, with 6,370 occurring within the Nairobi boundaries defined by our hexaclusters. We allocated each historical incident to its corresponding hexacluster based on recorded coordinates. To ensure that all areas have non-zero demand, preventing trivial exclusion from coverage considerations, we divided one unit of demand over all demand locations, adding 1/400th of an incident to each location. Figure 1a illustrates the resulting spatial demand distribution.



(a) Demand distribution (b) Ambulance distribution

Figure 1 **Demand and supply distribution. a) Historical demand distribution based on the period between August 2017 and August 2021. b) Home base of ambulances coordinated by Flare in Nairobi.**

Travel Time Data. The model requires travel times between potential ambulance locations $j \in J$ and demand locations $i \in I$ to determine the coverage sets J_i. We utilized detailed, publicly available travel time data aggregated by hexaclusters from Uber Movement for Nairobi. This dataset provides mean and standard deviation of travel times based on observed trips. To account for dispatch and crew mobilization time, we added a fixed pre-trip delay of 3 minutes to the Uber travel times, consistent with practical observations. A demand point i is considered covered by an ambulance at location j if the total response time (pre-trip delay plus mean travel time from j to i) is less than or equal to Flare's target of 15 minutes. This defines the sets J_i used in equations (1) and (3).

Ambulance Fleet Characteristics. Flare coordinates a large fleet of approximately 340 ambulances in Nairobi, operated by independent entities. Characterizing their commitment is crucial for

parameterizing our model, specifically q_k and p_{jk}.

Time Commitment (q_k): most ambulances on the platform are highly *time-uncommitted*, as they also serve calls made directly to them, as opposed to through the platform, or perform other duties. For our baseline analysis, we assume a uniform availability probability $q_k = q = 0.25$ for all time-uncommitted ambulances ($k \in K_2 \cup K_4$). Time-committed ambulances ($k \in K_1 \cup K_3$) have $q_k = 1$.

Spatial Commitment (p_{jk}): a significant portion of the fleet is owned by hospitals or specific providers with fixed bases; these are modeled as *location-committed* ($k \in K_1 \cup K_2$), with $p_{jk} = 1$ for their designated home base j (identified from Flare data) and $p_{jk} = 0$ otherwise. The remainder of the fleet lacks a fixed mandated location and exhibits roaming behavior. We assume that these *location-uncommitted* ambulances' ($k \in K_3 \cup K_4$) location distribution p_{jk} follows a *demand-searching* behavior; specifically, we set p_{jk} to be proportional to the historical demand that can be covered from location j: p_{jk} proportional to $\sum_{i:j \in J_i} d_i$. Figure 1b shows the distribution of identified home bases.

In summary, by combining Flare's operational data on incidents and ambulance bases with publicly available data for travel times and regional discretization, we parameterized the coverage evaluation and optimization models presented in Section 3. This data allows us to analyze the impact of temporal and spatial commitment in a realistic LMIC platform context as follows.

5. Numerical Experiments

This section leverages the modeling framework (Section 3) and the Nairobi case data (Section 4) to quantify the operational impact of ambulance commitment and evaluate potential improvement strategies. Our goal is to understand the value derived from increasing time and location commitment within a decentralized EMS platform context. The analysis adopts a tactical planning perspective, consistent with the challenges faced by platforms like Flare, which often lack reliable estimates on the day-to-day availability and location of uncommitted units. We first conduct a greenfield analysis (Section 5.1) to establish theoretical benchmarks and isolate the effects of different commitment types. We then analyze scenarios based on Flare's actual fleet and operational realities (Section 5.2) to assess the current system's efficiency and the potential gains from targeted interventions. Finally, Section 5.3 provides model validation by evaluating the the robustness of our results to considering simultaneous calls and stochasticity in demand and ambulance availability through simulation.

5.1. Greenfield Analysis

To understand the fundamental impact of commitment, we begin with a greenfield analysis, assuming no existing fleet. This allows us to establish benchmarks for efficient coverage in Nairobi and to

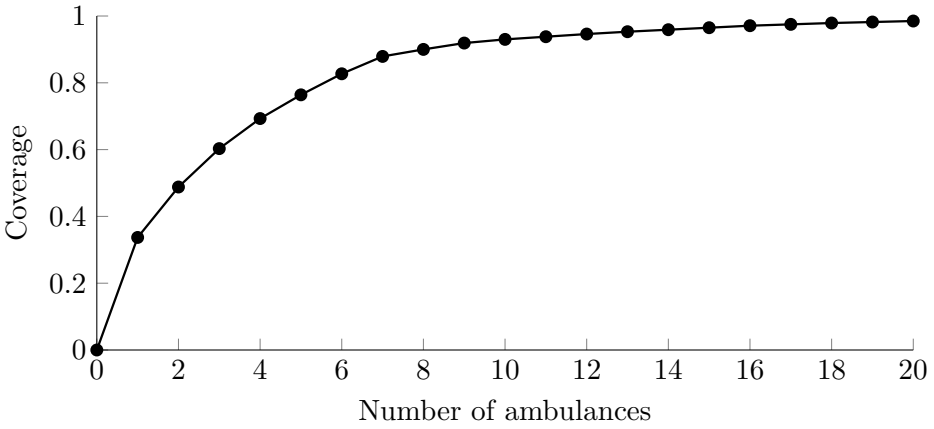


Figure 2 **Coverage for a different number of fully controllable ambulances in a greenfield scenario with a response time target of 15 minutes.**

isolate the value contributed by time and location commitment, independent of the complexities of the current fleet's configuration. This analysis reveals a key insight: relatively few fully committed ambulances can achieve high coverage levels for the current demand faced by Flare, highlighting the significant potential benefits of coordination.

Benchmark: Fully Committed Fleet. The most efficient theoretical configuration involves a fleet composed entirely of time-committed, location-flexible ambulances (K_5 type), over which the platform has full control. We use a simplified version of the optimization model (Section 3.2, with $U = 0$) to find the minimum number of such ambulances required to achieve different coverage levels. Figure 2 presents the results.

The results are striking: just 8 optimally located, fully committed ambulances achieve 90% expected coverage within the 15-minute target, and only 13 are needed for 95% coverage. This demonstrates that, theoretically, excellent coverage in a region like Nairobi does not necessarily require a massive fleet to meet the demand rate observed by Flare *if* that fleet is fully coordinated and committed. The marginal benefit diminishes significantly after the 13th ambulance, as covering the remaining, often more remote or sparsely populated areas (evident in Figure 1a), requires disproportionately more resources. While achieving full 100% coverage requires 37 optimally located, fully committed ambulances, the analysis underscores that high service levels are attainable with modest fleet sizes under ideal commitment conditions. We must interpret these low numbers with some caution; while the assumption of no simultaneous calls holds reasonably well for Flare's current dispersed operations (see Section 5.3 where we explore the loss from simultaneous calls for several fleet combinations), a small, highly utilized fleet of less than 10 ambulances would likely experience congestion, potentially requiring larger fleet sizes in practice than this idealized benchmark suggests.

Value of Time vs. Location Commitment. We now explore the distinct impacts of time and location commitment uncertainty in the greenfield setting. We compare fleets composed of ambulances with different commitment profiles but calibrated to have the same *expected* number of available units ($n \times q$). This experimental design allows us to isolate the impact of time commitment (higher q) versus sheer expected numbers. We analyze two spatial behaviors: location-uncommitted (*demand-searching*, $p_{jk} \propto \sum_{i:j \in J_i} d_i$) and location-flexible (optimally located via the Section 3.2 model).

Demand searching ambulances						Location-flexible ambulances				
$q = 1.0$	$q = 0.5$	$q = 0.25$	$q = 0.2$	$q = 0.1$	$n \times q$	$q = 1.0$	$q = 0.5$	$q = 0.25$	$q = 0.2$	$q = 0.1$
0.473	0.460	0.454	0.453	0.451	5	0.764	0.639	0.595	0.588	0.573
0.679	0.669	0.664	0.663	0.661	10	0.930	0.803	0.767	0.761	0.748
0.780	0.774	0.770	0.769	0.768	15	0.965	0.873	0.847	0.842	0.832

Table 1 The left side shows the coverage for a given expected number of demand searching ambulances with different availability probability and the right side for optimally located ambulances. We vary n to keep $n \times q$ fixed.

Table 1 reveals several important insights. First, comparing columns within each spatial behavior type, **increasing time commitment (higher q, lower n for fixed $n \times q$) consistently improves coverage**. For instance, with 10 expected available location-flexible ambulances ($n \times q = 10$), a fully time-committed fleet ($q = 1.0, n = 10$) achieves 93.0% coverage, significantly higher than the 74.8% achieved by a less reliable fleet ($q = 0.1, n = 100$). This highlights a non-linear benefit: even if two fleets have the same expected number of available ambulances, the one with more reliable ambulances will provide significantly better coverage. This is due to the non-linear nature of the coverage function (Eq. (1)), which depends on at least one ambulance being available and nearby. When time individual availability (q_k) is low, the negative compounding effect leads to a steep drop in system performance.

Second, comparing the demand-searching columns to the location-flexible columns demonstrates the substantial **value of spatial control (location flexibility)**. Optimally locating ambulances yields significantly higher coverage than relying on demand-searching behavior for the same number and type of ambulances, even when the demand-searching strategy is aligned with demand coverage.

Third, the results show a **strong complementarity between time commitment and location flexibility**. The coverage gain from increasing time commitment (moving left within a row) is much more pronounced for location-flexible ambulances than for demand-searching ones. Similarly, the gain from optimizing locations (moving from the left side to the right side of the table for a given q) increases as time commitment (q) increases. This suggests that the ability to strategically

place ambulances (location flexibility) is most valuable when those ambulances are reliably available (having a high time commitment). Similarly, information about availability is more valuable when it can be leveraged through location flexibility.

A critical implication arises when comparing these greenfield results to Flare's current fleet: the 13 fully committed, optimally placed ambulances achieving 95% coverage represent a potential fleet size reduction of over 95% compared to the 340 largely uncommitted ambulances Flare coordinates (detailed in Section 4). This stark contrast quantifies the immense "cost of lack of coordination" or, conversely, the potential value achievable through improved commitment and centralization in this LMIC context.

5.2. Analysis Based on Existing Fleet

Having established benchmarks, we now analyze scenarios grounded in Flare's current operational context. We aim to understand the effectiveness of interventions, specifically adding new, fully committed ambulances, given the characteristics of the existing fleet. This involves first defining baseline scenarios representing different commitment profiles that achieve similar coverage levels, and then evaluating the addition of ambulances.

Defining Comparable Fleet Scenarios. Flare's actual fleet consists of 340 largely time-uncommitted ($q \approx 0.25$) and location-committed (tied to specific bases) ambulances, achieving roughly 88% coverage according to our model (Eq. (2)) using the parameters from Section 4. To explore the impact of the existing fleet's nature, we construct five alternative hypothetical fleets that *also* achieve approximately 88% coverage but differ in their time and location commitment profiles, based on the types defined in Section 3. This experimental design allows a systematic comparison of intervention effectiveness across fleets with varying underlying commitment structures but equivalent starting performance. Table 2 summarizes these six scenarios (including the baseline representing Flare's current fleet), detailing the commitment characteristics and the resulting fleet size required to meet the $\approx 88\%$ coverage target.

Table 2 Six Fleet Scenarios Achieving Approx. 88% Coverage

Time Commitment	Location Commitment	Type	Spatial Behavior	Fleet Size (n)
Uncommitted ($q = 0.25$)	Committed (Baseline)	K_2	Fixed Bases (Flare)	340
	Uncommitted	K_4	Demand-Searching	125
	Flexible	K_6	Optimal Location	75
Committed ($q = 1.0$)	Committed	K_1	Fixed Bases (Avg.)[a]	83
	Uncommitted	K_3	Demand-Searching	30
	Flexible	K_5	Optimal Location	7

[a] Average over 100 random selections from the original 340 bases needed to reach 88% coverage.

Figure 3 illustrates the spatial distribution and resulting coverage patterns for these six baseline fleets.

The dramatic differences in fleet sizes required (Table 2) powerfully reiterate the value of commitment, particularly location flexibility combined with time commitment (7 ambulances vs. 340). Comparing the location-committed cases (Figure 3c, d) to the location-uncommitted (a, b) and location-flexible (e, f) highlights the inefficiency stemming from the suboptimal, clustered locations chosen by independent providers (visible in Figure 1b). Even simple demand-searching (a) is more efficient than the actual clustered home bases (c). This setup allows us to investigate how the *nature* of the existing fleet impacts the marginal value of adding new, optimized resources.

Impact of Fleet Expansion. We now analyze the value of adding up to five new, fully committed, location-flexible ambulances (K_5 type) to each of the six baseline fleets defined in Table 2. We use the optimization model (Section 3.2) to determine the optimal locations for these additional ambulances, maximizing the increase in coverage over the existing baseline. Table 3 shows the resulting total coverage.

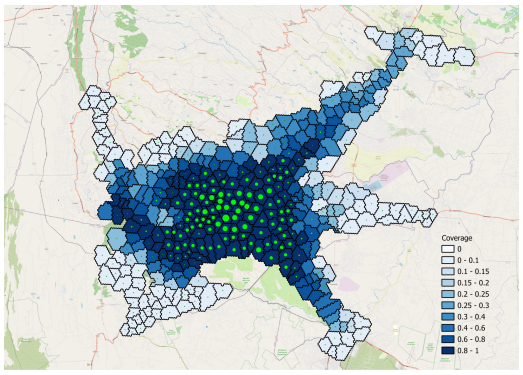
	Time-uncommitted ($q = 0.25$)			Time-committed ($q = 1.0$)		
	Location-uncommitted	Location-committed*	Location-flexible	Location-uncommitted	Location-committed	Location-flexible
Add 0	0.880	0.880	0.879	0.878	0.881	0.879
Add 1	0.892	0.906	0.887	0.890	0.924	0.900
Add 5	0.929	0.955	0.916	0.928	0.976	0.943

Table 3 Coverage after adding 1 or 5 time-committed, location-flexible ambulances to an existing fleet with different commitment levels. The response time target is 15 minutes.

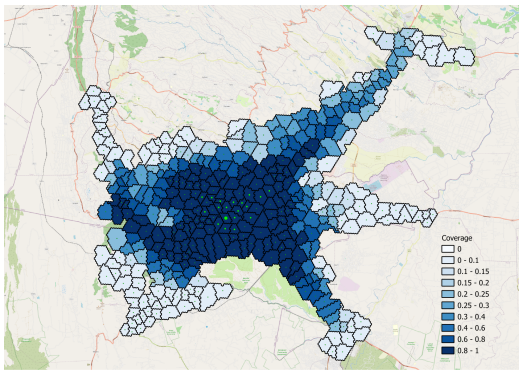
* The "Location-committed" column under "Time-uncommitted" represents the baseline scenario approximating Flare's current fleet (340 units). Other "Add 0" values correspond to the alternative scenarios from Table 2.

The results demonstrate that the **effectiveness of adding new, high-quality resources strongly depends on the commitment profile of the existing fleet**. The marginal gain from adding five committed, flexible ambulances varies significantly, from just 3.7 percentage points (0.916 - 0.879, when added to the location-flexible, time-uncommitted fleet) to 9.5 percentage points (0.976 - 0.881, when added to the location-committed, time-committed fleet).

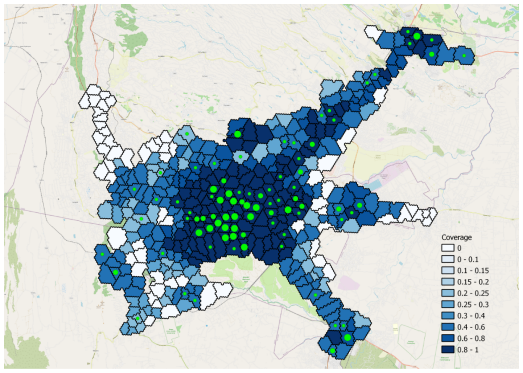
The highest gain occurs when improving upon the *location-committed* fleets. This is because these fleets, whether time-committed or not, have fixed, suboptimal base locations that leave predictable, often high-demand, geographical coverage gaps (visible in Figure 3c, d). The new flexible



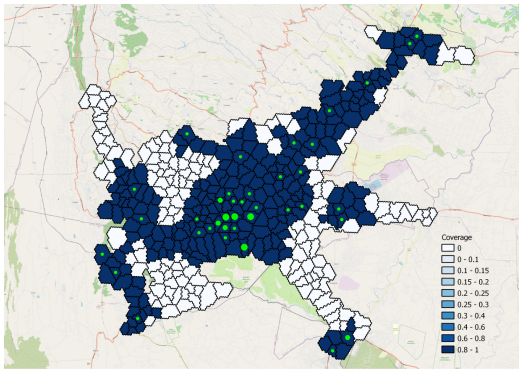
(a) Loc-uncommitted, time-uncommitted



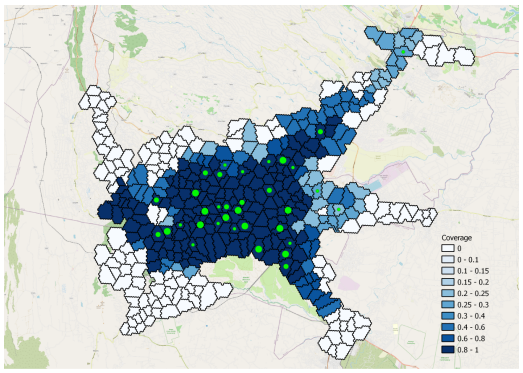
(b) Loc-uncommitted, time-committed



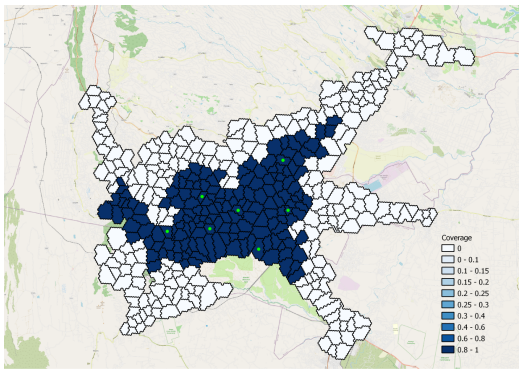
(c) Loc-committed, time-uncommitted



(d) Loc-committed, time-committed



(e) Loc-flexible, time-uncommitted



(f) Loc-flexible, time-committed

Figure 3 **Six different fleet configurations achieving approx. 88% coverage. Location commitment increases top to bottom; time commitment increases left to right. Subfigures show: (a) 125 Loc-Uncommitted, Time-Uncommitted ($q = 0.25$) ambulances (coverage 0.880); (b) 30 Loc-Uncommitted, Time-Committed ($q = 1.0$) ambulances (0.878); (c) 340 Loc-Committed, Time-Uncommitted ($q = 0.25$, Baseline) ambulances (0.880); (d) 83 Loc-Committed, Time-Committed ($q = 1.0$) ambulances (0.881); (e) 75 Loc-Flexible, Time-Uncommitted ($q = 0.25$) ambulances (0.879); (f) 7 Loc-Flexible, Time-Committed ($q = 1.0$) ambulances (0.879).**

ambulances can be precisely targeted by the optimization model (Eq. (3)) to fill these high-value gaps. Conversely, adding resources to an already optimized (location-flexible) or randomly dispersed (location-uncommitted/demand-searching) fleet yields lower marginal returns, as the remaining uncovered demand is likely in lower-demand or harder-to-reach areas. This suggests a clear strategic implication: for platforms facing fleets with suboptimal but fixed locations (similar to Flare's baseline), interventions that introduce *location flexibility*—whether through targeted incentives or dedicated platform units—likely offer the highest marginal return for improving coverage.

Interestingly, the gain is substantial even when adding to the baseline representing Flare's actual fleet (location-committed, time-uncommitted), increasing coverage from 88.0% to 92.9% with five additional optimized units. This underscores the practical implication: even modest investments in centrally coordinated, flexible resources can yield significant performance improvements in a largely uncoordinated system. This should strongly motivate platforms like Flare to explore strategies, such as targeted incentives or acquiring a small dedicated fleet, to introduce location flexibility and fill coverage gaps identified by the optimization model. Flare is indeed experimenting with such incentives based on initial findings from this work.

These numerical experiments, grounded in the Nairobi context, powerfully illustrate the substantial impact of both time and location commitment on EMS platform coverage and the potential gains from targeted interventions, such as adding location-flexible units. We observed significant inefficiencies in the current decentralized system and a strong complementarity between different forms of commitment. However, these findings are specific to the data and parameters of this case study. They raise broader questions about the fundamental trade-offs in systems with heterogeneous provider commitment. To explore these structural properties more formally and derive generalizable insights into when different commitment types might dominate in a decentralized equilibrium, we develop and analyze a stylized game-theoretic model in Section 6.

5.3. Model validation

In this section, we perform a simulation study to evaluate the performance of different fleet compositions and interventions from Section 5.2 in a more dynamic setting. This allows validating the assumption that platform-induced congestion does not play a major role. It further allows comparing these results to the expected coverage predicted by the static optimization results for key scenarios to assess the deterministic model's accuracy. Finally, the simulation allows us to assess the day-to-day variation in coverage due to random realizations of the demand (d_i) and ambulance time availability (q_k) and location (p_{jk}).

Simulation design. We perform an out-of-sample trace-driven simulation based on additional demand data gathered from Flare. We use the demand over the year 2022, as this is the first

full year after the data that was used in the evaluation and optimization model. That year, Flare handled 2866 calls in the Nairobi area, averaging 8 per day.

We explicitly simulate call arrivals based on historical data traces and track instances where the closest available ambulance is busy with a prior call. We quantify the actual coverage loss due to this effect, validating (or refining) the assumption made in Section 3.1, particularly for scenarios with smaller, potentially more utilized fleets (like the committed-fleet benchmarks).

In the simulation, we randomly sample the availability and location of ambulances from the distributions q_k and location p_{jk}. The availability of ambulances is accounted for by active sessions. Whenever an ambulance starts a session, we generate the duration of the session uniformly at random between 4 and 8 hours (mean of 6 hours). The off-session time between two sessions is also uniformly generated with a mean that ensures that the overall availability probability equals q_k: $\frac{6\times(1-q_k)}{q_k}$. At the start of each session, we use p_{jk} to randomly generate the location of this ambulance during the entire session. After serving a call, the ambulance will return to this base.

Based on the location of the incident, the closest available ambulance (on session and not busy serving other calls) is assigned to the incident. The ambulance is then busy with the call for the 3 minutes for preparation, the commute time to the incidents (which depends on the ambulance's and the incident's locations), and 75 minutes for transporting the patient to the the hospital and getting the ambulance ready for the next call.

To compute the impact of simultaneous calls, we also record the hypothetical response time of the closest ambulance that has an active session. This includes ambulances that are busy serving another platform call. Specifically, we record a call as lost due to simultaneous calls if one or more on-session ambulances have their location within a 15-minute radius, but all of them are busy with another platform call.

Impact of platform-induced congestion. In the first simulation experiment, we analyze the coverage obtained with the six fleets analyzed in Section 5.2. We simulate both the initial fleet and the resulting fleet after adding five optimally located time-committed, location-flexible ambulances. Table 4 shows the coverage for both the situation with and without simultaneous calls.

We first compare the results without simultaneous calls with the estimated coverage based on the optimization (top row of Table 3). The difference between these values comes from the difference in the demand distribution between the training (2017-2021) and the test (2022) data and the sampling of the availability and location of ambulances. This difference is typically around 2%, but slightly more for the time-committed location-flexible fleet. This result is expected since this fleet relies on a small number of precisely located ambulances, which makes it more vulnerable to the demand mismatch between training and test data.

		Time-uncommitted ($q = 0.25$)			Time-committed ($q = 1.0$)		
		Location-uncomm.	Location-committed	Location-flexible	Location-uncomm.	Location-committed	Location-flexible
Add 0	No sim. calls	0.869	0.855	0.861	0.866	0.856	0.843
	With sim. calls	0.864	0.851	0.849	0.861	0.851	0.768
	Lost coverage	0.005	0.004	0.011	0.005	0.006	0.074
Add 5	No sim. calls	0.918	0.948	0.895	0.914	0.973	0.930
	With sim. calls	0.913	0.943	0.888	0.909	0.963	0.877
	Lost coverage	0.005	0.005	0.008	0.004	0.009	0.052

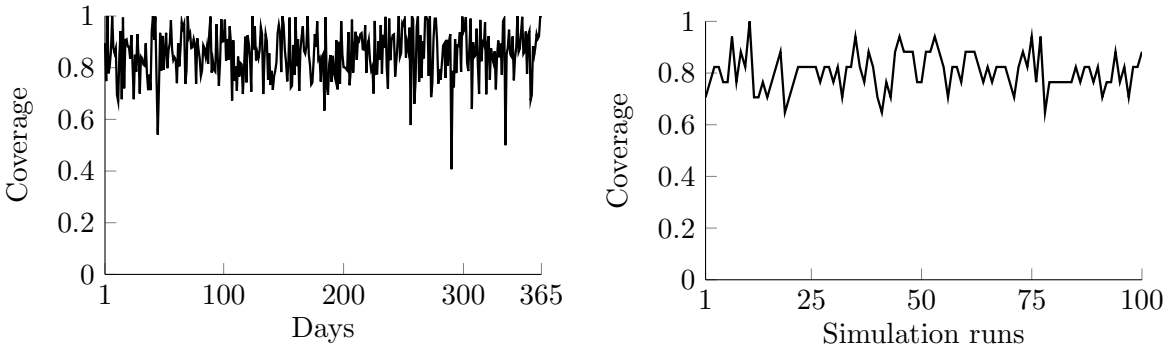
Table 4 Average coverage based on 100 simulation runs for fleets with different commitment levels with and without simultaneous calls. Lost coverage shows reduction in coverage as a results of simultaneous calls. The response time target is 15 minutes.

Second, we compare the resulting coverage after adding 5 time-committed location-flexible ambulances with the estimated coverage from the optimization (last row of Table 3). This shows that a slightly lower coverage is obtained, but that the different demand distribution and sampling of locations and availability of ambulances does not have a significant impact on the effectiveness of the intervention. Most of the estimated gains are indeed realized.

Finally, we evaluate the impact of simultaneous calls and observe that this impact is typically very small, but it does depend on the scenario. Scenarios with a high level of commitment have fewer ambulances, making them more vulnerable for simultaneous calls. Specifically, a fleet of 7 time-committed location-flexible ambulances experiences a 7.4% reduction in coverage because of simultaneous calls. This is in line with the literature as this fleet resembles a traditional EMS system with fully controllable ambulances. For these systems, it is well-established that simultaneous calls should be accounted for. However, in the setting that we analyze in Section 4, that is, in LMICs EMS systems' commitment levels are typically low, and losses from simultaneous calls are less than 1% (see the third and the last row of Table 3).

Impact of Stochasticity. Next, we analyze the impact of stochasticity on both the demand and capacity side. For this analysis, we provide an in-depth examination of the simulation results for the baseline scenario, which is the one most closely aligned with Flare's actual fleet. Across different days, the main driver of variation in coverage is the number of incidents and their location. When fixing a day, and thereby the occurrence of incidents, we can evaluate the variation in coverage due to random ambulance availability (q_k) and location (p_{jk}) realizations.

Figure 4a shows the average coverage over 100 simulation runs for each day in the test data. By averaging over 100 simulation runs, the capacity-side uncertainty is negligible, and the fluctuation (5%: 0.70, 95%: 0.997) shown is caused by stochasticity on the demand side. As simultaneous calls are rare, the variation is mostly due to the location of incidents, not to the daily call volume.



(a) Coverage variation across days (b) Coverage variation across runs

Figure 4 **Impact of stochasticity on coverage based on simulation for baseline fleet. Subfigures show: (a) Variation in coverage across different days, based on average coverage over 100 simulation runs.(b) Variation in coverage over 100 simulation runs for busiest day (March 4, 2022) with 17 incidents.**

Figure 4b shows the variation in coverage over 100 simulation runs for the busiest day in our test data. During this day, Flare experienced 17 incidents in the Nairobi area. The figure shows that even when we completely exclude demand-side stochasticity, we see large fluctuations (5%: 0.71, 95%: 0.94) in coverage induced by the uncertainty in availability and location of ambulances.

When we perform the same analysis on the baseline fleet after adding 5 optimally located, time-committed, location-flexible ambulances, we see that most of the variation vanishes. The coverage variation across days (5%: 0.81, 95%: 1.00) decreases significantly, and the coverage variation across runs for the busiest day vanishes entirely, with all 100 simulations runs giving full coverage. In other words, besides increasing the overall coverage, our results illustrate that adding committed ambulances to the fleet also reduces the uncertainty in the coverage.

6. Analytical Model

The numerical analysis in Section 5 provided detailed insights into the value of commitment within Flare's specific operational context. This section complements these numerical findings and derives more general structural insights through a stylized game-theoretic model. Our primary goal in this section is to understand under what conditions (regarding costs, coverage range, and time availability) does time commitment lead to better ambulance coverage than location commitment in a decentralized equilibrium?

Model Setup. We consider a stylized model of a circular city with perimeter normalized to 1. Demand is uniformly distributed around the circle. Ambulances can serve calls within their spatial reach, an interval of length $d \in (0, 1)$, centered at their location. If multiple ambulances cover a call, we assume the responding ambulance is chosen uniformly at random; this simplification ensures

analytical tractability compared to assuming the closest ambulance responds. We normalize the average call revenue to 1. The operational cost per unit of time is $a > 0$. Thus, the ambulance cost structure is aq, where $q \in (0, 1)$ is the fraction of the time the ambulance commits to the platform.

To isolate the fundamental trade-offs between temporal and spatial predictability, we focus on two ambulance archetypes representing extremes of commitment:

- **Roaming (Time-Committed, Location-Uncommitted):** These ambulances are fully temporally committed ($q = 1$). We model their location-uncommitted behavior by assuming they roam around the city searching for demand, effectively covering any point on the circle with probability d, independent of other ambulances. Their operational cost is a.
- **Stationed (Time-Uncommitted, Location-Committed):** These ambulances commit to specific locations but are only available to the platform a fraction q of the time. Their temporal availability is independent across ambulances. As is standard in the literature, we focus on a symmetric equilibrium and assume they position themselves equidistantly around the circle. Their operational cost, prorated by their time commitment, is aq.

Equilibrium Concept and Assumptions. We analyze a free-entry equilibrium where ambulances join the market (as either roaming or stationed) as long as their expected profit is non-negative. For analytical tractability, we make two standard simplifying assumptions: (i) *No congestion:* We assume non-overlapping calls, meaning an ambulance's availability is determined only by its exogenous time commitment (q) and not by being busy with prior platform calls. As discussed in Section 4 and validated in Section 5.3, this assumption aligns with the low platform call volume observed for Flare, allowing us to focus purely on commitment effects. (ii) *Continuous relaxation:* We relax the integrality constraint on the number of ambulances, allowing them to be non-negative real numbers. This fluid approximation is common for simplifying equilibrium analysis in location and market entry models (Gernert et al. 2024).

We characterize the pure equilibria induced by each ambulance type in Sections 6.1 and 6.2, respectively, compare their induced total coverage in Section 6.3, and numerically explore a mixed equilibrium where both ambulance types co-exist in Section 6.4. This analysis provides a theoretical foundation for the phenomena observed in Section 5. All the proofs are provided in Appendix 8.

6.1. Equilibrium with Roaming Ambulances Only

REMARK 1. *If k_r roaming ambulances join the market:*

(i) *The induced total coverage is $TC_r(k_r) = 1 - (1-d)^{k_r}$.*

(ii) *The induced profit for each ambulance is $\Pi_r(k_r) = \frac{TC_r(k_r)}{k_r} - a = \frac{1-(1-d)^{k_r}}{k_r} - a$.*

The total coverage $TC_r(k_r)$ follows directly from the independence assumption on the ambulance roaming behavior: the probability that no ambulance covers a given point is $(1-d)^{k_r}$. The profit $\Pi_r(k_r)$ arises because the total expected revenue generated by the system is $TC_r(k_r)$ (since revenue per call is normalized to 1), and under the uniform random dispatch assumption, this revenue is shared equally among the k_r identical ambulances present, less their operational cost a.

In Appendix 8.1, we formally establish the existence and uniqueness of the symmetric free-entry equilibrium number of roaming ambulances $\hat{k}_r \geq 0$ such that $\Pi_r(\hat{k}_r) = 0$.

6.2. Equilibrium with Stationed Ambulances Only

Due to their equidistant locations, the coverage calculation is significantly more complex for stationed ambulances than for roaming ambulances, as we formalize next.

THEOREM 1 **(Stationed Ambulances' Coverage Formula)**. *For any d, $q \in (0,1)$, if $k_s \in \mathbb{N}$ stationed ambulances are deployed equidistantly, then the total coverage attained is*

$$\begin{aligned} TC_s(k_s) &= (dk_s - \lfloor dk_s \rfloor)(1-(1-q)^{\lfloor dk_s \rfloor + 1}) + (1-(dk_s - \lfloor dk_s \rfloor))(1-(1-q)^{\lfloor dk_s \rfloor}) \\ &= 1-(1-q)^{\lfloor dk_s \rfloor}\,(1-(dk_s - \lfloor dk_s \rfloor)q)\,. \end{aligned} \tag{4}$$

The intuition behind the first equality in Eq. (4) arises from the symmetric ambulance locations. Any point on the circle is covered by either $N = \lfloor dk_s \rfloor$ or $N+1$ ambulances, depending on its position relative to the equidistant ambulance locations. A fraction $(dk_s - \lfloor dk_s \rfloor)$ of the circle is covered by $N+1$ ambulances, and the probability that at least one of these is available is $(1-(1-q)^{N+1})$. The remaining fraction $(1-(dk_s - \lfloor dk_s \rfloor))$ is covered by N ambulances, with coverage probability $(1-(1-q)^N)$. The total expected coverage is the weighted average of these probabilities. The second equality in Eq. (4) follows by rearranging terms.

Theorem 1 shows that the induced coverage by stationed ambulances is uneven unless $dk_s \in \mathbb{N}$, in which case the ambulance coverage matches the uniform demand distribution.

REMARK 2. *If k_s stationed ambulances join the market, the induced profit for each is*

$$\Pi_s(k_s) = \frac{TC_s(k_s)}{k_s} - aq = \frac{1-(1-q)^{\lfloor dk_s \rfloor}\,(1-(dk_s - \lfloor dk_s \rfloor)q)}{k_s} - aq. \tag{5}$$

Analogous to our discussion about roaming ambulances, the stationed ambulances' profit $\Pi_s(k_s)$ is the equally shared total revenue $TC_s(k_s)$ (due to symmetry) minus the individual operational cost aq, which accounts for their reduced time commitment.

In Appendix 8.3, we establish the existence and uniqueness of the symmetric free-entry equilibrium number of stationed ambulances $\hat{k}_s \geq 0$ such that $\Pi_s(\hat{k}_s) = 0$.

6.3. Total Coverage Comparison: Pure Roaming vs Pure Stationed Fleet

In this section, we investigate under which conditions the equilibrium total coverage provided by a roaming fleet exceeds that of a stationed fleet. This analysis provides insight into when ambulance time or location commitment yields higher coverage under different operational parameter regimes.

THEOREM 2. *For any d, $q \in (0,1)$, and $a > 0$, let $\hat{k}_r(d,a)$ and $\hat{k}_s(d,q,a)$ denote the free-entry equilibrium fleet sizes of roaming and stationed ambulances, respectively, and let $TC_r^*(d,a) := TC_r(\hat{k}_r(d,a))$ and $TC_s^*(d,q,a) := TC_s(\hat{k}_s(d,q,a))$ denote their induced total coverage. Then,*

(i) ***Entry condition.*** *$\hat{k}_r(d,a) > 0$ and $\hat{k}_s(d,q,a) > 0$ if and only if $a < d$.*

(ii) ***Unique time commitment threshold.*** *If $a < d$, then there exists a unique $\bar{q}(d,a) \in (0,1)$ such that $TC_r^*(d,a) = TC_s^*(d,\bar{q}(d,a),a)$ and $TC_r^*(d,a) < TC_s^*(d,q,a) \iff q > \bar{q}(d,a)$.*

Theorem 2(i) shows that the market can support either ambulance type if and only if their operating costs per unit of time, a, fall beneath their spatial-reach parameter, d. Within this parameter regime, Theorem 2(ii) characterizes a single stationed ambulances' time-commitment threshold $\bar{q}(d,a)$. A time commitment above $\bar{q}(d,a)$ makes the total coverage of stationed ambulances superior because their spatial predictability outweighs the unreliability of their time availability. A time commitment below $\bar{q}(d,a)$ restores the advantage of perfectly time-reliable roaming ambulances.

The following corollary enables us to shed further light on the insights from Theorem 2.

COROLLARY 1 **(Simplified Coverage Comparison)**. *The equilibrium total coverage induced by stationed ambulances matches the uniform demand distribution over the circle —formally, $d\hat{k}_s(d,\bar{q}(d,a),a) \in \mathbb{N}$— if and only if the time commitment threshold collapses to $\bar{q}(d,a) = d$.*

Corollary 1 shows that if the stationed ambulances' total coverage matches the uniform demand distribution, then total coverage dominance reduces to a simple commitment comparison: the roaming archetype delivers higher total coverage if and only if the ambulances' spatial reach d, i.e., their probability of matching demand due to a lack of location commitment, exceeds the stationed fleet's time commitment q.

Corollary 1, together with $d \leq \bar{q}(d,a)$ from Theorem 2(ii), also underscores why decentralized fleets rarely achieve first-best performance. Only in rare cases, when ambulance locations chosen to maximize individual profits collectively match the distribution of demand, does coverage performance boil down to a simple commitment comparison. In almost all instances, spatial misalignment with demand penalizes stationed fleets, and roaming ambulances can attain a higher total coverage even if their probability of matching demand due to a lack of location commitment, d, is less than the time commitment of stationed ambulances, q.

6.4. Numerical Results of Mixed Equilibrium with Both Ambulance Types

While comparing pure fleets is insightful, real-world systems likely involve a mix of ambulance types. Therefore, we analyze the free-entry equilibrium when both roaming and stationed ambulances can simultaneously enter the market, aiming to understand the marginal value of each ambulance type.

REMARK 3. If k_r roaming and k_s stationed ambulances join the market, let $P_r(k_r, k_s)$ and $P_s(k_r, k_s)$ denote the probability that a roaming and a stationed ambulance, respectively, gets assigned a call. Then,

(i) The induced Total Coverage is given by

$$TC_{rs}(k_r, k_s) = 1 - (1-d)^{k_r}(1-q)^{\lfloor dk_s \rfloor}(1 - (dk_s - \lfloor dk_s \rfloor)q). \tag{6}$$

(ii) The induced profit for each roaming and stationed ambulance is given by

$$\Pi_r(k_r, k_s) = \frac{P_r(k_r, k_s)}{k_r} - a, \quad \Pi_s(k_r, k_s) = \frac{P_s(k_r, k_s)}{k_s} - aq. \tag{7}$$

The total coverage formula (6) naturally extends the pure stationed case (4) by incorporating the impact of the k_r roaming ambulances. However, the profit functions (7) become analytically intractable. Specifically, calculating $P_r(k_r, k_s)$ and $P_s(k_r, k_s)$ requires complex combinatorial terms (detailed in Appendix 8.5). As a result, characterizing the mixed equilibrium (k_r^*, k_s^*) that simultaneously satisfies $\Pi_r(k_s^*, k_r^*) = 0$ and $\Pi_s(k_s^*, k_r^*) = 0$ is analytically intractable. Therefore, we turn to numerical analysis to understand the marginal value of each ambulance type in equilibrium.

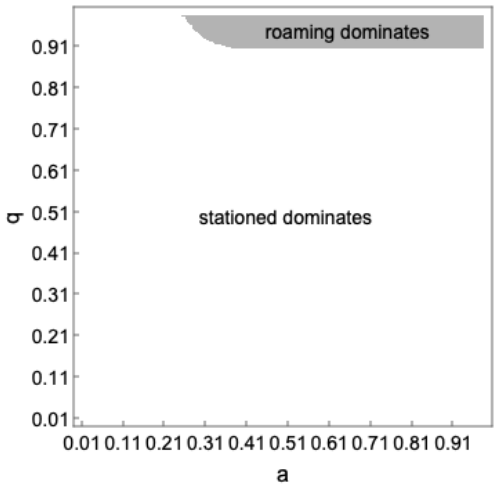


Figure 5 **In the grey parameter area, an additional roaming ambulance has a higher additional-coverage-to-cost ratio than an additional stationed ambulance. The opposite is true in the white area. Plot for $d = 0.5$.**

Figure 5 illustrates the results when the ambulances spatial reach $d = 0.5$. It is representative of the numerical results we find for different values of d. Figure 5 reveals a key insight. It illustrates when the return on investment (ROI), measured as the ratio of the increase in total coverage due to an additional ambulance to its cost per unit of time, is larger for each ambulance type. It suggests that to improve the equilibrium total coverage, investing in an additional stationed ambulance typically provides a higher ROI than investing in an additional roaming ambulance. These results align with the findings in Section 5.2, where the marginal value of adding time-uncommitted location-committed ambulances was consistently larger than the marginal value of adding time-committed location-uncommitted ambulances. In both cases, the generally superior performance of adding spatially committed ambulances compared to their temporally committed counterparts is mostly driven by the former's ability to fill coverage gaps created by decentralized location decisions more effectively.

7. Conclusion

This paper tackles the operational challenges of decentralized emergency medical services in low- and middle-income countries (LMICs), where supply uncertainty from heterogeneous provider commitment significantly impacts service quality. Through comprehensive analysis combining optimization, simulation, and game theory, we address a critical question: *What is the relative effectiveness of improving temporal versus spatial commitment in enhancing coverage for decentralized emergency response platforms?*

Our primary finding is that interventions that enhance spatial commitment through effective strategic positioning generally yield higher marginal returns for improving coverage compared to those that focus solely on increasing temporal commitment. This insight is particularly valuable in resource-constrained environments where targeted interventions must maximize impact. Our analysis quantifies the stark "cost of decentralization" - the coverage currently provided by approximately 340 largely uncommitted ambulances in Nairobi could potentially be matched by fewer than 15 optimally deployed, fully committed ambulances. We also find important complementarity effects: the value of time commitment is significantly amplified when coupled with location flexibility.

These findings offer important managerial implications for emergency platform operators. Decision makers should prioritize strategies enhancing spatial control and positioning, such as targeted relocation incentives (which our partner Flare has begun implementing based on our research), deploying platform-managed flexible units, or fostering partnerships that encourage coverage-oriented basing. Our modeling framework provides practical tools for identifying coverage gaps

and assessing potential benefits of interventions targeting either commitment dimension, even in settings with limited data availability.

From a theoretical perspective, this research contributes by analyzing the interplay of spatial and temporal commitment heterogeneity in decentralized service platforms, moving beyond common homogeneity assumptions in ride-hailing studies (Bimpikis et al. 2019, Chen et al. 2024). It informs the operational flexibility literature (Enayati et al. 2018, Shu et al. 2013) by demonstrating the differential value of location versus time flexibility in decentralized networks and extends emergency medical services literature (Boutilier and Chan 2020) by explicitly modeling the dual supply-side uncertainties prevalent in LMICs. Notably, the structural insights derived from our stylized game-theoretic model align with our case-based optimization and simulation analyses, reinforcing the robustness of our findings.

Our study has limitations, including the primarily tactical nature of the optimization models and the Nairobi-specific empirical context. These limitations suggest promising directions for future research: developing dynamic real-time relocation models incorporating commitment uncertainty, creating richer behavioral models of provider location choice, designing specific incentive contracts, incorporating equity considerations, and extending the analysis to other contexts or platform types.

In conclusion, effectively managing decentralized emergency systems requires understanding distinct operational leverage points. This paper demonstrates that ensuring resources are strategically positioned often yields greater benefits than solely increasing their time availability. By bridging operations management theory with practical implementation, our work provides guidance for enhancing critical emergency service platforms in resource-constrained environments, aiming to improve health outcomes for millions of people who currently lack reliable access to emergency medical services.

Acknowledgments

This research was financed in part by the Netherlands Organization for Scientific Research (NWO) in the form of a Veni grant [project number VI.Veni.191E.005] for P.L. van den Berg, and by the Natural Sciences and Engineering Research Council of Canada [Grant RGPIN-2025-05848] for G. Romero.

References

Afèche P, Liu Z, Maglaras C (2023) Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. *Manufacturing & Service Operations Management* 25(5):1890–1908.

Alanis R, Ingolfsson A, Kolfal B (2013) A markov chain model for an ems system with repositioning. *Production and Operations Management* 22(1):216–231, URL `http://dx.doi.org/https://doi.org/10.1111/j.1937-5956.2012.01362.x`.

Bai J, So KC, Tang CS, Chen X, Wang H (2019) Coordinating supply and demand on an on-demand service platform with impatient customers. *Manufacturing & Service Operations Management* 21(3):556–570.

Balseiro SR, Brown DB, Chen C (2021) Dynamic pricing of relocating resources in large networks. *Management Science* 67(7):4075–4094.

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

Benjaafar S, Wu S, Liu H, Gunnarsson EB (2021) Dimensioning on-demand vehicle sharing systems. *Management Science* 68(2):1218–1232.

Bimpikis K, Candogan O, Saban D (2019) Spatial pricing in ride-sharing networks. *Operations Research* 67(3):744–769.

Boutilier JJ, Chan TCY (2020) Ambulance emergency response optimization in developing countries. *Operations Research* 68(5):1315–1334, URL `http://dx.doi.org/10.1287/opre.2019.1969`.

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. *European Journal of Operational Research* 147(3):451–463.

Cachon GP, Daniels KM, Lobel R (2017) The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management* 19(3):368–384.

Chen Q, Lei Y, Jasin S (2024) Real-time spatial–intertemporal pricing and relocation in a ride-hailing network: Near-optimal policies and the value of dynamic pricing. *Operations Research* 72(5):2097–2118.

Das S, Desai R (2017) Emergence of EMS in India: India's budding EMS infrastructure is a model for the region despite challenges moving forward. *JEMS: A Journal of Emergency Medical Services* 42(4):55.

Daskin MS (1983) A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation Science* 17(1):48–70.

El Itani B, Abdelaziz FB, Masri H (2019) A bi-objective covering location problem: Case of ambulance location in the beirut area, lebanon. *Management Decision* 57:432–444.

Enayati S, Mayorga ME, Rajagopalan HK, Saydam C (2018) Real-time ambulance redeployment approach to improve service coverage with fair and restricted workload for ems providers. *Omega* 79:67–80, ISSN 0305-0483, URL `http://dx.doi.org/https://doi.org/10.1016/j.omega.2017.08.001`.

Gernert AK, Calmon AP, Romero G, Van Wassenhove LN (2024) Business model innovation for ambulance systems in low-and middle-income countries: "coordination and competition". *Production and Operations Management* URL `http://dx.doi.org/10.1177/10591478231224973`.

He L, Hu Z, Zhang M (2020) Robust repositioning for vehicle sharing. *Manufacturing & Service Operations Management* 22(2):241–256.

He L, Mak HY, Rong Y, Shen ZJM (2017) Service region design for urban electric vehicle sharing systems. *Manufacturing & Service Operations Management* 19(2):309–327.

Hosseini M, Milner J, Romero G (2024) Dynamic relocations in car-sharing networks. *Operations Research* .

Jin Z, Wang Y, Lim YF, Pan K, Shen ZJM (2023) Vehicle rebalancing in a shared micromobility system with rider crowdsourcing. *Manufacturing & Service Operations Management* 25(4):1394–1415.

Kabra A, Belavina E, Girotra K (2020) Bike-share systems: Accessibility and availability. *Management Science* 66(9):3803–3824, URL `http://dx.doi.org/10.1287/mnsc.2019.3407`.

Keskinocak P, Savva N (2020) A review of the healthcare-management (modeling) literature published in manufacturing & service operations management. *Manufacturing & Service Operations Management* 22(1):59–72.

Khanna V (2017) Now, Dial 4242 to book an ambulance. *The Hindu* (May 19).

Kuo L (2016) A startup in Kenya is launching "Uber for ambulances". *Quartz Africa* (December 16).

Macintyre K, Hotchkiss DR (1999) Referral revisited: community financing schemes and emergency transport in rural africa. *Social Science & Medicine* 49(11):1473–1487.

Marla L, Krishnan K, Deo S (2021a) Managing ems systems with user abandonment in emerging economies. *IISE Transactions* 53(4):389–406, URL `http://dx.doi.org/10.1080/24725854.2020.1802086`.

Marla L, Shin J, Deshpande S (2021b) Strategic service logistics games with customer-induced competition. *Available at SSRN: https://ssrn.com/abstract=3205999* URL `http://dx.doi.org/10.2139/ssrn.3205999`.

Marla L, Yue Y, Krishnan K (2017) Data-driven omniscient bounds and greedy policies for ambulance allocation and dynamic redeployment. *Available at SSRN: https://ssrn.com/abstract=3009043* URL `http://dx.doi.org/10.2139/ssrn.3009043`.

McLay LA, Mayorga ME (2013) A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing & Service Operations Management* 15(2):205–220.

Nasrollahzadeh AA, Khademi A, Mayorga ME (2018) Real-time ambulance dispatching and relocation. *Manufacturing & Service Operations Management* 20(3):467–480.

Shu J, Chou MC, Liu Q, Teo CP, Wang IL (2013) Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems. *Operations Research* 61(6):1346–1359.

Sudtachat K, Mayorga ME, Mclay LA (2016) A nested-compliance table policy for emergency medical service systems under relocation. *Omega* 58:154–168, ISSN 0305-0483, URL `http://dx.doi.org/https://doi.org/10.1016/j.omega.2015.06.001`.

Taylor TA (2018) On-demand service platforms. *Manufacturing & Service Operations Management* 20(4):704–720.

Thomson N (2005) Emergency medical services in Zimbabwe. *Resuscitation* 65(1):15–19.

van Barneveld T (2016) The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. *INFORMS Journal on Computing* 28(2):370–384, URL `http://dx.doi.org/10.1287/ijoc.2015.0687`.

van Barneveld T, van der Mei R, Bhulai S (2017) Compliance tables for an ems system with two types of medical response units. *Computers & Operations Research* 80:68–81, ISSN 0305-0548, URL `http://dx.doi.org/https://doi.org/10.1016/j.cor.2016.11.013`.

van den Berg PL, Aardal K (2015) Time-dependent mexclp with start-up and relocation cost. *European Journal of Operational Research* 242(2):383–389.

8. Analytical Results and Proofs

8.1. Pure Roaming Ambulances Equilibrium

PROPOSITION 1. *For any $d \in (0,1)$ and $a > 0$, there exists a unique symmetric free-entry equilibrium number of roaming ambulances $\hat{k}_r \geq 0$ such that $\Pi_r(\hat{k}_r) = 0$. Further,*

(i) *The equilibrium is positive, i.e., $\hat{k}_r > 0$, if and only if $a < -\log(1-d)$. If $a \geq -\log(1-d)$, then $\hat{k}_r = 0$.*

(ii) *If $a < -\log(1-d)$, then $\hat{k}_r$ is strictly decreasing in a and strictly increasing in d.*

Proof of Proposition 1. We first show part (i) of the proposition.

Note that $\Pi_r(k_r)$ is strictly decreasing on $k_r \in (0,\infty)$. Indeed,

$$\Pi_r'(k_r) = \frac{-(1-(1-d)^{k_r}) - k_r(1-d)^{k_r}\log(1-d)}{k_r^2} < 0 \text{ for all } k_r > 0. \tag{8}$$

To see the inequality in (8), let $h(k_r) = -(1-(1-d)^{k_r}) - k_r(1-d)^{k_r}\log(1-d)$. Note that $h(0) = 0$ and $h'(k_r) = -k_r(1-d)^{k_r}\log^2(1-d) < 0$ for all $d \in (0,1)$ and $k_r > 0$. Therefore, $h(k_r) < 0$, and thus $\Pi_r'(k_r) < 0$, for all $k_r \in (0,\infty)$.

We now characterize the values attained by $\Pi_r(k_r)$ at the extremes of the interval $(0,\infty)$. Indeed, using L'Hôpital's rule,

$$\lim_{k_r \downarrow 0} \Pi_r(k_r) = \lim_{k_r \downarrow 0} \frac{-(1-d)^{k_r}\log(1-d)}{1} - a = -\log(1-d) - a, \qquad \lim_{k_r \to \infty} \Pi_r(k_r) = -a < 0.$$

Therefore, we conclude that if $a \geq -\log(1-d)$, then $\Pi_r(k_r) \leq 0$ for all $k_r > 0$, so the unique free–entry equilibrium is that no ambulances enter the market, i.e., $k_r^* = 0$.

Further, if $a < -\log(1-d)$ then $\Pi_r(k_r)$ is positive for k_r close to 0 and negative for large k_r. Continuity and strict monotonicity of $\Pi_r(k_r)$ imply exactly one root $\hat{k}_r > 0$ satisfying $\Pi_r(\hat{k}_r) = 0$, completing the proof of part (i) of the proposition.

We now prove part (ii) of the proposition.

Assume $a < -\log(1-d)$ and define $F(k_r, a, d) := \frac{1-(1-d)^{k_r}}{k_r} - a$. Then $F(\hat{k}_r, a, d) = 0$ and,

$$\frac{\partial \hat{k}_r}{\partial a} = -\frac{\partial F/\partial a}{\partial F/\partial k_r} = -\frac{-1}{\Pi_r'(\hat{k}_r)} < 0,$$

where the first equality follows from the Implicit Function Theorem, and the inequality follows from (8). Similarly,

$$\frac{\partial \hat{k}_r}{\partial d} = -\frac{\partial F/\partial d}{\partial F/\partial k_r} = -\frac{(1-d)^{\hat{k}_r}}{\Pi_r'(\hat{k}_r)} > 0,$$

where, again, the first equality follows from the Implicit Function Theorem, and the inequality follows from (8), completing the proof. □

8.2. Proof of Theorem 1

Recall that, by symmetry, the k_s ambulances are stationed equidistantly in the unit circle. Therefore, to study the coverage in the entire circle, it is sufficient to examine the coverage in the interval $[0, 1/k_s)$, starting from an ambulance location. Let $C_s(x)$ denote the coverage at location $x \in [0, 1/k_s)$, starting from an ambulance location, when k_s ambulances are stationed equidistantly in the unit circle.

We start by stating and proving two technical results that will allow us to prove Theorem 1.Specifically, they will enable us to fully characterize $C_s(x)$ for $x \in [0, 1/k_s)$.

LEMMA 1. *For any $d,\ q \in (0,1)$, if $k_s \in \mathbb{N}$ stationed ambulances are deployed equidistantly, such that $dk_s \geq 1$, then the coverage attained at any ambulance location is*

$$C_s(0) = \begin{cases} 1-(1-q)^{\lfloor dk_s \rfloor + 1}, & \textit{if } \lfloor dk_s \rfloor \textit{ is even,} \\ 1-(1-q)^{\lfloor dk_s \rfloor}, & \textit{if } \lfloor dk_s \rfloor \textit{ is odd.} \end{cases}$$

Proof of Lemma 1. Every ambulance location b is covered by its own stationed ambulance plus the number of other stationed ambulances whose location is within a distance $d/2$ of b. Let $l \in \mathbb{N}$ denote the number of other stationed ambulances whose location is within a distance $d/2$ on one side of b, i.e., l is the unique integer such that $l/k_s \leq d/2 < (l+1)/k_s$, or equivalently $l = \left\lfloor \frac{dk_s}{2} \right\rfloor$. Hence, we conclude that the coverage at each ambulance location b is $C_s(0) = 1-(1-q)^{2l+1} = 1-(1-q)^{2\left\lfloor \frac{dk_s}{2} \right\rfloor + 1}$.

To conclude, we show that

$$C_s(0) = 1-(1-q)^{2\left\lfloor \frac{dk_s}{2} \right\rfloor + 1} = \begin{cases} 1-(1-q)^{\lfloor dk_s \rfloor + 1}, & \text{if } \lfloor dk_s \rfloor \text{ is even,} \\ 1-(1-q)^{\lfloor dk_s \rfloor}, & \text{if } \lfloor dk_s \rfloor \text{ is odd.} \end{cases}$$

Indeed, if $\lfloor dk_s \rfloor$ is even, then there exists $m \in \mathbb{N}$ such that $\lfloor dk_s \rfloor = 2m$. Hence, $2\left\lfloor \frac{dk_s}{2} \right\rfloor + 1 = 2\left\lfloor \frac{2m}{2} \right\rfloor + 1 = 2m+1 = \lfloor dk_s \rfloor + 1$, concluding the proof when $\lfloor dk_s \rfloor$ is even.

Finally, if $\lfloor dk_s \rfloor$ is odd, then there exists $m \in \mathbb{N}$ such that $\lfloor dk_s \rfloor = 2m+1$. Hence, $2\left\lfloor \frac{dk_s}{2} \right\rfloor + 1 = 2\left\lfloor \frac{2m+1}{2} \right\rfloor + 1 = 2m+1 = \lfloor dk_s \rfloor$, concluding the proof when $\lfloor dk_s \rfloor$ is odd, and thus the lemma. □

PROPOSITION 2. *For any d, $q \in (0,1)$, if $k_s \in \mathbb{N}$ stationed ambulances are deployed equidistantly, such that $dk_s \geq 1$, then the coverage attained at any point in the circle at a distance $x \in [0, 1/k_s)$ from an ambulance location is*

$$C_s(x) = \begin{cases} 1-(1-q)^{\lfloor dk_s \rfloor + 1}, & \text{if } x \in \left[0, \frac{dk_s - \lfloor dk_s \rfloor}{2k_s}\right), \\ 1-(1-q)^{\lfloor dk_s \rfloor}, & \text{if } x \in \left[\frac{dk_s - \lfloor dk_s \rfloor}{2k_s}, \frac{\lfloor dk_s \rfloor + 2 - dk_s}{2k_s}\right), \\ 1-(1-q)^{\lfloor dk_s \rfloor + 1}, & \text{if } x \in \left[\frac{2-(dk_s - \lfloor dk_s \rfloor)}{2k_s}, \frac{1}{k_s}\right). \end{cases}$$

if $\lfloor dk_s \rfloor$ is even, and it is

$$C_s(x) = \begin{cases} 1-(1-q)^{\lfloor dk_s \rfloor}, & \text{if } x \in \left[0, \frac{\lfloor dk_s \rfloor + 1 - dk_s}{2k_s}\right), \\ 1-(1-q)^{\lfloor dk_s \rfloor + 1}, & \text{if } x \in \left[\frac{\lfloor dk_s \rfloor + 1 - dk_s}{2k_s}, \frac{1 + dk_s - \lfloor dk_s \rfloor}{2k_s}\right), \\ 1-(1-q)^{\lfloor dk_s \rfloor}, & \text{if } x \in \left[\frac{1 + dk_s - \lfloor dk_s \rfloor}{2k_s}, \frac{1}{k_s}\right), \end{cases}$$

if $\lfloor dk_s \rfloor$ is odd.

Proof of Proposition 2. The case for $x = 0$ follows from Lemma 1.

By continuity, exactly the same proof as in Lemma 1 works for all $x \in [0, \min(x_1, x_2))$, where $x_1 = \left(\frac{dk_s}{2} - \left\lfloor \frac{dk_s}{2} \right\rfloor\right) \Big/ k_s$ and $x_2 = \left(\left\lfloor \frac{dk_s}{2} \right\rfloor + 1 - \frac{dk_s}{2}\right) \Big/ k_s$. Indeed, as x moves over $[0, 1/k_s)$, the only two changes in coverage that occur are that location x stops being covered by the ambulance further away from it when $x = x_1$, and it starts being covered by a new ambulance when $x = x_2$.

We now prove the case when $\lfloor dk_s \rfloor$ is even. Then, we have $x_1 = \frac{dk_s - \lfloor dk_s \rfloor}{2k_s} \leq \frac{\lfloor dk_s \rfloor + 2 - dk_s}{2k_s} = x_2$, where both equalities follow since $\lfloor dk_s \rfloor$ being even implies $2\left\lfloor \frac{dk_s}{2} \right\rfloor = \lfloor dk_s \rfloor$ (cf. proof of Lemma 1).

Hence, in this case we conclude that all locations $x \in [x_1, x_2)$ are covered by one less ambulance than the locations in the interval $[0, x_1)$, leading to a coverage $1-(1-q)^{\lfloor dk_s \rfloor}$ if $x \in [x_1, x_2)$.

Similarly, in this case all locations $x \in [x_2, 1/k_s)$ are covered by one more ambulance than the locations in the interval $[x_1, x_2)$, leading to a coverage $1-(1-q)^{\lfloor dk_s \rfloor + 1}$ if $x \in [x_2, 1/k_s)$, concluding the proof when $\lfloor dk_s \rfloor$ is even.

To conclude, we now prove the case when $\lfloor dk_s \rfloor$ is odd. Then, we have $x_2 = \frac{\lfloor dk_s \rfloor + 1 - dk_s}{2k_s} \leq \frac{1 + dk_s - \lfloor dk_s \rfloor}{2k_s} = x_1$, where both equalities follow since $\lfloor dk_s \rfloor$ being odd implies $2\left\lfloor \frac{dk_s}{2} \right\rfloor + 1 = \lfloor dk_s \rfloor$ (cf. proof of Lemma 1).

Hence, in this case we conclude that all locations $x \in [x_2, x_1)$ are covered by one additional ambulance than the locations in the interval $[0, x_2)$, leading to a coverage $1-(1-q)^{\lfloor dk_s \rfloor + 1}$ if $x \in [x_2, x_1)$.

Similarly, in this case all locations $x \in [x_1, 1/k_s)$ are covered by one less ambulance than the locations in the interval $[x_2, x_1)$, leading to a coverage $1-(1-q)^{\lfloor dk_s \rfloor}$ if $x \in [x_1, 1/k_s)$, concluding the proof when $\lfloor dk_s \rfloor$ is odd, and thus the proof of the proposition. □

We are now ready to prove the main result in this section.

Proof of Theorem 1. First assume that $dk_s < 1$, then $\lfloor dk_s \rfloor = 0$ and the formula in the theorem correctly evaluates to $TC_s(k_s) = dk_s q$. Indeed, if the ambulances locate themselves symmetrically around the circle, then the assumption $dk_s < 1$ implies that their coverage areas do not intersect; hence, each of the k_s ambulances provides a coverage q to a separate region of length d in the circle, for a total coverage $dk_s q$.

Now assume $dk_s \geq 1$ and note that, by symmetry, the total coverage is the same as the average coverage attained on the interval $[0, 1/k_s)$ starting from an ambulance location.

If $\lfloor dk_s \rfloor$ is even, then from Proposition 2 the total coverage is

$$\begin{aligned} TC_s(k_s) &= \frac{\frac{dk_s - \lfloor dk_s \rfloor}{2k_s}(1-(1-q)^{\lfloor dk_s \rfloor + 1}) + \frac{\lfloor dk_s \rfloor + 1 - dk_s}{k_s}(1-(1-q)^{\lfloor dk_s \rfloor}) + \frac{dk_s - \lfloor dk_s \rfloor}{2k_s}(1-(1-q)^{\lfloor dk_s \rfloor + 1})}{1/k_s} \\ &= (dk_s - \lfloor dk_s \rfloor)(1-(1-q)^{\lfloor dk_s \rfloor + 1}) + (1-(dk_s - \lfloor dk_s \rfloor))(1-(1-q)^{\lfloor dk_s \rfloor}), \end{aligned}$$

completing the proof when $dk_s \geq 1$ and $\lfloor dk_s \rfloor$ is even.

Finally, if $\lfloor dk_s \rfloor$ is odd, then from Proposition 2 the total coverage is

$$\begin{aligned} TC_s(k_s) &= \frac{\frac{\lfloor dk_s \rfloor + 1 - dk_s}{2k_s}(1-(1-q)^{\lfloor dk_s \rfloor}) + \frac{dk_s - \lfloor dk_s \rfloor}{k_s}(1-(1-q)^{\lfloor dk_s \rfloor + 1}) + \frac{\lfloor dk_s \rfloor + 1 - dk_s}{2k_s}(1-(1-q)^{\lfloor dk_s \rfloor})}{1/k_s} \\ &= (dk_s - \lfloor dk_s \rfloor)(1-(1-q)^{\lfloor dk_s \rfloor + 1}) + (1-(dk_s - \lfloor dk_s \rfloor))(1-(1-q)^{\lfloor dk_s \rfloor}), \end{aligned}$$

completing the proof when $dk_s \geq 1$ and $\lfloor dk_s \rfloor$ is odd, thus the proof of the theorem. □

8.3. Pure Stationed Ambulances Fleet Equilibrium

PROPOSITION 3. *For any d, $q \in (0,1)$, $a > 0$, there exists a unique symmetric free-entry equilibrium number of stationed ambulances $\hat{k}_s \geq 0$ such that $\Pi(\hat{k}_s) = 0$. Further,*

(i) *The equilibrium is positive, i.e., $\hat{k}_s > 0$, if and only if $a < d$. If $a \geq d$, then $\hat{k}_s = 0$.*

(ii) *If $a < d$, then $\hat{k}_s$ is continuous, strictly decreasing in a and q, and continuous strictly increasing in d.*

Proof of Proposition 3. We first show part (i) of the proposition.

Note that $\Pi_s(k_s)$ is continuous, constant for $k_s \in (0, 1/d)$, and strictly decreasing for $k_s \in (1/d, \infty)$. Indeed, it is continuous within all the intervals $(l/d, (l+1)/d)$, $l \in \mathbb{N}_0$, and at the breakpoints the left and right limits coincide, so the whole function is continuous. Moreover, $\Pi_s(k_s)$ is differentiable for all $(l/d, (l+1)/d)$, $l \in \mathbb{N}_0$, and

$$\Pi_s'(k_s) = \frac{TC_s'(k_s)k_s - TC_s(k_s)}{k_s^2} \begin{cases} = 0, & \text{if } k_s \in (0, 1/d) \\ < 0, & \text{if } k_s \in (l/d, (l+1)/d) \text{ for any } l \in \mathbb{N}, \end{cases} \tag{9}$$

where the cases follow since $TC_s(k_s)$ is piece-wise linear concave and $TC_s(0)=0$. Indeed, for all $k_s \in (l/d,(l+1)/d)$, $l \in \mathbb{N}_0$, $TC_s'(k_s)=(1-q)^{\lfloor dk_s \rfloor}dq$ is decreasing in k_s. Hence, $TC_s'(k_s)k_s = TC_s(k_s)$ for all $k_s \in (0,1/d)$, and $TC_s'(k_s)k_s < TC_s(k_s)$ for all $k_s \in (l/d,(l+1)/d)$, $l \in \mathbb{N}$.

We now characterize the values attained by $\Pi_s(k_s)$ at the extremes of the interval $(0,\infty)$. Indeed, using L'Hôpital's rule,

$$\lim_{k_s \downarrow 0} \Pi_s(k_s) = \lim_{k_s \downarrow 0} \frac{(1-q)^{\lfloor dk_s \rfloor}dq}{1} - aq = (d-a)q, \qquad \lim_{k_s \to \infty} \Pi_s(k_s) = -aq < 0.$$

Therefore, we conclude that if $a \geq d$, then $\Pi_s(k_s) \leq 0$ for all $k_s > 0$, so the unique free–entry equilibrium is that no ambulances enter the market, i.e., $k_s^* = 0$.

Further, if $a < d$ then $\Pi_s(k_s)$ is positive for $k_s \in (0,1/d)$ and negative for large k_s. Continuity and strict monotonicity of $\Pi_s(k_s)$ for $k_s \in (1/d,\infty)$ imply exactly one root $\hat{k}_s > 1/d$ satisfying $\Pi_s(\hat{k}_s)=0$, completing the proof of part (i) of the proposition.

We now prove part (ii) of the proposition.

Assume $a < d$ and define $F(k_s,a,q,d) := \frac{TC_s(k_s)}{k_s} - aq = \frac{1-(1-q)^{\lfloor dk_s \rfloor}(1-(dk_s-\lfloor dk_s \rfloor)q)}{k_s} - aq$. Then $F(\hat{k}_s,a,q,d)=0$ for a unique $\hat{k}_s \in (1/d,\infty)$ and,

$$\frac{\partial \hat{k}_s}{\partial a} = -\frac{\partial F/\partial a}{\partial F/\partial k_s} = -\frac{-q}{\Pi_s'(\hat{k}_s)} < 0,$$

where the first equality follows from the Implicit Function Theorem, and the inequality follows from (9) and $\hat{k}_s \in (1/d,\infty)$. Similarly,

$$\frac{\partial \hat{k}_s}{\partial d} = -\frac{\partial F/\partial d}{\partial F/\partial k_s} = -\frac{(1-q)^{\lfloor d\hat{k}_s \rfloor}q}{\Pi_s'(\hat{k}_s)} > 0,$$

where the first equality follows from the Implicit Function Theorem, and the inequality follows from (9) and $\hat{k}_s \in (1/d,\infty)$. Finally,

$$\frac{\partial \hat{k}_s}{\partial q} = -\frac{\partial F/\partial q}{\partial F/\partial k_s} = -\frac{\frac{\frac{\partial TC_s(\hat{k}_s)}{\partial q}}{\hat{k}_s} - a}{\Pi_s'(\hat{k}_s)} = -\frac{\frac{\partial TC_s(\hat{k}_s)}{\partial q} - \frac{TC_s(\hat{k}_s)}{q}}{\hat{k}_s \Pi_s'(\hat{k}_s)} < 0, \tag{10}$$

where the first equality follows from the Implicit Function Theorem, the third equality follows from $F(\hat{k}_s,a,q,d)=0$, i.e., $TC_s(\hat{k}_s) = aq\hat{k}_s$, and the inequality follows from (9), $\hat{k}_s \in (1/d,\infty)$, and since $TC_s(k^s)$ is concave in q for all $k_s \in (1/d,\infty)$, hence $\frac{\partial TC_s(\hat{k}_s)}{\partial q}q < TC_s(\hat{k}_s)$ for all $k_s \in (1/d,\infty)$. Indeed,

$$\begin{aligned}\frac{\partial^2 TC_s(k_s)}{\partial q^2} =& -(dk_s - \lfloor dk_s \rfloor)(\lfloor dk_s \rfloor + 1)\lfloor dk_s \rfloor (1-q)^{\lfloor dk_s \rfloor - 1} \\ & -(1-(dk_s-\lfloor dk_s \rfloor))\lfloor dk_s \rfloor (\lfloor dk_s \rfloor - 1)(1-q)^{\lfloor dk_s \rfloor - 2}) < 0, \text{ for all } k_s \in (1/d,\infty),\end{aligned}$$

completing the proof. □

8.4. Proof of Theorem 2

We start by stating and proving three technical results that will allow us to prove Theorem 2.

PROPOSITION 4. *For any $d,\ q \in (0,1)$, and $0 < a < d$, let $\hat{k}_s(d,q,a)$ denote the free-entry equilibrium fleet size of stationed ambulances and $TC_s^*(d,q,a) := TC_s(\hat{k}_s(d,q,a))$ denote their induced total coverage. Then, $\frac{\partial TC_s^*(d,q,a)}{\partial q} > 0$.*

Proof of Proposition 4. For any $d,\ q \in (0,1)$ and $0 < a < d$, recall from Proposition 3 that the free entry equilibrium condition for stationed ambulances, $\Pi_s(\hat{k}_s) = 0$, is equivalent to $TC_s(\hat{k}_s) = aq\hat{k}_s$, and it is satisfied by a unique $\hat{k}_s \in (1/d, \infty)$. Hence,

$$\frac{\partial TC_s^*(d,q,a)}{\partial q} = \frac{\partial aq\hat{k}_s}{\partial q} = a\hat{k}_s + aq\frac{\partial \hat{k}_s}{\partial q}. \tag{11}$$

Therefore, we conclude the following chain of equivalences,

$$\begin{aligned}
&\frac{\partial TC_s^*(d,q,a)}{\partial q} > 0 \\
\iff &\frac{\partial \hat{k}_s}{\partial q} = -\frac{\frac{\partial TC_s(\hat{k}_s)}{\partial q} - \frac{TC_s(\hat{k}_s)}{q}}{\hat{k}_s\Pi_s'(\hat{k}_s)} > -\frac{\hat{k}_s}{q} \\
\iff &q\left(\frac{\partial TC_s(\hat{k}_s)}{\partial q} - \frac{TC_s(\hat{k}_s)}{q}\right) > \hat{k}_s^2\Pi_s'(\hat{k}_s) = TC_s'(\hat{k}_s)\hat{k}_s - TC_s(\hat{k}_s) \\
\iff &q\frac{\partial TC_s(\hat{k}_s)}{\partial q} > TC_s'(\hat{k}_s)\hat{k}_s \\
\iff &\lfloor d\hat{k}_s\rfloor(1-q)^{\lfloor d\hat{k}_s\rfloor - 1}\left(1 - (d\hat{k}_s - \lfloor d\hat{k}_s\rfloor)q\right) + (1-q)^{\lfloor d\hat{k}_s\rfloor}(d\hat{k}_s - \lfloor d\hat{k}_s\rfloor) > (1-q)^{\lfloor dk_s\rfloor}d\hat{k}_s \\
\iff &d\hat{k}_s < \lfloor d\hat{k}_s\rfloor + 1,
\end{aligned}$$

where the first equivalence follows from (11), and the first equality follows from (10). The second equivalence follows from rearranging terms, and the second equality follows from (9). The third equivalence follows from eliminating the common term $TC_s(\hat{k}_s)$. The fourth equivalence follows from substituting the expressions for $\frac{\partial TC_s(\hat{k}_s)}{\partial q}$ and $TC_s'(\hat{k}_s)$. Finally, the last equivalence follows from simplifying the expression, and the last inequality always holds, completing the proof. □

LEMMA 2. *For any $d \in (0,1)$ and $x \geq 0$,*

$$(x - \lfloor x\rfloor)(1-(1-d)^{\lfloor x\rfloor+1}) + (\lfloor x\rfloor + 1 - x)(1-(1-d)^{\lfloor x\rfloor}) \leq 1 - (1-d)^x, \tag{12}$$

and

$$(x - \lfloor x\rfloor)(1-(1-d)^{\lfloor x\rfloor+1}) + (\lfloor x\rfloor + 1 - x)(1-(1-d)^{\lfloor x\rfloor}) = 1 - (1-d)^x \iff x \in \mathbb{N}. \tag{13}$$

Proof of Lemma 2. Let $h(x) := 1-(1-d)^x$ and note that

$$(12) \iff (x-\lfloor x\rfloor)h(\lfloor x\rfloor+1)+(\lfloor x\rfloor+1-x)h(\lfloor x\rfloor)\le h(x).$$

Moreover, note that x can be expressed as a convex combination of $\lfloor x\rfloor$ and $\lfloor x\rfloor+1$, with weights $(\lfloor x\rfloor+1-x)$ and $(x-\lfloor x\rfloor)$, respectively, i.e., $x=(x-\lfloor x\rfloor)(\lfloor x\rfloor+1)+(\lfloor x\rfloor+1-x)\lfloor x\rfloor$. Then, (12) follows since $h(x)$ is concave. Indeed, $h''(x)=-(1-d)^x\log^2(1-d)<0$ for any $d\in(0,1)$ and $x\ge 0$.

Further, since $h(x)$ is strictly concave, then the inequality (12) is strict whenever both $\lfloor x\rfloor$ and $\lfloor x\rfloor+1$ have positive weights in the convex combination, i.e., whenever $(\lfloor x\rfloor+1-x)>0$ and $(x-\lfloor x\rfloor)>0$, or equivalently $\lfloor x\rfloor<x<\lfloor x\rfloor+1$, namely whenever $x\notin\mathbb{N}$, proving (13) and completing the proof of the lemma. □

PROPOSITION 5. *For any $d\in(0,1)$ and $0<a<d$, if $q=d$, then $TC_r(\hat{k}_r)\ge TC_s(\hat{k}_s)$, with equality holding if and only if $d\hat{k}_s\in\mathbb{N}$.*

Proof of Proposition 5. For any $d\in(0,1)$ and $0<a<d<-\log(1-d)$, recall from Proposition 1 that the free entry equilibrium condition for roaming ambulances, $\Pi_r(\hat{k}_r)=0$, is equivalent to

$$TC_r(\hat{k}_r)=1-(1-d)^{\hat{k}_r}=a\hat{k}_r, \tag{14}$$

and it is satisfied by a unique $\hat{k}_r\in(0,\infty)$.

For any $d\in(0,1)$ and $0<a<d$, if $q=d$ then note from Proposition 3 that the free entry equilibrium condition for stationed ambulances, $\Pi_s(\hat{k}_s)=0$, is equivalent to

$$TC_s(\hat{k}_s)=(d\hat{k}_s-\lfloor d\hat{k}_s\rfloor)(1-(1-d)^{\lfloor d\hat{k}_s\rfloor+1})+(\lfloor d\hat{k}_s\rfloor+1-d\hat{k}_s)(1-(1-d)^{\lfloor d\hat{k}_s\rfloor})=ad\hat{k}_s, \tag{15}$$

and it is satisfied by a unique $\hat{k}_s\in(1/d,\infty)$.

To conclude, note that

$$\begin{aligned} ad\hat{k}_s &= (d\hat{k}_s-\lfloor d\hat{k}_s\rfloor)(1-(1-d)^{\lfloor d\hat{k}_s\rfloor+1})+(\lfloor d\hat{k}_s\rfloor+1-d\hat{k}_s)(1-(1-d)^{\lfloor d\hat{k}_s\rfloor}) \\ &\le 1-(1-d)^{d\hat{k}_s}=TC_r(d\hat{k}_s), \end{aligned} \tag{16}$$

where the first equality follows from (15), and the inequality follows from Lemma 2. By definition, (16) is equivalent to $\Pi_r(d\hat{k}_s)\ge 0$, and since $\Pi_r(k_r)$ is strictly decreasing from (8), we conclude that $\hat{k}_r\ge d\hat{k}_s$, or equivalently from (14) and (15), $TC_r(\hat{k}_r)\ge TC_s(\hat{k}_s)$. From Lemma 2, the same analysis holds with equality if and only if $d\hat{k}_s\in\mathbb{N}$, completing the proof of the proposition. □

We are now ready to prove the main result in this section.

Proof of Theorem 2. Part (i) of the theorem follows from Propositions 1(i) and 3(i), and the observation that $d < -\log(1-d)$ for all $d \in (0,1)$.

We now prove part (ii) of the theorem.

For any $d \in (0,1)$ and $0 < a < d$, let $\Delta(q) := TC_s^*(d,q,a) - TC_r^*(d,a)$ for $q \in (0,1)$. Note that $\Delta(q)$ is continuous and strictly increasing for $q \in (0,1)$. Indeed, $TC_s(k_s)$ is continuous by definition, from Proposition 3(ii) we have that $\hat{k}_s(d,q,a)$ is continuous in q when $a < d$, and $TC_r^*(d,a)$ is independent of q, thus $\Delta(q)$ is continuous. Further, $\Delta'(q) > 0$ since from Proposition 4 we have $\frac{\partial TC_s^*(d,q,a)}{\partial q} > 0$ and $TC_r^*(d,a)$ is independent of q.

We now characterize the values attained by $\Delta(q)$ at the extremes of the interval $(0,1)$. Recall from Proposition 1 that if $a < d < -\log(1-d)$ then $\hat{k}_r > 0$ and thus $TC_r^*(d,a) \in (0,1)$ by definition. Then, we conclude that $\Delta(0) = -TC_r^*(d,a) < 0$ and $\Delta(1) = 1 - TC_r^*(d,a) > 0$.

Hence, if $a < d$ then the continuity and strict monotonicity of $\Delta(q)$ imply exactly one root $\bar{q}(d,a) \in (0,1)$ such that $\Delta(\bar{q}(d,a)) = 0$, and $\Delta(q) > 0 \iff q > \bar{q}(d,a)$.

Further, by setting $q = d$ we obtain $\Delta(d) = TC_s^*(d,d,a) - TC_r^*(d,a) \leq 0$, where inequality follows from Proposition 5, with equality holding if and only if $d\hat{k}_s \in \mathbb{N}$. Therefore, we conclude $d \leq \bar{q}(d,a)$, with equality holding if and only if $d\hat{k}_s \in \mathbb{N}$ (proving Corollary 1), completing the proof of the theorem. □

Proof of Corollary 1. Provided within the proof of Theorem 2. □

8.5. Equilibrium with both roaming and stationed ambulances

Assume that k_r roaming ambulances and k_s stationed ambulances join the market equilibrium, then the profits of each of the k_r roaming ambulances are then given by

$$\Pi_r(k_r,k_s) =$$
$$\frac{(dk_s - \lfloor dk_s \rfloor)}{k_r} \sum_{i=0}^{k_r} \sum_{j=0}^{\lfloor dk_s \rfloor + 1} \binom{k_r}{i} d^i (1-d)^{k_r - i} \binom{\lfloor dk_s \rfloor + 1}{j} q^j (1-q)^{\lfloor dk_s \rfloor + 1 - j} \frac{i}{i + j + \mathbb{1}_{\{i=0, j=0\}}}$$
$$+ \frac{(\lfloor dk_s \rfloor + 1 - dk_s)}{k_r} \sum_{i=0}^{k_r} \sum_{j=0}^{\lfloor dk_s \rfloor} \binom{k_r}{i} d^i (1-d)^{k_r - i} \binom{\lfloor dk_s \rfloor}{j} q^j (1-q)^{\lfloor dk_s \rfloor - j} \frac{i}{i + j + \mathbb{1}_{\{i=0, j=0\}}} - a,$$

where we use the fact that stationed ambulances are located equidistantly in the circle in a symmetric equilibrium and the assumption that if more than one ambulance is within reach of a call, the ambulance that responds to it is selected at random among them.

Similarly, the profits of each of the k_s stationed ambulances are then given by

$$\Pi_s(k_r,k_s) =$$
$$\frac{(dk_s - \lfloor dk_s \rfloor)}{k_s} \sum_{i=0}^{k_r} \sum_{j=0}^{\lfloor dk_s \rfloor + 1} \binom{k_r}{i} d^i (1-d)^{k_r - i} \binom{\lfloor dk_s \rfloor + 1}{j} q^j (1-q)^{\lfloor dk_s \rfloor + 1 - j} \frac{j}{i + j + \mathbb{1}_{\{i=0, j=0\}}}$$

$$+\frac{(\lfloor dk_s \rfloor + 1 - dk_s)}{k_s} \sum_{i=0}^{k_r} \sum_{j=0}^{\lfloor dk_s \rfloor} \binom{k_r}{i} d^i (1-d)^{k_r - i} \binom{\lfloor dk_s \rfloor}{j} q^j (1-q)^{\lfloor dk_s \rfloor - j} \frac{j}{i+j+\mathbb{1}_{\{i=0,j=0\}}} - aq,$$

where we again use the fact that stationed ambulances are located equidistantly in the circle in a symmetric equilibrium and the assumption that if more than one ambulance is within reach of a call, the ambulance that responds to it is selected at random among them.

Therefore, the number of ambulances of each type that join the free-entry market equilibrium is given by the values k_r^* and k_s^* such that $\Pi_r(k_s^*, k_r^*) = 0$ and $\Pi_s(k_s^*, k_r^*) = 0$ simultaneously hold, which is analytically intractable.