International Journal of Paramedicine

Published by the National EMS Management Association (USA)

RESEARCH REPORTS

EXPERIENCED AGENCIES, FASTER CHUTE TIMES: DETERMINANTS OF EMS RESPONSE IN THE UNITED STATES

Jennifer Ziemke, PhD, EMT-P*[1,2]

Author Affiliations: 1. Political Science, John Carroll University, University Heights, OH, USA; 2. Lifecare Ambulance, Elyria, OH, USA.

Recommended Citation: Ziemke, J. (2026). Experienced agencies, faster chute times: Determinants of EMS response in the United States. International Journal of Paramedicine. (15). 53-69. https://doi.org/10.56068/DKGJ4494

Keywords: NEMSIS, response time, chute, ambulance, agency, experience, fixed effects, emergency medical services, EMS, paramedicine

Acknowledgements: Dedicated to Aaron and for all the medics who taught us and continue to shape us in the field.

Disclaimer: Ethics approval: This study only analyzed publicly-available, de-identified data, and did not meet the criteria for IRB review according to 45 CFR 46.102.

Disclosures: None.

Funding: External funding was not used to support this work.

Received: December 14, 2025
Revised: March 20, 2026
Accepted: March 24, 2026
Published: July 15, 2026

**Corresponding Author:*
jziemke@jcu.edu

ABSTRACT

Why do emergency medical units respond more rapidly to some dispatched calls than others? This study examines over 355 million emergency medical services (EMS) activations recorded in the National Emergency Medical Services Information System (NEMSIS) database between 2018 and 2023. The results show that teams operating in the top 5% of agencies by call volume—representing the most experienced systems—initiate response more quickly and post consistently shorter chute times.

Experienced responders more quickly recognize the nature of the call, develop confidence in both the likely course of events and their own ability to manage it, and initiate mobilization with minimal hesitation. As these behaviors are repeated, they become embedded in routines, coordination, and shared expectations. In this way, experienced responders shape agency-level culture from the ground up, contributing to persistent differences in performance across agencies, captured in agency-level fixed effects.

INTRODUCTION

What explains the variation in the amount of time it takes emergency medical responders to begin responding once dispatched? This question is examined using NEMSIS data comprising over 355 million 911 activations routed to EMS in the United States between 2018-2023. This study shows that teams operating within the highest call volume agencies—those in the top 5% of total volume—more quickly initiate their response, resulting in substantially shorter EMS chute times. Through repeated exposure, responders learn to anticipate call demands, coordinate more effectively, and initiate mobilization with fewer delays—patterns that become visible in the timestamp data as faster and more consistent chute times.

In practice, first responders must rapidly interpret limited dispatch information (e.g., age and call type) to decide how urgently to mobilize. These split-second judgments determine whether crews depart immediately or delay briefly (for instance, to finish washing their hands in the bathroom) before responding. To what extent do patient factors and dispatch call type help predict the speed of response? After all, a call for a child in full arrest might yield a slightly different behavior pattern than for abdominal pain.

SECONDS MATTER

For true emergencies, every second matters. For example, consider a patient in cardiac arrest. Given that "shockable" arrests presenting with Ventricular Tachycardia or Ventricular Fibrillation only "stay shockable" for a limited time before progressing to asystole or Pulseless Electrical Activity arrest (which are not shockable rhythms and where the odds of survival are much lower), being on time to deliver a shock can mean life or death for a patient in cardiac arrest. One study found that for every additional minute it takes for an ambulance to respond, the odds of a patient presenting with a shockable rhythm decrease by 8% per minute. (Renkiewicz, et al., 2014). In another study, the odds of surviving a cardiac arrest increase 24% if ambulances respond 1 minute faster than the mean (O'Keeffe et al., 2011).

Other studies emphasize the importance of response speed after critical MVAs (Motor Vehicle Accidents), where delays can be fatal due to rapid hemorrhage Helander et al., 2024). A growing body of research demonstrates how even short delays in care can impact patient outcomes. For instance, during cerebrovascular events (CVAs), a patient experiencing a typical ischemic stroke loses 1.9 million neurons every minute. Over that same minute, an estimated 14 billion synapses and 7.5 miles of myelinated fibers are also lost (Saver, 2006).

LITERATURE

A number of studies examine the length of time it takes for an ambulance to respond. This variable is referred to in several different ways, including squad chute time, call-response time, activation time, Ambulance Response Time, response time, and EMS Chute Initiation Time (Adeyemi et al., 2022, Medrano et al., 2022). Explanations for why EMS practitioners respond to certain calls more quickly than others are broadly divided in the literature into patient-level and system-level factors. (Abbas et al., 2022, Do et al., 2013, Nehme et al., 2016).

Patient-level factors include the information regularly dispatched to a responding unit: "Medic 1, I show you responding to 123 Main Street for a 27-year-old female, unresponsive." With this information, both dispatchers and first responders share information on the nature of the call, the age of the patient, and the gender of the patient. The dataset does not contain the dispatched address as doing so would compromise patient privacy; in addition, national changes in how and the extent to which gender is coded in each year necessitated dropping gender as a variable from the analysis.

Literature in this field indicates there are also systemic factors that can help explain variation in EMS chute times, like season, time of day, geographic region and urbanicity score. For example, a recent multivariate analysis of response times in China demonstrates how the pace of unit mobilization changes across the seasons (Chen et al., 2024). Temporal patterns in emergency call data reveal a consistent nationwide diurnal rhythm in call volume. For example, Helander et al. find that, regardless of dispatched call type, nationwide call volume consistently peaks around 15:00 local time (2024). Exceptions to this pattern include cases of: overdoses, which peak later, around 19:00, stabbings and shootings, which peak around 20:00, and assaults (peaking around 20:30, local time). Previous research has also found that the pace of 911 response is intimately linked to overall call volume at the agency level. Some studies have teased out how average hospital delay times in the previous hour as well as the workload of the squad in the previous hour impacts the pace of response (Nehme, et al., 2016). Finally, literature from around the world has deployed location-based analyses to test a number of hypotheses. Some studies examine how the geographic density of emergency stations in an area influences response time. Others use geography and income together to help tease out effects. For example, Hsia et al. found that the average income of a zip code helped predict the speed of an ambulance's response to a full arrest. Ambulances respond more slowly to 911 emergencies in poor areas, especially when compared with their counterparts working in wealthy communities (2018, see also Chen et al. 2024 and Earnest et al., 2011). While prior work has examined both patient and system-level constraints, less attention has been paid to how first responders accumulate knowledge about EMS gained through repeated exposure to different kinds of emergency calls, and how this learning translates into faster and more consistent response times. This study builds on this gap by examining whether agency-level experience—measured through call volume—serves as a primary determinant of EMS chute times.

A SYSTEM UNDER STRESS

More broadly, this analysis is motivated by a desire to better understand what systems do when placed under incredible strain. The emergency response system was designed so those experiencing true life-threatening emergencies could receive pre-hospital care and quick transit to the hospital. Instead, many service areas often find themselves inundated with calls that are not true emergencies. Indeed, many calls occur as a direct outcome of what is widely noted as widespread systemic failures of the health care system. Unable to pay for your medications and now feel sick? Don't have a ride to the pharmacy to pick up the medications prescribed to you in the emergency department last week? Tired of waiting in pain and discomfort over 3 months for a primary care visit, only to be directed to a series of specialists that are scheduling another 3 months out? Enough is enough, their family says. Call an ambulance.

First responders and emergency room employees operate under significant personnel and resource constraints; they work long hours and often do what should be the job of 2 or 3 others. On many occasions, all beds in the emergency room are full, often with a substantial number of people queuing to be triaged in the waiting room. For their part, pre-hospital first responders go from one call to another, often without time to start their report or adequately restock the ambulance before being paged to see if they are available to take the next.

To make matters worse, many of the people being seen are well known to the first responders and the hospital staff that greet them. A relatively limited number of people account for a substantial number of EMS calls. Studies examining the abuse of the 911 system by "frequent fliers" fill an important gap in our understanding of this problem (Gehrt 2025).

ABUSE OF THE 911 SYSTEM

While NEMSIS data do not collect "frequent flier" data directly, it is plausible that, taken together with the aforementioned stressors on the EMS system, first responders would move more slowly when dispatched to most frequent call types, due to call type burnout ("one of those calls again"). In much the same way that responders might groan when being dispatched to a frequent flier, one may hypothesize that high-frequency call types might yield a similar kind of sentiment, especially because over 40% of all EMS calls come from just three categories: Sick Person (20%), Respiratory Issues (11%), and Falls (10%), as Table 1 shows.

Dispatch Call Type	Frequency	Percent
Sick Person	41,153	20.1%
Respiratory	22,311	10.9%
Falls	20,633	10.1%
Chest Pain	14,455	7.1%
Unclear	12,389	6.0%
Traffic Incident	11,875	5.8%
Unconscious	10,852	5.3%
Abdominal Pain/Problems	7,678	3.7%
Seizure	7,449	3.6%
Psych	7,274	3.5%
Unknown/Person Down	6,536	3.2%
Traumatic Injury	5,511	2.7%
Stroke	5,172	2.5%
Hemorrhage/Laceration	4,600	2.2%
Overdose	4,467	2.2%
Heart Problems	3,238	1.6%
Diabetic	2,909	1.4%
Back Pain	2,554	1.2%
Cardiac Arrest	2,473	1.2%
Assault	2,445	1.2%
Allergic Reaction	1,537	0.7%
OBGYN	1,393	0.7%
Headache	1,219	0.6%
Stab/GSW	921	0.4%

Table 1. Distribution of top 25 EMS dispatch call types.

Next, the paper details all model variables and specifies hypotheses to be formally tested.

The dependent variable, Squad Response Time, is measured as the time elapsed between when a unit is assigned a call and the unit's departure, in seconds.

HYPOTHESES

Attributes of Responding Unit

H1: "Top 5% Experienced Agency": Agencies with greater call volume have more operational experience and are expected to demonstrate faster chute times.

An agency-level dataset was constructed using deidentified keys linked to each agency. Teams operating within agencies in the top 5% of call volume in the sample are classified as "most experienced" under a binary threshold model. While experience cannot be directly observed, agencies that handle more calls are exposed to a wider range of emergency scenarios, allowing them to develop and reinforce routines, coordination, and institutional knowledge. Call volume therefore provides a practical and observable proxy for accumulated operational experience.

Truck Disposition at Dispatch

H2: "In-Vehicle When Dispatched": To account for operational differences in unit status at the time of dispatch, a binary variable identifies cases in which first responders are already in the response vehicle when the call is received. All observations with zero-second chute times are coded as 1 for this variable. This control distinguishes true mobilization performance from cases in which no mobilization phase is required and accounts for the possibility that high-volume agencies may be more likely to have units already in motion.

Call Type Characteristics

H3: "Dispatched Call Type": Higher-acuity dispatch types (e.g. Full Arrest, Choking, and Stabbing/GSW) will be associated with shorter chute times relative to lower-acuity calls (e.g., Back Pain). Each dispatch call type is included as an independent predictor of chute time. "Abdominal Pain/Problems" is set as the reference category.

H4: "Call Type Fatigue": Response times will be longer for the most common call types (Respiratory, Falls, and Sick Person) than for other call types, reflecting potential fatigue with these call types.

Characteristics Relayed to EMS

H5: "Patient Age": Response times will increase with dispatch-reported patient age.

Timing of Dispatched Response

H6: "COVID Period": Response times will be longer during the period set equal to 1 for all dates inclusive of March 11, 2020 and May 1, 2023, marking the beginning and end of CDC-announced COVID-era restrictions.

H7: "Night Response": Response times will be longer during overnight hours. Coded as 1 for calls occurring between midnight and 07:30, and 0 otherwise.

H8: "Weekend Response": Response times will be longer on weekends than weekdays. Coded 1 for calls taking place on Saturday and Sunday and 0 otherwise.

Other Control Variables

H9: "Region of EMS Response": NEMSIS deidentifies location data to preserve patient privacy. However, each call is coded for region (South, East, West, Great Lakes, or Western Plains) and urbanicity (Wilderness, Rural, Suburban or Urban). Dummy variables are included for Western, Southern, Western Plains, and Great Lakes regions, with the East set as the reference category.

H10: "Urbanicity": Dummy variables are included for Suburban, Rural, and Wilderness areas with Urban set as the reference category.

Region and location variables act as important controls given the geographic distribution of EMS calls and resources.

What stands out immediately after exploration of Figure 1 is that most of the calls in the United States take place in urban areas with high-call volume environments, where resources are more likely to be constrained and demand high. This concentration of calls in urban environments suggests that many agencies operate under conditions of sustained demand, requiring a multivariate model to control for the independent effect of urbanicity on squad response time. To evaluate these hypotheses, two complementary modeling approaches are used using data described below.

DATA AND SAMPLING

First responders generate Patient Care reports (PCRs) with detailed information about each patient interaction. Stations submit these reports to the National EMS Information

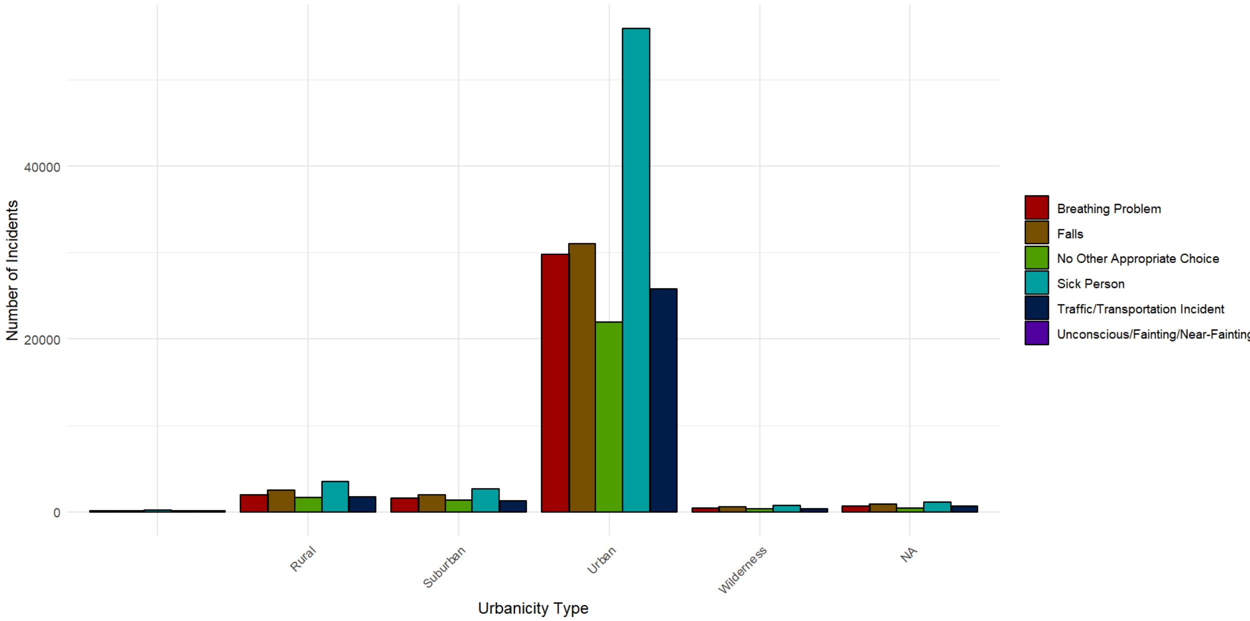

Figure 1. Top five dispatched call types by urbanicity score

System (National Association of State EMS Officials, 2024). Each 911 call in the database includes timestamps (specific to the second) that indicate the local time that the trunk system alerted local EMS dispatchers, the time the call was dispatched to a specific responding unit, and the time that first responders marked themselves en route toward the dispatched location.

The NEMSIS repository contains hundreds of millions of 911 ambulance patient care reports (PCRs), organized by year into approximately 40 relational tables. Each table contains hundreds of interrelated variables. Given the sheer complexity and raw size of the dataset, conventional approaches to random sampling that involve first reading the dataset into memory are unwieldy. To construct a representative analytic sample from the full dataset, reservoir sampling was implemented. As records were read sequentially in batches, candidate observations were probabilistically retained or replaced within a fixed-size reservoir, ensuring that each observation in the full dataset had an equal probability of selection without requiring full in-memory storage.

To begin, 100,000 records per year were selected from the 354.8 million rows over the years 2018-2023, and then a master file was created that appended all years to one another. Next, a second sample of 100,000 observations was taken from each year and appended to the combined dataset. Table 2 describes the workflow that led to the final analytic sample and Table 3 presents descriptive statistics for the main variables used in the analysis.

MODEL ESTIMATION

The distribution of the dependent variable, Squad Response Time, is strongly right skewed with a long upper tail, consistent with the asymmetric nature of operational time data. Given this skewness and the presence of extreme values, robust modeling approaches were used, and observations exceeding 10 minutes were excluded to limit the influence of atypical cases that likely reflected non-routine operational conditions and coding errors rather than typical response behavior.

NEMSIS Dataset (≈354 million records)
↓
Reservoir Sample (≈600,000 records)
↓
Valid Response Times (0 < time ≤ 600 seconds) (N ≈ 205,000 records)
↓
Final Analytic Sample (striking covariate missing values) (N ≈ 185k (Model 1), N ≈ 183.5k (Model 2)

Table 2. Data construction workflow.

RANDOM LINEAR MODEL AND FIXED EFFECTS OLS MODELS

The logic described in Pek et al.'s study guided model selection (2018). First, a baseline multivariate model using a robust linear approach was estimated using RLM in R (Model 1). Electing to use robust models helps us rein in those extreme observations (e.g., the handful of extreme outliers where chute times are measured in hours or even days rather than seconds) that would otherwise normally substantially influence or skew the results. Second, use of a fixed effects model helps account for expected substantial variation in the dependent variable best explained by agency-level fixed effects, using FEOLS in R (Model 2).

Variable	Mean/%(SD) [Median]
Main Variables	
Squad Response Time (sec)	66.11 (78.81) [50]
Patient Age	55.86 (22.92) [58]
Experienced Agency	56.9%
In Vehicle When Dispatched	11.7%
COVID Period	54.2%
Night Response	18.2%
Weekend Response	27.3%
Urbanicity	
Urban	85.1%
Suburban	5.2%
Rural	6.2%
Wilderness	1.4%
Missing	2.0%
Region	
South	38.8%
East	22.7%
West	19.6%
Great Lakes	11.5%
Western Plains	7.3%

Table 3. Descriptive statistics.

DISCUSSION

Highly Experienced Agencies have the fastest EMS Chute Times

The most significant finding of this research project is that, even when controlling for urbanicity and other factors, **experience is the strongest predictor of the speed of EMS response**. Importantly, these effects persist even after controlling for agency-level fixed effects, indicating that the relationship is not solely driven by differences across agencies.

In a simple bivariate model, agency experience is strongly associated with faster chute times. A one-unit increase in agency experience rank (0–1 scale) is associated with a 135-second reduction in chute time ($p < 0.001$), with experience alone explaining approximately 7.5% of the variation in response times.

In the multivariate model, a one-unit increase in agency experience rank (0–1 scale) is associated with an 87-second reduction in chute time, while a one standard deviation increase corresponds to approximately a 15–16 second reduction in response time.

These results suggest that what is being captured is not simply exposure to a higher volume of calls, but the accumulation of experiential learning in the field. Through repeated exposure, responders learn to anticipate call demands, coordinate more effectively, and

$$\text{ResponseTime}_i = \beta_0 + \beta_1\text{Top5PctAgency}_i + \beta_2\text{InVehicle}_i$$
$$+\beta_3\text{CallTypeFatigue}_i + \sum_k \beta_k\text{CallType}_{ik} + \beta_4\text{Age}_i$$
$$+\beta_5\text{COVID}_i + \beta_6\text{Night}_i + \beta_7\text{Weekend}_i$$
$$+\sum_r \gamma_r\text{Region}_{ir} + \sum_u \delta_u\text{Urbanicity}_{iu} + \varepsilon_i$$

Model 1. Robust linear model.

$$\text{ResponseTime}_i = \beta_0 + \beta_1\text{InVehicle}_i$$
$$+\beta_2\text{CallTypeFatigue}_i + \sum_k \beta_k\text{CallType}_{ik} + \beta_3\text{Age}_i$$
$$+\beta_4\text{COVID}_i + \beta_5\text{Night}_i + \beta_6\text{Weekend}_i$$
$$+\sum_r \gamma_r\text{Region}_{ir} + \sum_u \delta_u\text{Urbanicity}_{iu}$$
$$+\alpha_{\text{agency}(i)} + \varepsilon_i$$

Model 2. OLS with agency-level fixed effects.

initiate mobilization with fewer delays. We posit that these learned behaviors are not directly observed but are revealed in the timestamp data through faster and more consistent chute times, as illustrated in Figure 2.

Figure 2 illustrates the relationship between agency call volume and mean chute time. Each point represents an agency in the dataset. The relationship is nonlinear, with the largest reductions in chute time occurring at lower levels of volume with increasing returns giving way to diminishing returns at higher levels. There is also evidence of a tightening in the distribution of chute times at higher levels of agency volume, with fewer extreme values.

Figure 2 demonstrates that high-volume agencies aren't just faster, they are more consistent in their performance. Taken together, the figure provides visual evidence that accumulated exposure to emergency calls is associated with both faster and more predictable response times.

However, at the most extreme end of total call volume and agency experience, any additional exposure does not significantly impact response speed. Figure 3 presents a binned version of the relationship between agency call volume and mean chute time, helping to

Hypothesis	Variable	Model 1 Robust Linear Model	Model 2 OLS Fixed Effects
H1 Attributes of Responding Unit	Agency experience	-97.68*** (0.76)	—
H2 Truck Disposition at Dispatch	In Vehicle When Dispatched	-70.38***	-86.37***
H3 Call Type as Dispatched	Cardiac Arrest	-4.06*** (1.14)	-6.79*** (1.43)
H4 Call Type Fatigue	Falls	0.79 (0.67)	-1.06 (0.84)
	Respiratory	0.63 (0.66)	-1.666* (0.826)
	Sick Person	0.85 (0.62)	0.66 (0.78)
H5	Patient Age	0.032*** (0.005)	0.008 (0.007)
H6 Timing of Dispatched Response	COVID period	-1.12*** (0.23)	-0.27 (0.31)
H7	Night Response	17.87*** (0.29)	23.22*** (0.36)
H8	Weekend response	0.44 (0.25)	0.03 (0.31)
H9 Region: East is Reference Category	Western US	-16.25*** (0.35)	-2.62 (1.52)
	Southern US	3.61*** (0.30)	-3.73** (1.38)
	Western Plains US	-1.93*** (0.48)	—
	Great Lakes US	-5.13*** (0.40)	—
H10 Dispatched Location Type: Urban is reference category	Suburban area	10.57*** (0.50)	4.83** (1.85)
	Rural area	14.77*** (0.48)	13.14*** (1.69)
	Wilderness area	19.99*** (0.96)	17.30*** (3.46)
		(0.35)	(0.53)
	BIC	2,127,039	—
	Agency fixed effects	No	Yes (5,518 agencies)
	Observations	185,015	183,592
	RMSE	45.58	57.10
	R^2 / Adj. R^2	—	0.432

*Notes. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001. Maximum Variance inflation factors (VIFs) = 1.31, indicating no evidence of multicollinearity. For complete model estimation results, reference Appendix 1.*

Table 4. Determinants of EMS chute time (seconds).

clarify the underlying pattern observed in Figure 2. Mean chute time declines sharply at lower levels of call volume, indicating that early increases in exposure yield substantial gains in mobilization speed. As call volume increases further, this relationship flattens, suggesting diminishing returns to additional experience.

This pattern is consistent with a learning curve: initial exposure to emergency response generates rapid improvements, while additional experience yields smaller, incremental gains. At the highest levels of call volume, the slight plateau suggests that operational constraints may limit further improvements in mobilization speed. Together, Figures 2 and 3 indicate that experience improves response performance, but that the marginal benefit of additional exposure decreases as agencies become highly experienced.

To further clarify this pattern, Figure 3 shows that mean chute times decline sharply at lower levels of volume and then flatten. Figure 4 turns to agency-level fixed effects to capture persistent differences in mobilization performance.

Figure 4 displays the estimated agency-level fixed effects, capturing persistent deviations in chute time across agencies after controlling for observed characteristics. The wide distribution of these effects—ranging from agencies that mobilize substantially faster than

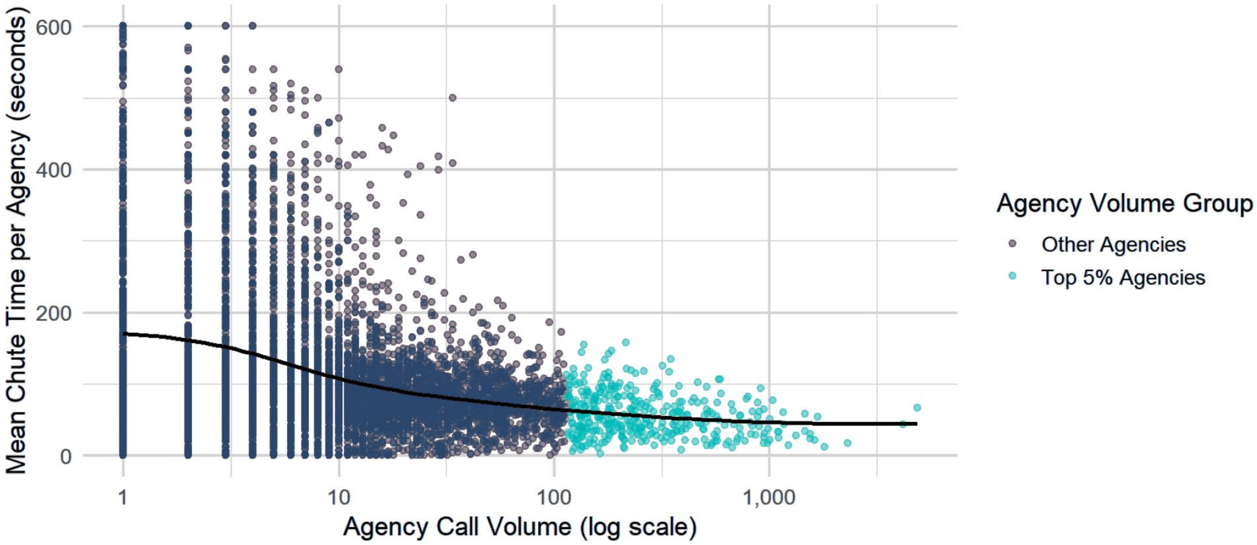

Figure 2. Agency volume vs. chute time.

Note: Each point represents an agency.

average to those that are consistently slower—reveals large and systematic differences in operational performance.

These differences are not explained by call characteristics, geography, or timing, but instead reflect underlying organizational factors such as routines, coordination practices, and institutional norms. In this sense, the fixed effects capture each agency's "operational signature"—a stable pattern of behavior that governs how quickly units mobilize. The magnitude of these differences suggests that organizational factors play a central role in determining EMS response speed. While experience contributes to improved performance, Figure 4 demonstrates that agencies differ in systematic ways beyond observable characteristics, reinforcing the interpretation that learned behaviors become embedded in agency-level culture over time.

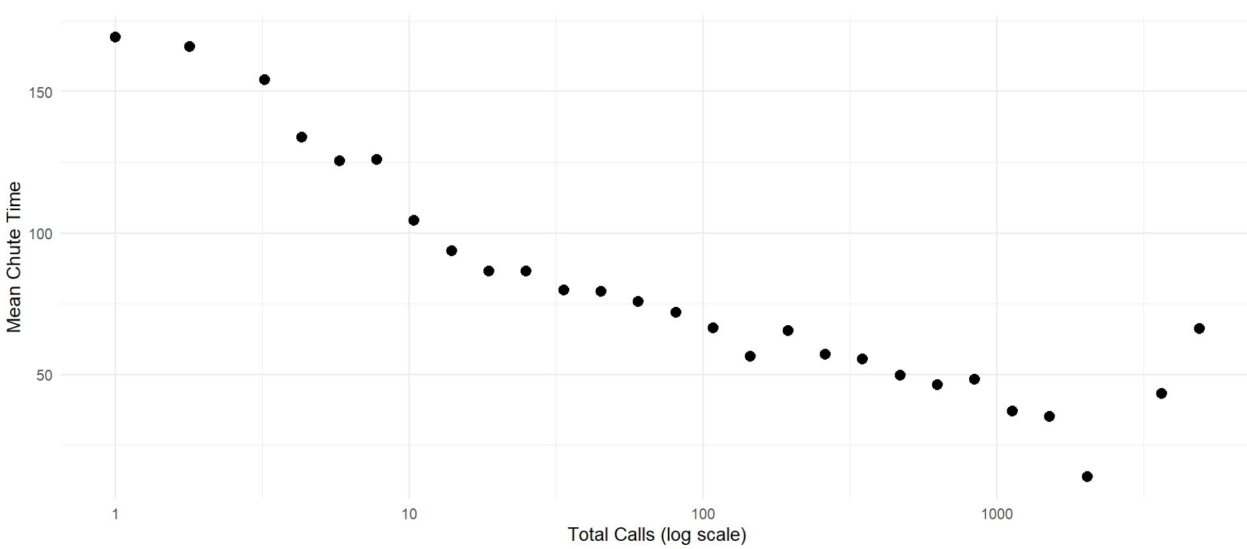

Figure 3. Binned relationship between volume and chute time.

Note: Points represent average chute time within bins of agency call volume (log scale).

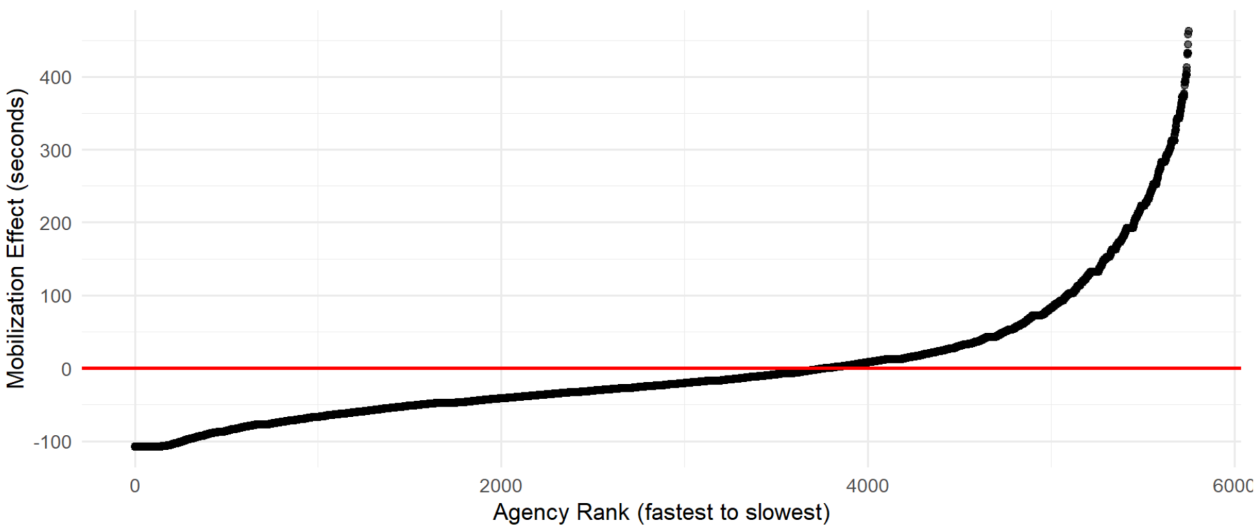

Figure 4. Agency-level fixed effects on chute time.

Note: Each point represents an agency's estimated deviation from the average mobilization time

To account for operational differences in unit status, an indicator is included for cases in which the responding unit was already in motion at the time of dispatch. An assumption was made that chute times that were exactly zero seconds in length represented in-vehicle status. This helps parse between call types in which the unit is at the station versus those already on the road. As expected, these observations are associated with substantially shorter chute times (approximately 70 seconds), reflecting a fundamentally different operational state rather than faster mobilization. Including this control improves overall model fit and helps more accurately assess other parameter estimates.

One assumes that high-volume agencies might be faster simply because they are more likely to be in the vehicle, clearing from the past call, than low volume agencies. Contrary to expectations, high-volume, high-experience agencies are actually substantially less likely to be in motion at the time of dispatch compared with everyone else, suggesting that their faster chute times are not driven by a greater reliance on immediate response, but instead are truly reflective of more efficient mobilization processes (Table 5).

Agency Experience	Unit in Motion at Dispatch = yes
Other Agencies	15.7%
Top 5% Most Experienced Agencies	8.7%

Table 5. Unit status at dispatch by agency experience level.

AGENCY-LEVEL FIXED EFFECTS

Beyond experience, agency-level fixed effects capture stable, agency-specific patterns of behavior that explain a substantial portion of the variation in the dependent variable. Approximately 27.9% of the global variation in EMS 911 chute times can be attributed to agency-level fixed effects. In a nutshell: organizations exhibit stable, characteristic patterns of behavior. Agencies operate differently and have their own distinct operational signatures, and these differences systematically explain a large proportion of response-time variation. Crucially, once agency-level fixed effects are included in the model, many variables that appeared to have substantial independent effects across multiple model specifications no longer remain significant.

This model also finds that, when first responders are dispatched to a fall, respiratory call, or "sick person," Call Type Fatigue does not set in, nor does the presence of these call types have an impact on EMS chute times (in either direction). Likewise, the model finds no robust or statistically significant evidence for the impact of age on response times, nor does the COVID period appear to affect the speed of response. Additionally, there is no measurable difference between weekend responses and responses that take place during weekday hours. Compared with the Eastern US, while the South technically posts longer chute times, the differences are negligible (approximately 3 seconds). In short, regional variation does not meaningfully explain variation once agency-level differences are accounted for.

Urban areas exhibit faster chute times (approximately 4–20 seconds, on average) compared with their suburban, rural, and wilderness counterparts. These differences are statistically significant, robust across model specifications, and persist after the inclusion of agency-level fixed effects. EMS units also mobilize more quickly during daytime hours than during overnight periods, with delays of approximately 17–23 seconds observed at night.

Some evidence suggests that responders mobilize more quickly for certain high-acuity calls. In particular, cardiac arrest calls are associated with modest but statistically significant reductions in chute time (approximately 5–7 seconds). However, most other dispatch call types do not independently predict chute time once other factors are controlled for.

LIMITATIONS

This study has several important limitations. First, while efforts were made to restrict the analysis to 911 responses by specifically excluding routine and transport and administrative calls from the dataset, it remains possible that some non-emergency, prescheduled, or administratively oriented runs remain in the dataset. As such, some heterogeneity in call context may persist despite best efforts.

Second, experience is measured at the agency level rather than the level of the responding unit, using call volume as a proxy for experience. Further disaggregation to the unit level would provide a more precise measure but is not currently feasible given the scale and computational demands of the dataset.

Third, while agency-level fixed effects capture persistent differences across organizations, they do not identify the reason why, or help understand the specific causal mechanisms underlying these differences. Further research incorporating qualitative or operational data would be necessary to better understand how organizational practices, protocols, and constraints contribute to variation in response times.

An additional limitation of this study concerns the lack of detailed dispatch context within the NEMSIS dataset. The data do not distinguish between different types of call origins or operational workflows, such as whether a call was initiated through the 911 system or represented a pre-scheduled or administratively assigned event. As a result, the dataset cannot fully capture situations in which crews may report "en route" simultaneously, particularly for pre-scheduled movements or stacked assignments. These reporting patterns may produce artificially short recorded chute times that are more reflective of documentation practices rather than true mobilization behavior.

More broadly, "en route" timestamps may reflect differences in reporting practices across agencies, including local conventions or dispatcher-specific decision rules governing when units are marked as responding. These systematic reporting differences or individual decision-rules (i.e. "I've always just coded it that way…") may introduce measurement error that is not fully captured by observed variables and may instead be absorbed within agency-level fixed effects. While a control variable was introduced to account for cases in which units are already in motion at the time of dispatch, some degree of misclassification may remain due to the absence of a directly observed measure of unit status.

POLICY IMPLICATIONS

These findings highlight an important limitation in current EMS data systems: the inability to truly distinguish between units that are already in motion at dispatch and those that initiate mobilization from a stationary position. To partially address this, chute times equal to zero were assumed to represent units already in motion at the time of dispatch. However, explicitly recording unit status (e.g., in-vehicle vs. stationary) at dispatch would substantially improve measurement precision in future research. Because these operational states produce fundamentally different "chute time" profiles, future data collection efforts should explicitly record unit status at the moment of dispatch.

A simple modification—such as adding a structured field indicating whether a unit is "in vehicle and mobile," "at station," "in hospital," or "other" at the time of dispatch—would substantially improve the interpretability of EMS response time data. Improving timestamp granularity in this way would enhance both research validity and system-level performance monitoring, enabling more precise identification of operational inefficiencies and best practices across agencies.

CONCLUSION

In conclusion, this analysis demonstrates that EMS chute times are shaped far less by individual call characteristics—such as patient age or dispatched call type—than by the expertise that accrues as a result of repeated exposure to a vast array of unique and diverse 911 calls.

These findings point to a fundamental insight: response speed reflects not only the nature of the call, but the accumulated learning, anticipatory capacity, and confidence in execution of the responders themselves. As these capabilities develop over time, they become embedded in routines, coordination, and shared expectations, producing persistent differences in performance across agencies. Improving EMS response times, therefore, requires understanding how emergency response is practiced on the ground—insights best learned from those working in the field every day.

REFERENCES

Abbas, A. Y., Odom, E. C., & Nwaise, I. (2022). Association between dispatch complaint and critical prehospital time intervals in suspected stroke 911 activations in the National Emergency Medical Services Information System, 2012–2016. *Journal of Stroke and Cerebrovascular Diseases, 31*(3), 106228. https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106228

Adeyemi, O. J., Paul, R., DiMaggio, C., Delmelle, E., & Arif, A. (2022). The association of crash response times and deaths at the crash scene: A cross-sectional analysis using the 2019 National Emergency Medical Service Information System. *Journal of Rural Health, 38*(4), 1011–1024. https://doi.org/10.1111/jrh.12666

Chen, S., Zhang, Y., Wang, Y., et al. (2024). A study of prehospital EMS response time and influencing factors in the main urban area of Chongqing, China. *International Journal for Quality in Health Care, 36*(3), mzae065. https://doi.org/10.1093/intqhc/mzae065

Do, Y. K., Foo, K., Ng, Y. Y., & Ong, M. E. H. (2013). A quantile regression analysis of ambulance response time. *Prehospital Emergency Care, 17*(2), 170–176. https://doi.org/10.3109/10903127.2012.729127

Earnest, A., Ong, M. E. H., Shahidah, N., Ng, W. M., Foo, C., & Nott, D. J. (2011). Spatial analysis of ambulance response times related to prehospital cardiac arrests in the city-state of Singapore. *Prehospital Emergency Care, 16*(2), 256–265. https://doi.org/10.3109/10903127.2011.615974

Frequent Caller National Network. (2022). *Frequent Caller National Network (FreCaNN)* – Written evidence (AES0008). https://committees.parliament.uk/writtenevidence/113035/pdf

Gehrt, T. B. (2025). Frequent callers to the emergency medical services: A mixed-methods study of call patterns, reasons for calling, and wellbeing. *Social Science & Medicine, 382*, 118326. https://doi.org/10.1016/j.socscimed.2025.118326

Helander, M. E., Formica, M. K., & Bergen-Cico, D. K. (2024). The daily patterns of emergency medical events. *Journal of Biological Rhythms, 39*(1), 79–99. https://doi.org/10.1177/07487304231193876

Hsia, R. Y., Huang, D., Mann, N. C., et al. (2018). Association between income and ambulance response time in cardiac arrest: A US national study. *JAMA Network Open, 1*(7), e185202. https://doi.org/10.1001/jamanetworkopen.2018.5202

Jones, D., Massey, H., Smith, C., Freeman, A., & Jones, E. (2025). Reducing frequent caller use of ambulance services: A review of effective strategies. *Journal of Paramedic Practice, 17*(4), 158–166. https://doi.org/10.12968/jpar.2024.0031

Medrano, N. W., Villarreal, C. L., Mann, N. C., Price, M. A., Nolte, K. B., & MacKenzie, E. J. (2022). Activation and on-scene intervals for severe trauma EMS interventions: An analysis of the NEMSIS database. *Prehospital Emergency Care, 27*(1), 46–53. https://doi.org/10.1080/10903127.2022.2053615

National Association of State EMS Officials. (2024). *EMS time interval definitions: NEMSIS data dictionary NHTSA V3.5 EMS data standard*. Data Managers Council.

Nehme, Z., Andrew, E., & Smith, K. (2016). Factors influencing the timeliness of emergency medical service response to time critical emergencies. *Prehospital Emergency Care, 20*(6), 783–791. https://doi.org/10.3109/10903127.2016.1164776

O'Keeffe, C., Nicholl, J., Turner, J., & Goodacre, S. (2011). Role of ambulance response times in the survival of patients with out-of-hospital cardiac arrest. *Emergency Medicine Journal, 28*(8), 703–706. https://doi.org/10.1136/emj.2009.086363

Pek, J., Wong, O., & Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology, 9*, 2104. https://doi.org/10.3389/fpsyg.2018.02104

Renkiewicz, G. K., Hubble, M. W., Wesley, D. R., Dorian, P. A., Losh, M. J., Swain, R., & Taylor, S. E. (2014). Probability of a shockable presenting rhythm as a function of EMS response time. *Prehospital Emergency Care, 18*(2), 224–230. https://doi.org/10.3109/10903127.2013.851308

Saver, J. L. (2006). Time is brain—quantified. *Stroke, 37*(1), 263–266. https://doi.org/10.1161/01.STR.0000196957.55928.ab

APPENDIX. Determinants of EMS Chute Time (seconds): Full Model

Variable	Model 1 – Robust Linear Model	Model 2 – Fixed Effects
Allergic Reaction	-3.33* (1.39)	-1.48 (1.75)
Animal Bite	0.07 (2.92)	-2.81 (3.70)
Assault	-1.32 (1.17)	*-0.21 (1.46)*
Back Pain	3.26** (1.14)	*1.61 (1.43)*
Burns/Explosion	2.50 (3.27)	*-5.20 (4.16)*
Cardiac Arrest	-4.06*** (1.14)	*-6.79*** (1.43)*
Chest Pain	-0.97 (0.70)	*-1.61 (0.88)*
Choking	-1.23 (2.28)	*-3.57 (2.85)*
Diabetic	1.08 (1.09)	*-1.17 (1.37)*
Falls	0.79 (0.67)	*-1.06 (0.84)*
Fire	0.60 (3.38)	*6.09 (4.32)*
Headache	2.86 (1.54)	*2.32 (1.93)*
Heart Problems	3.43** (1.05)	*-0.89 (1.33)*
Heat/Cold Exposure	-5.58* (2.83)	*-9.85** (3.55)*
Hemorrhage/Laceration	-0.46 (0.93)	*-0.41 (1.16)*
Medical Alarm	4.40* (1.88)	*-0.65 (2.41)*
OBGYN	-1.41 (1.47)	*0.41 (1.82)*
Overdose	-2.09* (0.94)	*-2.81* (1.17)*
Pandemic	-7.35*** (1.99)	*0.70 (2.50)*
Psych	-3.82*** (0.82)	*2.71** (1.02)*
Respiratory	0.63 (0.66)	*-1.67* (0.83)*
Seizure	-2.27** (0.81)	*-3.95*** (1.02)*
Sick Person	0.85 (0.62)	*0.66 (0.78)*
Stabbing/GSW	-8.58*** (1.71)	*-0.40 (2.14)*
Standby	3.64 (2.53)	*-2.29 (3.22)*
Stroke	-0.74 (0.90)	*-4.01*** (1.13)*
Traffic Incident	0.34 (0.74)	*-2.61** (0.93)*
Traumatic Injury	-2.37** (0.88)	*-1.70 (1.11)*
Unclear	-0.99 (0.72)	*1.00 (0.96)*
Unconscious	-1.32 (0.74)	*-3.34*** (0.93)*
Unknown/Person Down	-2.36** (0.84)	*-1.97 (1.07)*
Well Check	-0.39 (2.27)	*-5.52 (2.92)*
Patient Age	0.032*** (0.005)	*0.008 (0.007)*
COVID period	-1.12*** (0.23)	*-0.27 (0.31)*
Night response	17.87*** (0.29)	*23.22*** (0.36)*
Weekend response	0.44 (0.25)	*0.03 (0.31)*
West region	-16.25*** (0.35)	*-2.62 (1.52)*
South region	3.61*** (0.30)	*-3.73** (1.38)*
Western Plains	-1.93*** (0.48)	—
Great Lakes	-5.13*** (0.40)	—
Suburban area	10.57*** (0.50)	*4.83** (1.85)*
Rural area	14.77*** (0.48)	*13.14*** (1.69)*
Wilderness area	19.99*** (0.96)	*17.30*** (3.46)*
Agency experience	-97.68*** (0.76)	—
In Vehicle When Dispatched	-70.38***	*-86.37****
AIC	2,126,573	—
BIC	2,127,039	—
Agency fixed effects	No	*Yes (5,518 agencies)*
Variance Explained by Fixed Effects	—	*27.9%*
Observations	185,015	*183,592*
RMSE	45.58	*57.10*
R^2 / Adj. R^2	—	*0.432*

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Variance inflation factors (VIFs) were examined. The maximum VIF was 1.31, indicating no evidence of multicollinearity. Reference categories were set to avoid multicollinearity: “East” was the reference category set for “region”. “Urban” was the reference category for “urbanicity”, and “Abdominal Pain/Problems” was the reference category selected for “Dispatched Call Type.”


EMS|MC
EMS|MC