TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
2024, Vol. 2678(11) 1378–1394

DOI: 10.1177/03611981241243329
journals.sagepub.com/home/trr
Sage

Examining the Relationship between Connected Vehicle Driving Event Data and Police-Reported Traffic Crash Data at the Segment- and Event Level

Nischal Gupta[1], Hisham Jashami[1], Peter T. Savolainen[1], Timothy J. Gates[1], Timothy Barrette[2], and Wesley Powell[2]

Abstract

Police-reported crash data have been the de facto element used by the transportation agencies in developing and implementing traffic safety projects. This approach is reactive in nature and can lead to suboptimal investment decisions owing to inherent challenges in crash data analysis. Because of their large-scale and near real-time availability, connected vehicle (CV) driving event data have emerged as a promising means of addressing these challenges. This study utilized CV event data for three different event types, namely, acceleration, braking, and cornering at three severity levels (easy, normal, and harsh), to examine the viability of using these data in traffic safety analysis. The results showed a strong correlation between crash frequency and CV driving event frequency. CV event data also improved the goodness-of-fit of crash frequency models. The results also showed that the relationship between CV driving events and traffic volume and roadway geometric data were generally consistent with the trends that crash data usually exhibit with the same predictors. This was true at both segment level and individual event level, as well as when the data were examined across different event/crash types. Overall, the results showed a strong case for these data to be used in traffic safety analyses as a complement to, or in lieu of, crash data.

Keywords

data and data science, connected vehicle data applications, safety, surrogate safety measures

More than 36,000 fatalities and 5 million injuries are sustained in motor vehicle crashes in the United States every year (*1*). For every crash-related fatality, eight people are hospitalized, and 100 are treated and released from hospitals (*2*). Additionally, the United States incurs a huge economic and societal cost from traffic crashes, equivalent to approximately 1.6% of the gross domestic product (*3*). In response to this public health dilemma, advances in vehicle safety features, improved roadway design, and the introduction of various policies and programs to address driver behavioral issues have significantly reduced crashes, injuries, and fatalities over the years. However, these metrics have generally plateaued in recent years, providing motivation for further efforts to address this public health and economic issue. In 2020, despite a decrease in vehicle miles traveled (VMT) because of the COVID-19 pandemic, vehicle-related deaths were up by 8% in the United States (*1*).

Transportation agencies invest considerable resources to reduce both the frequency of crashes, as well as the degree of injury sustained by crash victims. A diverse range of highway safety stakeholders have adopted the national strategy of "Towards Zero Deaths," which was initiated by the Federal Highway Administration in 2009. These same stakeholders have developed strategic highway safety plans that outline comprehensive frameworks to help reduce traffic crashes and fatalities on public roads. These plans provide guidance as to the identification of emphasis areas where crash risks are

[1]Department of Civil and Environmental Engineering, Michigan State University, East Lansing, MI
[2]Global Data Insights & Analytics, Ford Motor Company, Dearborn, MI

Corresponding Author:
Peter T. Savolainen, pete@msu.edu

most pronounced, as well as specific strategies that present the greatest potential for near- and long-term improvements in traffic safety.

Historically, the most critical element of these data systems is police-reported crash data. In consideration of resource constraints, it is imperative that agencies are able to proactively identify crash countermeasures and candidate locations that present the greatest opportunities for improvement. To this end, the *Highway Safety Manual* (*4*) outlines best practices for data-driven and proactive methods of safety management. These practices are based on the availability of high-quality, properly maintained, and regularly updated police-reported crash data. These data records are compiled by law enforcement agencies and describe the location, circumstances, persons, and vehicles involved in the crashes. However, reliance on crash data presents several challenges from issues such as underreporting of crashes, differences in minimum crash reporting requirements across states, the relative infrequency of crashes at individual locations on an annual basis, and the time lag introduced in making proactive decisions owing to crash data being available generally at year-end (*5*). Collectively, these issues impede the ability of transportation agencies to quickly and proactively respond to emerging traffic safety issues. This is particularly true during periods of extraordinary events that significantly affect travel patterns on a systemwide level, such as the large-scale travel restrictions induced in response to the COVID-19 pandemic in 2020 (*6*). These changes in travel behavior, in turn, affected traffic speeds and safety trends across the transportation network.

To this end, various surrogate measures of road safety have emerged in recent years that have shown promise in overcoming some of the limitations of crash data. These surrogate measures include traffic conflicts and various other types of near-crash events. The advantage of these metrics is that they tend to occur more frequently than crashes, allowing for safety issues to be identified more quickly as compared to reliance on police-reported crash data. Much of the early work in this area focused on facility-level observations, such as monitoring individual road locations through field observation or the use of cameras. Alternately, the observation of traffic over time and space provides a means of network-level analysis. Recent examples include the second Strategic Highway Research Program (i.e., SHRP 2) Naturalistic Driving Study, which included voluntary participation from 3,400 drivers using a series of cameras and sensors installed on the vehicles of study participants (*7*). Although more efficient, these methods also tend to be resource-intensive and are difficult to implement at scale.

In contrast, connected vehicle (CV) driving event data have emerged as a promising surrogate safety measure that allows the leveraging of data using equipment already installed in the vehicles on the road today. These data can provide real-time information about vehicle performance and underlying aspects of driver behavior, including travel speeds, engine status, and the use of various vehicle systems. These data present a more objective lens than relying on subjective assessment of a crash scene. Moreover, CV event data are more frequently updated, providing significant advantages as compared to police-reported crash data for analysis purposes.

However, the use of such data for proactive safety analyses is still largely in its nascent stages, and the research is limited. Nevertheless, some studies have identified risky or aggressive driving behaviors based on the magnitude of acceleration, braking, and steering employed by the driver (*8–12*). These risky driving behaviors, in turn, have been found to be positively correlated with the likelihood of a crash or near-crash (*13–18*). Braking behavior has also been found to be significantly associated with traffic safety. Generally, drivers who were involved in crashes tended to brake more frequently and severely (*19*). Also, fatigued or distracted drivers tended to show more aggressive braking behavior (*20*, *21*).

The studies in the existing literature primarily focus on correlating driver behavior with crash risk. However, to use these driving event data as an alternative to crash data or as a supplement to the crash data, it is important to demonstrate that these event data behave similarly to crash data at an aggregate level as well as at an individual event level. For example, a study conducted using truck braking events at roundabouts found that harsh braking events tended to be influenced by traffic and geometric parameters similar to crashes (*22*). Similarly, a strong positive correlation was also found between hard braking events and crashes in work zones (*23*).

To that end, the present study utilizes CV driving data to demonstrate the usefulness of using such data in traffic safety analyses. The objective is to show that the CV driving event data follow similar relationships as crashes with respect to traffic volume, roadway geometric, and other safety-related parameters at both aggregated- and individual event levels. Count models and severity models are developed for both crashes and CV driving event data using the same predictors and the results are compared.

Data Preparation and Data Summary

The data used for analyses in the present study were drawn from the roadway network in the Southeast Michigan Council of Governments (SEMCOG) region of Michigan. SEMCOG is a regional planning partnership made up of seven counties in the Detroit metropolitan area, namely Livingston, Macomb, Monroe, Oakland, St. Clair, Washtenaw, and Wayne. The following subsections detail the datasets used in the present study and the

manner in which these sources were integrated to create the analysis dataset.

Ford Journeys Data

The CV driving event data were obtained from Ford Motor Company. Ford collects CV event data from Ford vehicles equipped with its FordPass mobile application. This includes information about vehicle kinematic performance, including spatial and temporal information about three major driving event categories: acceleration, braking, and cornering. This dataset is referred to as FordPass Journeys data. The relative magnitude of the acceleration or the angular velocity recorded for each individual event is used to further divide the three driving events into three severity categories: easy, normal, or harsh. Table 1 provides the thresholds of the three driving events that define these three categories for each event type.

The data for the frequency and severity of these events were provided by Ford in an aggregated and deidentified format for calendar year 2020 for the SEMCOG region. Figure 1 shows the distribution of Ford CV driving events for the entire calendar year of 2020 by type and severity of events. After removing events where data were missing or incomplete, more than 13 million events were included across the SEMCOG region for the analysis period.

Roadway and Crash Data

The crash data for calendar year 2020 were obtained from Michigan State Police for the entire SEMCOG region. The data, obtained at individual crash level, included details about crash severity on the KABCO scale (further details in the section summarizing the data) and characteristics of the vehicles and drivers involved in the crashes. These crashes were aggregated at segment level and integrated with a detailed roadway information database maintained by the Michigan Department of Transportation. This database included information about annual average daily traffic (AADT), national functional class of roadways, roadway geometric characteristics including number and width of lanes, type and width of medians, shoulder width, and the presence of features such as signals, turn lanes, passing lanes. Information on the contextual classification of these roadways was not available. However, it should be noted that the traffic volume and speed limit information, which are considered in the analyses, are generally correlated with roadway context and tend to capture its effects to some degree.

Table 1. Event Thresholds for FordPass Journeys Data

Event type	Event thresholds		
	Easy, ft/s^2	Normal, ft/s^2	Harsh, ft/s^2
Acceleration	3.9	5.6	7.2
Braking	5.9	8.2	10.2
Cornering	3.0	5.2	7.2

Data Summary

The number of crashes and CV driving events occurring on each segment were calculated and integrated with roadway information data to create a segment-level dataset. Initial investigation considered three roadway facility types: limited-access freeways, multilane highways, and two-lane roads. However, the frequency of occurrence of CV driving events on limited-access freeways was relatively small. The CV driving events occurred at a rate of 366.99 and 2,149.19 events per million VMT on multilane and two-lane roads. In comparison, the rate of occurrence on limited-access facilities was only 33.29 per million VMT. Thus, only multilane roads and two-lane roads were considered for analysis, as limited-access facilities did not show any meaningful, significant relationships. Table 2 presents the descriptive statistics of variables used in the segment-level analysis.

Figure 2 shows the relationship between crash occurrence and CV driving event occurrence during the same time period (2020) separately for multilane and two-lane roads. These plots show very strong linear and positive relationships between crashes and the CV driving data, as indicated by R^2. Similar trends were obtained when crash counts were plotted against CV driving events by type and severity. However, the goodness-of-fit was better for both acceleration and braking events compared with cornering events. When broken down based on their respective types, the relationship between certain crash types and CV driving events improved further. For example, rear-end crashes tend to occur when drivers fail to brake at appropriate time. It was therefore expected that braking events would occur much more frequently at locations where rear-end crashes occur.

Figure 3 shows the relationship between rear-end crashes and braking events (including easy, normal, and harsh braking events). Similar relationships were also found between rear-end crashes and acceleration events. Angle crashes showed a positive but weak relationship with braking events (R^2 of 0.44). Cornering events did not show as strong a relationship with various crash types. Collectively, the results showed significant potential for using CV driving event data as a supplement or proxy to police-reported crash data.

The relationship between crashes and CV driving events was also investigated spatially at county level. To

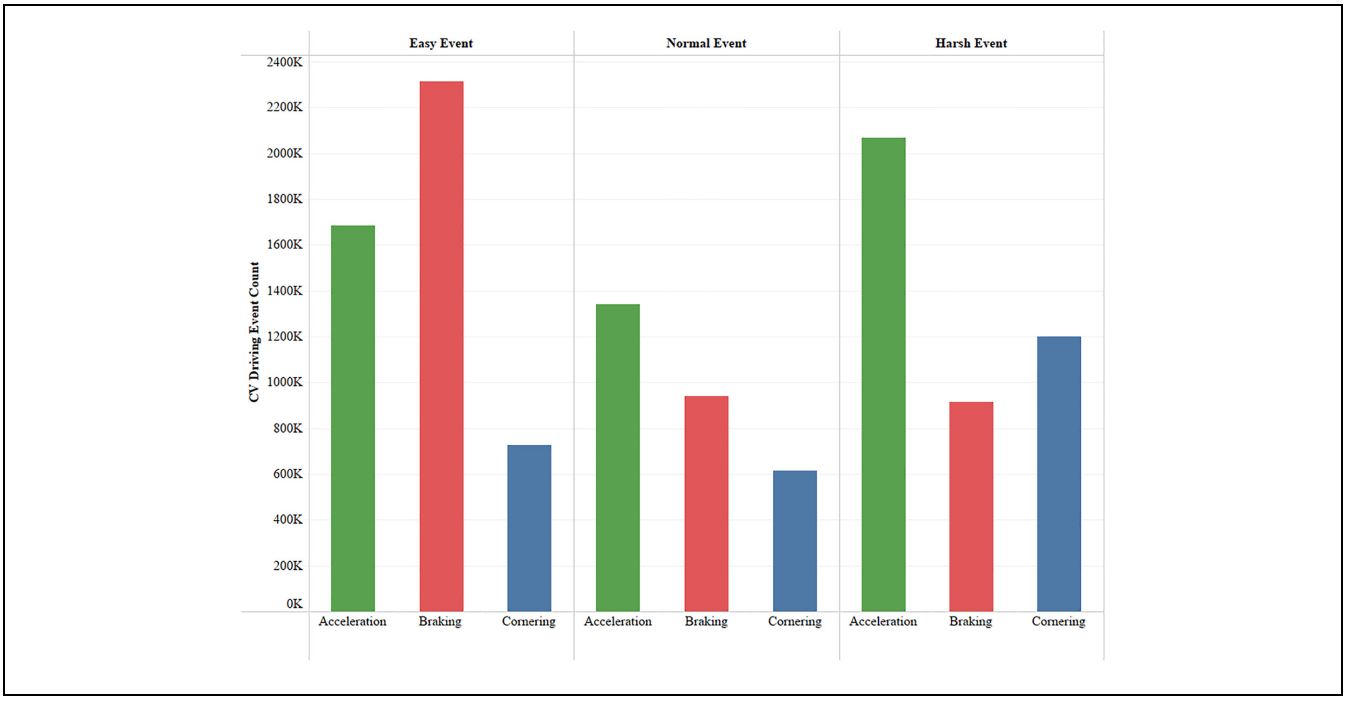

Figure 1. Distribution of Ford CV driving events by event type and severity.
Note: CV = connected vehicle.

Table 2. Descriptive Statistics of Pertinent Variables for Segment-Level Dataset

Parameter	Multilane roads		Two-lane roads	
	Mean	SD	Mean	SD
Crash count (2020)	24.3	26.3	7.97	9.38
CV driving event count (2020)	2,522.0	2,558	686.10	926.70
Segment length (mi)	0.99	0.79	1.28	1.22
AADT (vpd)	21,964.82	10,628.05	9,402.53	6,903.65
Undivided road (1 if yes, 0 otherwise)	0.42	0.49	na	na
Ditch-type median (1 if yes, 0 otherwise)	0.14	0.35	na	na
Other type of median (1 if yes, 0 otherwise)	0.44	0.50	na	na
Speed limit 55 mph or more (1 if yes, 0 otherwise)	0.16	0.37	0.44	0.50
Speed limit 40 to 50 mph (1 if yes, 0 otherwise)	0.62	0.49	0.35	0.48
Speed limit 35 mph or less (1 if yes, 0 otherwise)	0.22	0.42	0.21	0.42
Signal present (1 if yes, 0 otherwise)	0.93	0.26	0.40	0.49
Sample size	619 segments, 610 mi		214 segments, 273 mi	

Note: CV = connected vehicle; AADT = annual average daily traffic; SD = standard deviation; vpd = vehicles per day; na = not applicable.

that end, Figure 4 shows the distribution of police-reported crashes and CV driving events based on the SEMCOG counties. The figure shows that the spatial distribution of the two types of data are similar. For each of the seven SEMCOG counties separately, Figure 5 additionally shows the relationship between total crash counts and total CV driving event counts for the calendar year 2020. The figure shows strong and consistently positive relationships between police-reported crashes and CV driving event data when the data were disaggregated spatially. The figure also shows that this relationship was stronger for Macomb, Oakland, St. Clair, and Washtenaw. Monroe and Livingston counties demonstrated relatively weaker relationships between crashes and CV driving events as indicated by the slope of the trend line.

When the CV driving event data were further disaggregated, they still showed trends similar to the crash data. Figure 6 provides a comparison of the degree to which total CV driving events (i.e., of all types and

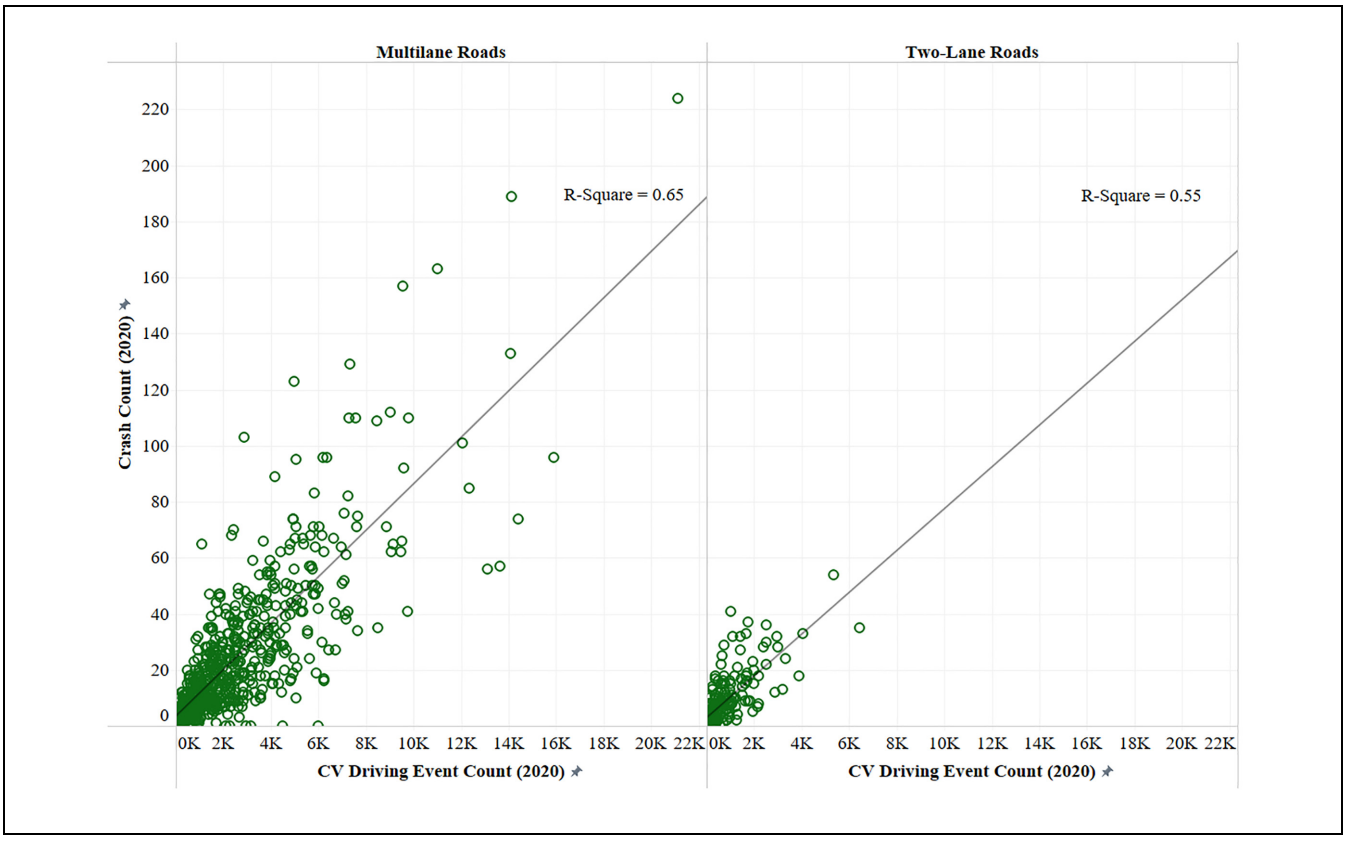

Figure 2. Plot of total crash count versus CV driving event count by road type.
Note: CV = connected vehicle.

severity levels) compared with crashes are affected at the individual road segment level by time of day. In this case, the data were drawn from four roadway corridors that include a series of signalized intersections. These included sections of Gratiot Avenue, Groesbeck Highway, M-53, and M-59. Figure 6 illustrates trends in crash and CV driving event frequency by the time of day on all four routes combined. This figure clearly shows similar patterns in relation to the relative frequency of both crashes and CV driving events. Both safety indicators were shown to peak around the same time. The morning peak occurred between 8 and 9 a.m., whereas the evening peak occurred between 4 and 5 p.m. It is interesting to note that CV driving events occurred at a ratio of between 50 to 1 and 100 to 1. This illustrates one of the primary advantages of such data in that CV driving events occur significantly more frequently and, as such, they may allow for more efficient identification of safety issues as compared to crash data.

Crash-level and event-level datasets were also prepared for severity analysis. There are five severity levels for crashes on the KABCO scale, where K represents fatal crashes, A denotes serious injuries (e.g., severe lacerations, burns, broken bones), B denotes minor injuries (i.e., any injury evident less than K or A), C denotes possible injuries (no evident external injury, but potential injury noted by crash-involved victim), and O denotes property damage only (PDO) crashes (i.e., no injury). Because of the lower occurrence of K and A crashes, the two categories were combined. Similarly, B and C crashes were combined. Thus, the final data had three levels of crash severity. The three severity levels were coded in order of their increasing severity from 1 (PDO crash) to 3 (K or A crash). Since crashes that involve pedestrians, bicyclists, motorcycles, or animals tend to skew severity distributions, they were excluded from the severity analysis. In total, 14,088 crashes and 1,562,786 CV driving events were analyzed on multilane roads. On two-lane roads, 1,112 crashes and 146,825 CV driving events were included in the severity analysis.

As stated earlier, the CV driving data had events categorized into three levels of severity (easy, normal, and harsh). For severity analysis purposes, the three severities were also coded in order of their increasing severity: 1 for easy event, 2 for normal event, and 3 for harsh event. Table 3 presents the descriptive statistics of the crash- and event-level datasets. The crash-level information variables such as weather and lighting conditions, and

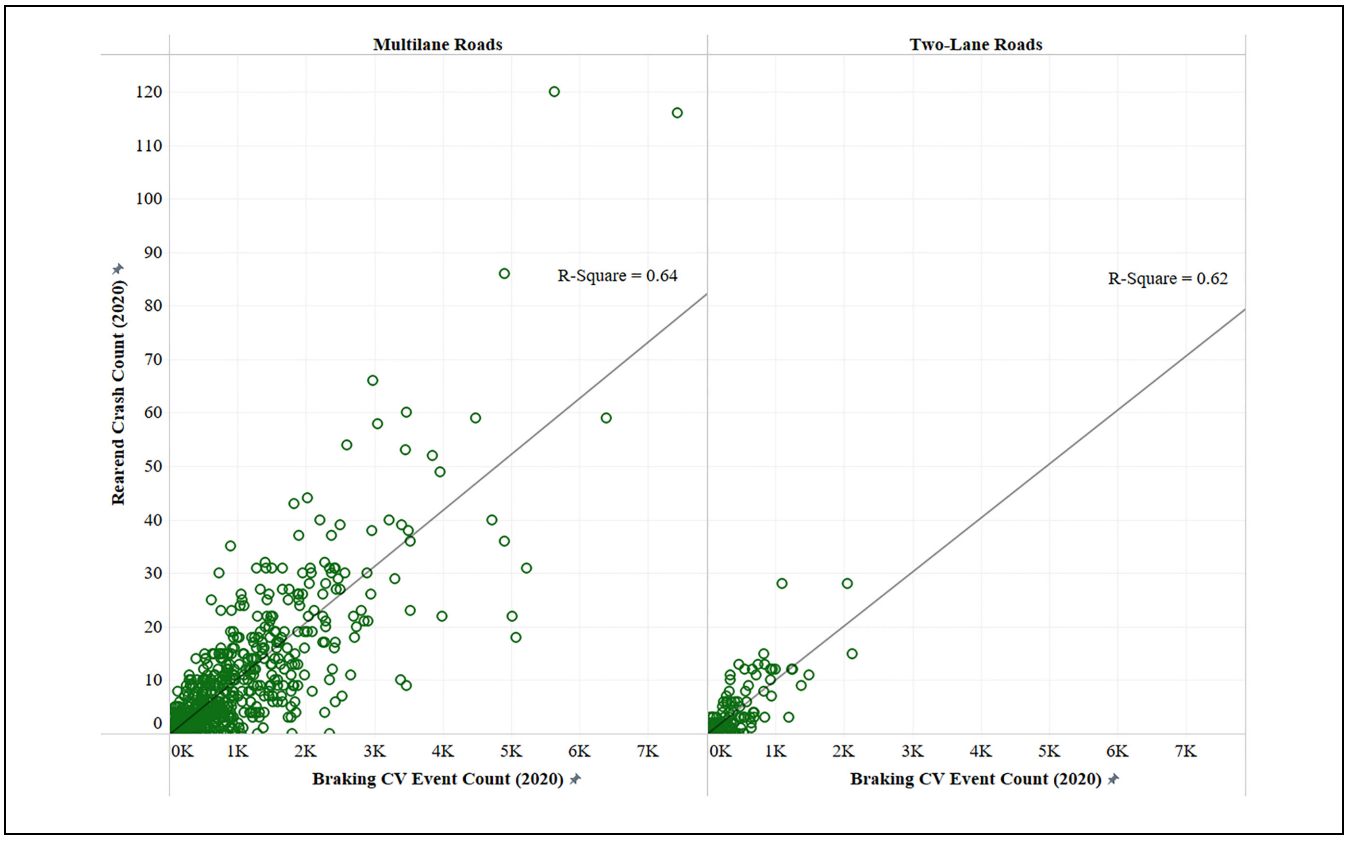

Figure 3. Plot of rear-end crashes versus braking event count by road type.

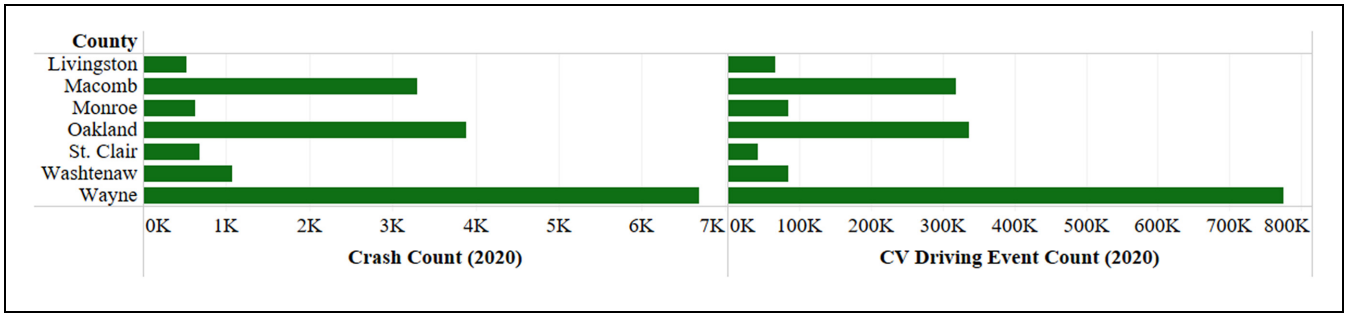

Figure 4. Distribution of crashes and CV driving events by county.
Note: CV = connected vehicle.

driver sobriety were not available for the CV driving event dataset and thus are marked as "NA" in Table 3. One important point that warrants discussion here is the frequency of crashes and CV driving events during the months of March to May 2020. These were the months when the effects of the COVID-19 pandemic were greatest on the transportation system nationwide. In the state of Michigan, stay-at home orders were issued on March 23, 2022 and stayed in effect until April 30 of that year (*24*). Significantly fewer crashes and CV driving events occurred during this period owing to lower traffic volumes. This is evident from the descriptive statistics shown in Table 3.

Network Screening Applications

Typically, transportation agencies and practitioners utilize police-reported crash data to identify and rank high-risk locations to create a prioritization scheme for implementation of countermeasures. To that end, the simple ranking method based on raw crash counts, or crashes normalized by segment length or VMT is the simplest

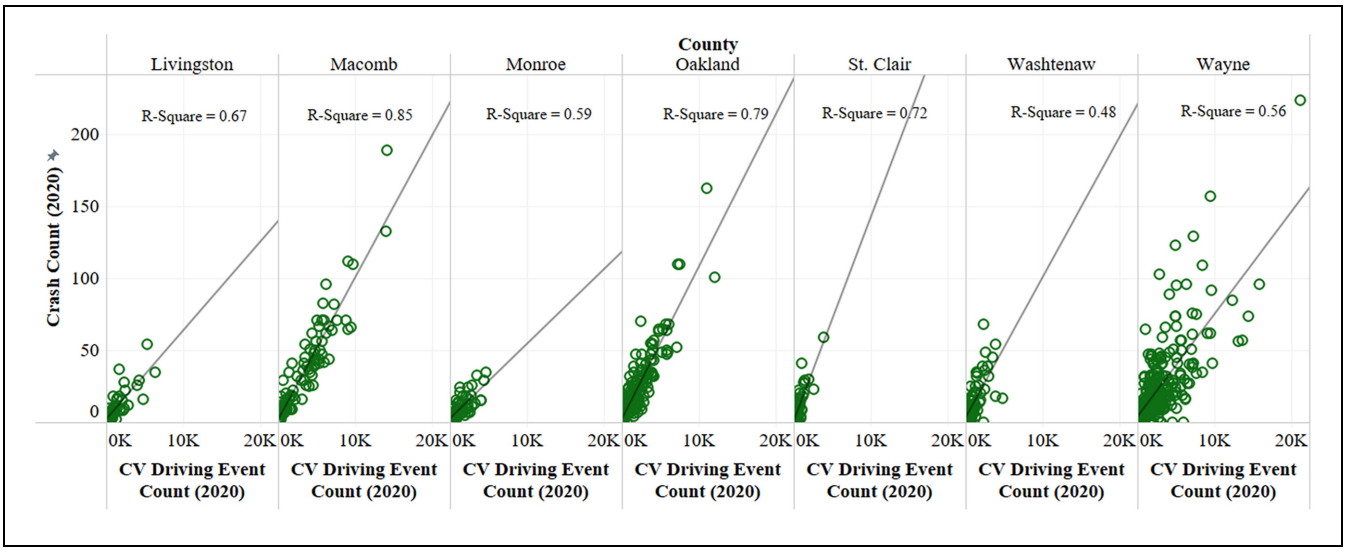

Figure 5. Plots of total crash count versus CV driving event count by county.
Note: CV = connected vehicle.

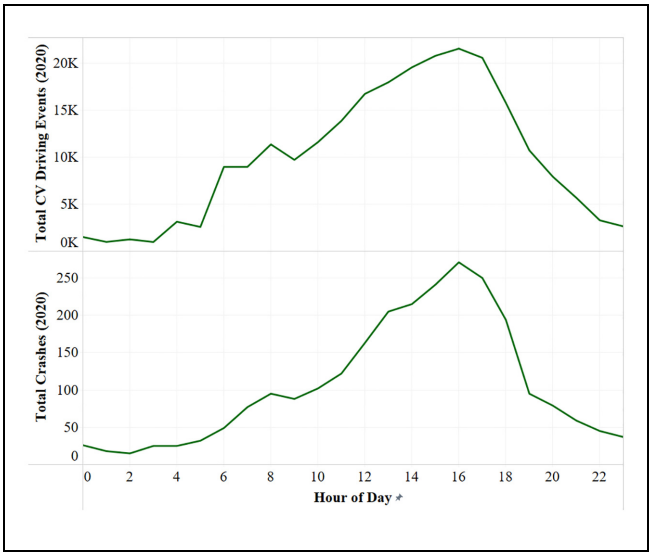

Figure 6. Frequency of crashes and CV driving events by time of day on signalized corridors.
Note: CV = connected vehicle.

method listed in the *Highway Safety Manual* (*25*). A network screening exercise for the state-maintained multilane and two-lane roadway networks in the SEMCOG region was conducted using both the crash data and the CV driving event data. Two separate ranked lists were created, one based on crash data and the other based on CV driving event data. Sites of very short segment length (<0.1 mi) or very low traffic volume (<1,000 vpd) were excluded from this analysis, as their inclusion may have resulted in overestimations of crash rates on normalization. The top 10% of the total number of sites in each list were considered to be high-risk locations and were thus filtered and compared. Table 4 presents the results of this exercise and shows the number of segments that were common to both lists. Ranking was based on three metrics: raw counts, counts normalized by segment length, and counts normalized by VMT.

The table shows promising results from the network screening perspective since both the crash data and the CV driving event data identified similar sites as high-risk locations. Of the top 10% sites (n = 62) identified by the crash data as high-risk locations on the multilane roads, 39 sites (63%) were also identified as high-risk locations based on the CV driving event data when the ranking of sites was based on raw counts. When normalized by segment length and AADT, 58% of the sites were still common to both the ranked lists. Similar results were obtained for two-lane roads where majority of the high-risk sites were common to both lists prepared based on crash data and CV driving event data regardless of the normalization scheme. These results provide a strong case for the use of CV data in network screening applications, particularly in cases in which the availability of crash data is limited.

Statistical Analysis

To gain further insights into the nature of the relationships between crashes and CV driving data, a series of regression models was developed. As stated earlier, frequency models were estimated to assess the degree to which CV driving events were predictive of traffic crashes. Since both crash frequency and CV driving event frequency on any given road segment take the

Table 3. Descriptive Statistics of Pertinent Variables for Severity Analysis

Parameter	Event-level dataset		Crash-level dataset	
	Mean	SD	Mean	SD
Multilane roads				
Time period of event				
January 1 to March 15, 2020 (1 if yes, 0 otherwise)	0.26	0.44	0.26	0.44
March 16 to May 31, 2020 (1 if yes, 0 otherwise)	0.15	0.36	0.10	0.31
June 1 to September 30, 2020 (1 if yes, 0 otherwise)	0.34	0.47	0.37	0.48
October 1 to December 31, 2020 (1 if yes, 0 otherwise)	0.25	0.43	0.27	0.44
Speed limit				
Speed limit 55 mph or more (1 if yes, 0 otherwise)	0.14	0.34	0.07	0.25
Speed limit 40 to 50 mph (1 if yes, 0 otherwise)	0.77	0.42	0.75	0.43
Speed limit 35 mph or less (1 if yes, 0 otherwise)	0.09	0.29	0.18	0.38
Road Type				
Undivided road (1 if yes, 0 otherwise)	0.41	0.49	0.45	0.50
Ditch-type median (1 if yes, 0 otherwise)	0.09	0.28	0.06	0.24
Raised island with curb or other type of median (1 if yes, 0 otherwise)	0.50	0.50	0.49	0.50
Time of day				
Midnight to 3 a.m. (1 if yes, 0 otherwise)	0.01	0.11	0.04	0.19
3 to 6 a.m. (1 if yes, 0 otherwise)	0.02	0.15	0.03	0.16
6 to 9 a.m. (1 if yes, 0 otherwise)	0.11	0.31	0.09	0.28
9 a.m. to midday (1 if yes, 0 otherwise)	0.15	0.36	0.13	0.33
Midday to 3 p.m. (1 if yes, 0 otherwise)	0.22	0.42	0.22	0.41
3 to 6 p.m. (1 if yes, 0 otherwise)	0.27	0.44	0.29	0.45
6 to 9 p.m. (1 if yes, 0 otherwise)	0.16	0.36	0.15	0.35
9 p.m. to midnight (1 if yes, 0 otherwise)	0.05	0.22	0.07	0.25
Crash-level information				
Driver intoxicated (1 if yes, 0 otherwise)	NA	NA	0.03	0.23
Adverse weather (1 if yes, 0 otherwise)	NA	NA	0.13	0.34
Truck or bus involved (1 if yes, 0 otherwise)	NA	NA	0.05	0.23
Daylight condition (1 if yes, 0 otherwise)	NA	NA	0.77	0.50
Dark lighted condition (1 if yes, 0 otherwise)	NA	NA	0.19	0.40
Dark unlighted condition (1 if yes, 0 otherwise)	NA	NA	0.04	0.20
Sample size	1,562,786		13,907	
Two-lane roads				
Time period of event				
January 1 to March 15, 2020 (1 if yes, 0 otherwise)	0.23	0.42	0.29	0.46
March 16 to May 31, 2020 (1 if yes, 0 otherwise)	0.15	0.36	0.10	0.30
June 1 to September 30, 2020 (1 if yes, 0 otherwise)	0.38	0.49	0.35	0.48
October 1 to December 31, 2020 (1 if yes, 0 otherwise)	0.23	0.42	0.26	0.44
Speed Limit				
Speed limit 55 mph or more (1 if yes, 0 otherwise)	0.40	0.46	0.51	0.50
Speed limit 40 to 50 mph (1 if yes, 0 otherwise)	0.43	0.49	0.31	0.46
Speed limit 35 mph or less (1 if yes, 0 otherwise)	0.18	0.38	0.18	0.39
Turn lanes				
No turn lane (1 if yes, 0 otherwise)	0.08	0.27	0.08	0.27
Only left-turn lane present (1 if yes, 0 otherwise)	0.18	0.39	0.19	0.39
Only right-turn lane present (1 if yes, 0 otherwise)	0.06	0.24	0.05	0.23
Both left- and right-turn lanes present (1 if yes, 0 otherwise)	0.68	0.47	0.68	0.47
Time of day				
Midnight to 3 a.m. (1 if yes, 0 otherwise)	0.01	0.09	0.03	0.18
3 to 6 a.m. (1 if yes, 0 otherwise)	0.03	0.18	0.02	0.13
6 to 9 a.m. (1 if yes, 0 otherwise)	0.11	0.31	0.13	0.33
9 am to midday (1 if yes, 0 otherwise)	0.15	0.36	0.12	0.33
Midday to 3 p.m. (1 if yes, 0 otherwise)	0.23	0.42	0.22	0.42
3 to 6 p.m. (1 if yes, 0 otherwise)	0.27	0.45	0.27	0.45
6 to 9 p.m. (1 if yes, 0 otherwise)	0.15	0.35	0.15	0.35
9 pm to midnight (1 if yes, 0 otherwise)	0.05	0.21	0.06	0.24
Crash-level information				
Driver intoxicated (1 if yes, 0 otherwise)	NA	NA	0.05	0.22
Adverse weather (1 if yes, 0 otherwise)	NA	NA	0.16	0.37

(continued)

Table 3. (continued)

Parameter	Event-level dataset		Crash-level dataset	
	Mean	SD	Mean	SD
Truck or bus involved (1 if yes, 0 otherwise)	NA	NA	0.05	0.22
Daylight condition (1 if yes, 0 otherwise)	NA	NA	0.77	0.50
Dark lighted condition (1 if yes, 0 otherwise)	NA	NA	0.09	0.28
Dark unlighted condition (1 if yes, 0 otherwise)	NA	NA	0.14	0.35
Sample size	146,825		1,097	

Note: NA = not available; SD = standard deviation.

Table 4. Network Screening Results

Performance metric	Number of high-risk sites common in lists based on crash data and CV driving event data	
	Multilane roads	Two-lane roads
Raw counts	39 (63%)	11 (58%)
Counts normalized by segment length (per mile)	32 (52%)	10 (53%)
Counts normalized by VMT (per MVMT)	36 (58%)	13 (68%)
Total number of sites	614	184
Number of high-risk sites (10% of total)	62	19

Note: CV = connected vehicle; VMT = vehicle miles traveled; MVMT = million vehicle miles traveled.

form of discrete and nonnegative integers, count data models such as Poisson and negative binomial models were appropriate choice for modeling. In the present study, negative binomial models were selected to analyze the data since they account for the overdispersion generally found in crash data, which occurs when the variance of the expected crash counts is greater than its conditional mean. The probability of the number of crashes, y, occurring on a road segment i, during a specific time period is given in Equation 1,

$$P(y_i) = \frac{e^{-\lambda}\lambda_i^{y_i}}{y_i!} \tag{1}$$

where λ_i is the average number of crashes for segment i with similar characteristics. In negative binomial modeling frameworks, parameter λ_i can be calculated as shown in Equation 2.

$$\lambda_i = EXP(\beta X_i + \varepsilon_i) \tag{2}$$

where
X_i is vector of explanatory variables;
β is vector of parameters to be estimated that quantify the effect of these variables; and
$EXP(\varepsilon_i)$ is gamma-distributed with a mean and variance equal to 1 and α, respectively, where α is overdispersion parameter.

The random effects framework was not considered in the analysis since data from only 1 year were analyzed, thereby eliminating any potential correlation among crash counts or CV driving event counts over time.

In addition to frequency models, severity models were also estimated separately for both crashes and CV driving events using ordinal logistic regression. Ordinal models are derived by defining an unobserved variable as a linear function for each observation, as shown in Equation 3,

$$z = \beta X + \varepsilon \tag{3}$$

where
X is vector of variables determining the discrete ordering for observation n,
β is vector of estimable parameters, and
ε is random disturbance (*26*).
The observed ordinal data, y, for each observation is defined as follows:

$$\begin{aligned} &y = 1 \text{ if } z \leqslant \mu_0 \\ &y = 2 \text{ if } \mu_0 \leqslant z \leqslant \mu_1 \\ &y = \ldots \\ &y = I \text{ if } z \geqslant \mu_{I-1} \end{aligned} \tag{4}$$

where μ are thresholds that define y, which corresponds to integer ordering, with I being the highest integer ordered

Table 5. Comparison of Negative Binomial Parameter Estimates Between Crashes and CV Driving Events on Multilane Roads

Parameters	Response variable = total crash count		Response variable = total CV driving events	
	Estimate (SE)	*p*-value	Estimate (SE)	*p*-value
Intercept	−6.24 (0.51)	<0.001	−1.22 (0.49)	<0.001
Ln(CV driving event count)	0.10 (0.03)	<0.001	na	na
Ln(AADT)	0.79 (0.06)	<0.001	0.82 (0.05)	<0.001
Speed limit (mph)				
55 or more	Base condition		Base condition	
40 to 50	0.39 (0.11)	<0.001	0.24 (0.11)	<0.001
35 or less	0.73 (0.12)	<0.001	0.29 (0.12)	<0.001
Signal present (1 if yes, 0 otherwise)	0.61 (0.13)	<0.001	0.88 (0.10)	<0.001
Median type				
Undivided	Base condition		Base condition	
Ditch-type median	−0.35 (0.12)	0.003	−0.25 (0.12)	0.031
Other type of median	−0.18 (0.05)	<0.001	−0.02 (0.06)	0.680
Goodness-of-fit (−2×log-likelihood)				
With CV driving event data	4,486.60		10,297	
Without CV driving event data	4,496.86		na	
Likelihood ratio test for Ln(CV driving event count)				
Chi-square on 1 degree of freedom	10.26	0.001	na	

Note: Ln(CV driving event count) = natural logarithm of CV driving event count; Ln(AADT) = natural logarithm of AADT; CV = connected vehicle; SE = standard error; na = not applicable.

response. μ and β are estimated jointly, which reduces the estimation problem to determining the probability of *I*-specific ordered responses for each observation, *n*. If ε is assumed to be logistically distributed across observations, an ordered logit model ensues, with ordered selection probabilities as in Equation 5,

$$\begin{aligned} P(y=1) &= \Lambda(\mu_1 - \beta X) \\ P(y=2) &= \Lambda(\mu_2 - \beta X) - \Lambda(\mu_1 - \beta X) \\ P(y=I) &= 1 - \Lambda(\mu_{m-1} - \beta X) \end{aligned} \quad (5)$$

where $\Lambda()$ is cumulative logistic distribution.

As stated earlier, the geometric characteristics of segments on which the crash or CV driving event occurs were merged with the event-level datasets. Since multiple crashes or CV driving events could occur on the same segments in a given period of time, there may be some correlations in severity of these events because of certain unobserved site-specific factors. To account for any such correlations, a random effects framework was adopted which includes a series of site-specific random effects. This modifies the constant term in Equation 3 as follows:

$$\beta_{0i} = X\beta_0 + \omega_i \quad (6)$$

where ω_i is randomly distributed random effect for roadway segment and all other variables are as previously defined.

Results and Discussion

As stated earlier, frequency and severity models were estimated separately for crashes and CV driving events using the same set of predictors for calendar year 2020. This allowed us to compare the similarities and differences that exist between both datasets. The following subsections present and discuss the model results.

Frequency Models

Tables 5 and 6 present the results of the negative binomial models estimated for both crashes and CV driving events for multilane roads and two-lane roads, respectively. The parameter estimates, along with the standard error in parenthesis *p*-values are provided for each estimate. Broadly speaking, the model results showed that both the crash counts and CV driving event counts exhibited very similar trends with respect to each of the independent variables included in the model formulation. On both facility types, the crash model was compared with and without the CV driving event data. The relationships with other variables were shown to be stable regardless of whether the CV driving event data counts were included in the analysis. This suggested that CV driving events provided additional explanatory power as compared to the other variables, that is, sites that experience higher numbers of CV events also tend to experience higher numbers of crashes even after controlling for other important predictor variables. A 1% increase in CV driving events was associated with a 0.1% increase in crashes on both facilities. This makes sense given that CV driving events occur much more frequently than crashes. Goodness-of-fit was additionally found to increase when CV driving event data were included as

Table 6. Comparison of Negative Binomial Parameter Estimates Between Crashes and CV Driving Events on Two-Lane Roads

	Response variable = total crash count		Response variable = total CV driving events	
Parameters	Estimate (SE)	*p*-value	Estimate (SE)	*p*-value
Intercept	−6.30 (0.62)	<0.001	4.78 (0.43)	<0.001
Ln(CV driving event count)	0.13 (0.04)	0.003	na	na
Ln(AADT)	0.76 (0.07)	<0.001	0.07 (0.05)	0.154
Speed limit (mph)				
55 or more	Base condition		Base condition	
40 to 50	0.13 (0.09)	0.137	0.75 (0.14)	<0.001
35 or less	0.27 (0.12)	0.026	0.82 (0.20)	<0.001
Signal present (1 if yes, 0 otherwise)	0.52 (0.10)	<0.001	1.37 (0.15)	<0.001
Goodness-of-fit (−2×log-likelihood)				
With CV driving event data	1,004.76		3,061	
Without CV driving event data	1,013.50		na	
Likelihood ratio test for Ln(CV driving event count)				
Chi-square on 1 degree of freedom	8.74	0.003	na	

Note: Ln(CV driving event count) = natural logarithm of CV driving event count; Ln(AADT) = natural logarithm of AADT; CV = connected vehicle; SE = standard error; na = not applicable.

independent variable in the crash frequency model, as indicated by $-2 \times$ log-likelihood. The likelihood ratio test was also conducted to compare the crash frequency model with and without the CV driving event data. In all cases, the results of the test showed that the CV driving event data contributed significantly to the model.

Turning to other variables of interest, both crashes and CV driving events exhibited strong and nearly elastic relationships with AADT on multilane roads. On two-lane roads, however, CV driving events showed a weak and statistically insignificant relationship with AADT, which may be reflective of the lower penetration rates of Ford CV vehicles on these facilities. Both crashes and CV driving events were also significantly higher on segments with a signalized intersection. With respect to speed limits, both crashes and CV driving events were found to decrease with speed limits. This seems counterintuitive at first, however, this was primarily driven by a greater proportion of low-severity crashes and CV driving events occurring at lower speed limits, and also reflects the relationships between speed limits and other segment-specific factors such as access point density and the level of roadside development.

On multilane roads, the type of median also had a significant relationship with both crash frequency and CV driving event frequency. Compared with undivided roads, divided roads exhibited significantly fewer crashes and CV driving events. When opposing directions of travel are separated by a median, this helps reduce cross-median crashes—also reflected in the model (*27*). Moreover, segments with ditch-type medians had fewer instances of crash and CV driving events compared with segments encompassing other types of medians such as guardrails, concrete barriers, or raised curbs. Graded ditch-type medians allow the driver to regain control of their vehicle in cases when the vehicle has departed the travel lane, thereby reducing roadway departure crashes. Concrete barriers, guardrails, or other types of median may result in a greater number of low-severity crashes compared with ditch-type medians.

Frequency models were also developed based on crash and CV driving event type. Tables 7 and 8 show the negative binomial regression models for rear-end crashes and braking events for two-lane roads and multilane roads, respectively.

The results showed similar relationship with traffic volume and roadway geometric characteristics as seen previously for total crashes and total CV driving events. Both rear-end crashes and braking events increased with traffic volume. At higher traffic volumes, congestion may increase, which results in more braking events and a greater risk of rear-end crashes; this explains the higher sensitivity of rear-end crashes and braking events to traffic volume compared with the total crashes and total CV driving events, respectively. With respect to speed limit, both crashes and braking events occurred much more frequently as speed limits reduced. This was true on both multilane roads and two-lane roads. As mentioned previously, these relationships with speed limits are largely a function of the greater proportion of PDO crashes and CV braking events occurring on lower-speed-limit roadways. Lastly, segments with a signalized intersection showed a greater number of rear-end crashes and braking events, which was expected since intersections tend to experience queue formation, which results in a higher risk of being rear-ended and greater number of braking maneuvers.

These results indicated that the CV driving events tended to hold their relationships with various parameters even when categorized into event types. This may

Table 7. Comparison of Negative Binomial Parameter Estimates Between Rear-End Crashes and Braking Events on Multilane Roads

Parameters	Response variable = rear-end crash count		Response variable = braking events	
	Estimate (SE)	*p*-value	Estimate (SE)	*p*-value
Intercept	−12.92 (0.75)	<0.001	−2.60 (0.54)	<0.001
Ln(braking event count)	0.14 (0.04)	<0.001	na	na
Ln(AADT)	1.30 (0.08)	<0.001	0.85 (0.06)	<0.001
Speed limit (mph)				
55 or more	Base condition		Base condition	
40 to 50	0.50 (0.15)	<0.001	0.23 (0.11)	0.040
35 or less	0.80 (0.16)	<0.001	0.22 (0.12)	0.068
Signal present (1 if yes, 0 otherwise)	0.76 (0.22)	<0.001	0.91 (0.11)	<0.001
Median type				
Undivided	Base condition		Base condition	
Ditch-type median	−0.15 (0.16)	0.358	−0.23 (0.12)	0.056
Other type of median	−0.09 (0.07)	0.202	−0.02 (0.06)	0.732
Goodness-of-fit (−2×log-likelihood)				
With braking event data	3,383.36		4,560.4	
Without braking event data	3,397.80		na	
Likelihood Ratio Test for Ln(braking event count)				
Chi-square on 1 degree of freedom	14.44	<0.001	na	

Note: Ln(braking event count) = natural logarithm of braking event count; Ln(AADT) = natural logarithm of AADT; SE = standard error; na = not applicable.

Table 8. Comparison of Negative Binomial Parameter Estimates Between Rear-End Crashes and Braking Events on Two-Lane Roads

Parameters	Response variable = rear-end crash count		Response variable = braking events	
	Estimate (SE)	*p*-value	Estimate (SE)	*p*-value
Intercept	−17.97 (1.55)	<0.001	−2.56 (0.46)	<0.001
Ln(braking event count)	0.37 (0.09)	<0.001	na	na
Ln(AADT)	1.70 (0.17)	<0.001	0.20 (0.05)	<0.001
Speed limit (mph)				
55 or more	Base condition		Base condition	
40 to 50	0.42 (0.16)	<0.001	0.73 (0.15)	<0.001
35 or less	0.75 (0.21)	<0.001	0.86 (0.21)	<0.001
Signal present (1 if yes, 0 otherwise)	0.55 (0.20)	0.005	1.34 (0.15)	<0.001
Goodness-of-fit (-2×log-likelihood)				
With braking event data	−582.33		−1,298.8	
Without braking event data	−599.00		na	
Likelihood ratio test for Ln(braking event count)				
Chi-square on 1 degree of freedom	16.67	<0.001	na	

Note: Ln(braking event count) = natural logarithm of braking event count; Ln(AADT) = natural logarithm of AADT; SE = standard error; na = not applicable.

prove useful in safety analyses in which specific crash types are of particular interest, such as near intersections where a specific crash type may be overrepresented. Models were also developed for acceleration events, which exhibited similar trends to those of the rear-end crashes for both multilane roads and two-lane roads. Cornering events did not show any meaningful trends when modeled separately.

Severity Models

In addition to the segment-level analysis, a severity analysis at the event level was also carried out. Tables 9 and 10 present the results of the random effects ordinal logit regression models estimated for crashes and CV driving events on multilane and two-lane roads, respectively. The standard errors are provided in parenthesis, and statistically significant estimates at the 95% confidence level are marked with an asterisk. As stated earlier, both crash- and CV driving event severity were categorized into three levels in order of their increasing severity. When interpreting the results, a positive (or negative) coefficient for a given parameter means that the severity of crashes (or CV driving events) increases (or decreases) as that variable increases. To assist in interpreting the model results, odds ratios are also provided in Tables 9 and 10, which

Table 9. Comparison of Random Effects Ordinal Logit Parameter Estimates Between Crashes and CV Driving Events on Multilane Roads

	Response variable = crash severity (without crash-level information)		Response variable = crash severity (with crash-level information)		Response variable = CV Driving event severity	
Parameters	Estimate (SE)	Odds ratio	Estimate (SE)	Odds ratio	Estimate (SE)	Odds ratio
Speed limit (mph)						
55 or more	Base condition		Base condition		Base condition	
40 to 50	0.02 (0.09)	1.02	0.05 (0.09)	1.05	−0.14 (0.02)*	0.87
35 or less	0.06 (0.10)	1.07	0.11 (0.11)	1.12	−0.28 (0.03)*	0.75
Time period						
Jan 1–Mar 15	Base condition		Base condition		Base condition	
Mar 16–May 31	0.13 (0.08)	1.14	0.15 (0.08)	1.16	0.13 (0.005)*	1.13
Jun 1–Sep 30	0.26 (0.05)*	1.30	0.29 (0.06)*	1.34	0.02 (0.004)*	1.02
Oct 1–Dec 31	0.08 (0.06)	1.09	0.08 (0.06)	1.09	−0.04 (0.004)*	0.96
Time of day						
Midnight to 3 a.m.	Base condition		Base condition		Base condition	
3 to 6 a.m.	0.03 (0.15)	1.03	0.12 (0.15)	1.13	0.02 (0.02)	1.02
6 to 9 a.m.	−0.56 (0.12)*	0.57	−0.23 (0.14)	0.79	−0.06 (0.01)*	0.94
9 a.m. to midday	−0.46 (0.11)*	0.63	−0.11 (0.14)	0.89	−0.05 (0.01)*	0.95
Midday to 3 p.m.	−0.51 (0.10)*	0.60	−0.16 (0.13)	0.85	−0.07 (0.01)*	0.93
3 to 6 p.m.	−0.58 (0.10)*	0.56	−0.24 (0.13)	0.78	−0.11 (0.01)*	0.90
6 to 9 p.m.	−0.53 (0.11)*	0.59	−0.31 (0.12)*	0.73	−0.06 (0.01)*	0.94
9 p.m. to midnight	−0.15 (0.15)	0.86	−0.07 (0.12)	0.93	−0.02 (0.01)	0.98
Driver sobriety						
Sober			Base condition			
Intoxicated	na	na	1.15 (0.10)*	3.17	na	na
Lighting conditions						
Daylight			Base condition			
Dark lighted	na	na	0.16 (0.08)*	1.17	na	na
Dark unlighted	na	na	0.16 (0.12)	1.18	na	na
Threshold for severity						
1|2	0.95 (0.13)*	na	1.36 (0.16)*	na	−0.77 (0.02)*	na
2|3	4.10 (0.15)*	na	4.54 (0.18)*	na	0.19 (0.02)*	na
Random effect						
Variance of intercept	0.05 (0.02)*	na	0.06 (0.02)*	na	0.03 (0.002)*	na

Note: CV = connected vehicle; SE = standard error; na = not applicable.
*Statistically significant parameter estimate at 95% confidence level.

represent the change in odds of crashes or CV driving events being in the higher-severity category owing to the parameter of interest.

Two model specifications were provided for the crash severity models. The first model included only parameters that were available for the event severity analyses. These included speed limit information and time of occurrence of crash or CV driving event. The same parameters were thus also included in the CV driving event severity model. The second crash severity model considered parameters related to crash-level data, such as weather and lighting conditions at the time of the crash, driver sobriety, and type of vehicle involved in the crash. Other roadway geometric information such as the number of lanes, shoulder width, and median width did not have any significant impact on the severity of either crash or CV driving events. Both crash models generally exhibited similar trends. One exception was the time of day effect on the crash severity of two-lane roads. When driver sobriety was considered in the model, crash severity tended to increase during nighttime. When driver sobriety was not included, crash severity reduced during the same time period. Although these effects were not statistically different from zero in either case. The subsequent discussion is based on a comparison of the results without considering the crash-level information, since this allows for direct comparisons between the crash severity and CV driving event severity models.

In general, the results from the analysis of both multilane and two-lane roads showed that the severity of crashes and CV driving events tended to be affected by roadway geometric characteristics in a similar manner. With respect to speed limits, the severity of crashes and CV driving events tended to increase as the speed limit increased. This was expected since crashes occurring at higher speeds tend to be more severe. Similarly, CV driving events occurring at higher speeds, such as braking

Table 10. Comparison of Random Effects Ordinal Logit Parameter Estimates Between Crashes and CV Driving Events on Two-Lane Roads

	Response variable = crash severity (without crash-level information)		Response variable = crash severity (with crash-level information)		Response variable = CV driving event severity	
Parameters	Estimate (SE)	Odds ratio	Estimate (SE)	Odds ratio	Estimate (SE)	Odds ratio
Speed limit (mph)						
55 or more	Base condition		Base condition		Base condition	
40 to 50	−0.12 (0.19)	0.88	−0.12 (0.19)	0.89	−0.01 (0.05)	0.99
35 or less	−0.43 (0.23)	0.65	−0.41 (0.23)	0.67	−0.43 (0.06)*	0.65
Time period						
Jan 1–Mar 15	Base condition		Base condition		Base condition	
Mar 16–May 31	0.17 (0.27)	1.19	0.10 (0.27)	1.11	0.09 (0.02)*	1.10
Jun 1–Sep 30	0.19 (0.18)	1.21	0.18 (0.19)	1.20	0.03 (0.01)*	1.03
Oct 1–Dec 31	−0.02 (0.20)	0.98	−0.02 (0.20)	0.98	−0.06 (0.01)*	0.95
Time of day						
Midnight to 3 a.m.	Base condition		Base condition		Base condition	
3 to 6 a.m.	−0.13 (0.70)	0.88	0.01 (0.71)	1.01	0.33 (0.06)*	1.39
6 to 9 a.m.	−0.09 (0.46)	0.91	0.18 (0.48)	1.19	0.11 (0.06)	1.11
9 to midday	0.13 (0.46)	1.14	0.43 (0.47)	1.54	0.17 (0.06)*	1.19
Midday to 3 p.m.	0.08 (0.43)	1.09	0.39 (0.45)	1.48	0.20 (0.06)*	1.22
3 to 6 p.m.	−0.12 (0.43)	0.88	0.19 (0.45)	1.21	0.14 (0.05)*	1.14
6 to 9 p.m.	0.15 (0.45)	1.16	0.34 (0.46)	1.41	0.14 (0.06)*	1.16
9 p.m. to midnight	0.37 (0.49)	1.45	0.31 (0.50)	1.36	0.19 (0.06)*	1.21
Driver sobriety						
Sober			Base condition			
Intoxicated	na	na	1.36 (0.30)*	3.89	na	na
Threshold for severity						
1|2	1.23 (0.42)*	na	1.55 (0.44)*	na	−0.35 (0.06)*	na
2|3	3.97 (0.47)*	na	4.34 (0.49)*	na	0.67 (0.06)*	na
Random effect						
Variance of intercept	0.11 (0.08)	na	0.11 (0.08)	na	0.09 (0.01)*	na

Note: CV = connected vehicle; SE = standard error; na = not applicable.
*Statistically significant parameter estimate at 95% confidence level.

and turning, will be severe (harsh). The one exception was the crash severity model for multilane roads, which showed opposite trends. Other roadway geometric variables such as lane width, shoulder width, median width, and type were not found to be statistically significant predictors of severity for either crashes or CV driving events.

Perhaps the most interesting results were seen with respect to the time period. The travel restrictions imposed as a result of the COVID-19 pandemic resulted in lower VMT and fewer trips being generated across the entire United States (*6*). In the state of Michigan, stay-at-home orders were issued on March 23, 2020, and stayed in effect until May 31 with some restrictions loosened starting April 24 of the same year (*24*). During this period, crash frequency reduced as a result of lower traffic volumes. However, this also resulted in higher travel speeds, thereby increasing the frequency of more severe crashes (*28*, *29*). The model results in the present analysis showed similar trends. The severity of both crashes and CV driving events was significantly higher during the early months of the pandemic (March 15 to May 31, 2020) compared with the prepandemic period. Interestingly, the parameter estimates for this group of variables were not statistically different from zero for the crash severity model, especially with respect to two-lane roads. This was primarily because of the lower sample size of crashes on these roadways, as only 1 year of crash data were being analyzed. CV driving event data, on the other hand, showed statistically significant relationships even though only 1 year of data were used. This was because of the greater frequency of these events compared with crashes. This reinforces the usefulness of such data in cases in which sufficient crash data are unavailable.

With respect to time of day, both crash severity and CV driving event severity exhibited similar trends. On multilane roads, severities tended to be lower during daytime and peak time periods, which may be reflective of the level of congestion on these roadways during these periods. On two-lane roads, severity was highest from 9 p.m. to midnight, however, the severity of both crashes and CV driving events was also higher during certain hours throughout the day, which may be the result of

certain artifacts of the data. Again, these trends were nonsignificant for the crash severity model, particularly for the two-lane roads. CV driving event data, on the other hand, showed relationships that were generally statistically significant.

In addition to roadway geometric data, several parameters related to crash- and driver-level data, such as weather and lighting conditions at the time of the crash, driver sobriety, and type of vehicle involved were included in the crash severity model. Some of these parameters showed significant effects on severity, however, the inclusion of these parameters did not affect the parameter estimates of the current model. Thus, to allow for direct comparisons between the crash severity model and CV driving event severity model, these variables were excluded from the analysis shown here.

Conclusions

Police-reported crash data have been the most critical element of any traffic safety analysis. However, inherent limitations of crash data make it challenging to conduct predictive analysis in certain scenarios. This study has demonstrated the utility of CV driving event data in safety analysis by assessing whether the CV driving events have similar relationships to crashes with respect to traffic volume and roadway geometric characteristics. These relationships were evaluated by estimating a series of frequency- and severity models for both crash- and CV driving event data using the same set of predictors. An investigation into the applicability of CV driving event data in network screening procedures was also undertaken; this showed promising results, as more than 50% of the sites identified as high-risk locations based on crash data were also identified as high-risk using the CV driving event data, regardless of the normalization scheme. This will assist transportation agencies in preparing countermeasure prioritization schemes and allocating resources for their roadway network by utilizing CV driving event data as a supplement to the traditional crash data or as an alternative to crash data in cases in which sufficient crash data is lacking.

The results from the safety analyses showed significant positive correlations between crash frequency and CV driving event frequency on different classes of roadway facilities. The statistical models showed that the CV driving events tended to be influenced by traffic volume and roadway geometric characteristics in a way similar to crashes. This was true at both segment- and individual event levels. Moreover, the goodness-of-fit of crash frequency models was improved when CV driving event frequency was used as a predictor for crashes. When separated into subsets, the relationships were still found to be strong between CV driving event types and associated crash types. This showed that the CV driving event data provided a statistically significant surrogate safety measure as a complement to police-reported crash data. The results from the severity analysis further demonstrated the usefulness of Ford CV driving event data in traffic safety analysis beyond the typical crash frequency models. As more detailed event-level information was utilized in the analysis, including driver and vehicle characteristics, the severity analysis could be further strengthened.

Interestingly, CV driving event data were also able to show trends in severity resulting from the COVID-19 pandemic. The travel restrictions imposed because of the pandemic resulted in lower traffic volumes, which translated into higher speeds and, consequently, a greater severity of crashes and CV driving events. Because of their smaller sample size, the crash severity models did not show statistical significance in such trends. However, the CV driving events clearly exhibited these trends, thereby strengthening their usefulness in traffic safety analysis.

With time, as the penetration rate of CVs capable of generating driving event data increases, the performance of CV driving event data will also improve significantly. One of the limitations in the current study design is that the CV event data were obtained from Ford vehicles that have the FordPass mobile application enabled. Thus, the current sample of vehicles does not represent the entire vehicle fleet on any typical roadway network. Leveraging information about the sociodemographic characteristics of FordPass mobile app users could overcome this limitation to some degree. These data could potentially be utilized to create a representative sample of CV driving data based on geographic region and other sociodemographic characteristics. Integrating information about the number of Ford CVs on the roadways might further improve the results. Lastly, the study considered only 1 year of Ford CV driving event data owing to data availability issues. Replicating this study with additional years of CV driving data might further reinforce the trends observed in this study.

Nevertheless, the results indicated that crash- and the CV driving event data can be used interchangeably in traffic safety analysis. The results strongly supported the use of CV driving event data as a surrogate safety measure. Crash data analyses are often plagued by the limitations of data input from inaccurate recording on site, inaccurate data manipulation, and the inability to account for near misses. However, with the use of CV driving event data, transportation agencies will be given insights into events that could potentially lead to crash reductions (e.g., behaviors), ultimately looking at safety from a systemic approach rather than the typical systematic approach used by many professionals.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: N. Gupta, P. Savolainen; data collection: N. Gupta, H. Jashami, T. Barrette, W. Powell; analysis and interpretation of results: N. Gupta, H. Jashami, P. Savolainen, T. Gates, T. Barrette, W. Powell; draft manuscript preparation: N. Gupta, H. Jashami, P. Savolainen, T. Gates, T. Barrette, W. Powell. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: T. Barrette and W. Powell are employees of Ford Motor Company, which funded this research.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The connected vehicle event data and the funding for this research were provided by Ford Motor Company (Project No. MSU0157).

ORCID iDs

Nischal Gupta https://orcid.org/0000-0001-7033-1663
Hisham Jashami https://orcid.org/0000-0002-5511-7543
Peter T. Savolainen https://orcid.org/0000-0001-5767-9104
Timothy J. Gates https://orcid.org/0000-0002-7429-0990
Timothy Barrette https://orcid.org/0000-0002-7656-3454
Wesley Powell https://orcid.org/0009-0001-9939-5721

References

1. National Highway Traffic Safety Administration. FARS Encyclopedia. https://www-fars.nhtsa.dot.gov/Main/index.aspx. Accessed August 2, 2022.
2. National Center for Injury Prevention and Control. Motor Vehicle Crash Injuries: Costly but Preventable. *Atlanta: Centers for Disease Control and Prevention*. https://archive.cdc.gov/#/details?url=https://www.cdc.gov/vitalsigns/crash-injuries/index.html. Accessed August 2, 2022.
3. Blincoe, L., T. R. Miller, E. Zaloshnja, and B. A. Lawrence. *The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised)*. National Center for Statistics and Analysis, Washington, D.C., 2015.
4. American Association of State Highway and Transportation Officials (AASHTO). *Highway Safety Manual*. American Association of State Highway and Transportation Officials, Washington, D.C., 2010.
5. Savolainen, P. T., F. L. Mannering, D. Lord, and M. A. Quddus. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accident Analysis & Prevention*, Vol. 43, No. 5, 2011, pp. 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025.
6. Bamney, A., H. Jashami, S. Sonduru Pantangi, J. Ambabo, M.-U. Megat-Johari, Q. Cai, N. Gupta, and P. T. Savolainen. Examining Impacts of COVID-19-Related Stay-at-Home Orders Through a Two-Way Random Effects Model. *Transportation Research Record: Journal of the Transportation Research Board*, 2023. 2677: 255–266.
7. Wang, B., S. Hallmark, P. Savolainen, and J. Dong. Crashes and Near-Crashes on Horizontal Curves Along Rural Two-Lane Highways: Analysis of Naturalistic Driving Data. *Journal of Safety Research*, Vol. 63, 2017, pp. 163–169. https://doi.org/10.1016/J.JSR.2017.10.001.
8. Liu, J., and A. J. Khattak. Delivering Improved Alerts, Warnings, and Control Assistance Using Basic Safety Messages Transmitted between Connected Vehicles. *Transportation Research Part C: Emerging Technologies*, Vol. 68, 2016, pp. 83–100. https://doi.org/10.1016/J.TRC.2016.03.009.
9. Khattak, A. J., and B. Wali. Analysis of Volatility in Driving Regimes Extracted from Basic Safety Messages Transmitted Between Connected Vehicles. *Transportation Research Part C: Emerging Technologies*, Vol. 84, 2017, pp. 48–73. https://doi.org/10.1016/J.TRC.2017.08.004.
10. Jahangiri, A., V. J. Berardi, and S. G. MacHiani. Application of Real Field Connected Vehicle Data for Aggressive Driving Identification on Horizontal Curves. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 19, No. 7, 2018, pp. 2316–2324. https://doi.org/10.1109/TITS.2017.2768527.
11. Talebpour, A., H. Mahmassani, F. Mete, and S. Hamdar. Near-Crash Identification in a Connected Vehicle Environment. *Transportation Research Record: Journal of the Transportation Research Board*, 2014. 2424: 20–28.
12. Musicant, O., H. Bar-Gera, and E. Schechtman. Electronic Records of Undesirable Driving Events. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 13, No. 2, 2010, pp. 71–79. https://doi.org/10.1016/J.TRF.2009.11.001.
13. Simons-Morton, B. G., M. C. Ouimet, Z. Zhang, S. E. Klauer, S. E. Lee, J. Wang, P. S. Albert, and T. A. Dingus. Crash and Risky Driving Involvement Among Novice Adolescent Drivers and Their Parents. *American Journal of Public Health*, Vol. 101, No. 12, 2011, p. 2362. https://doi.org/10.2105/AJPH.2011.300248.
14. Simons-Morton, B. G., Z. Zhang, J. C. Jackson, and P. S. Albert. Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes? *American Journal of Epidemiology*, Vol. 175, No. 10, 2012, p. 1075. https://doi.org/10.1093/AJE/KWR440.
15. Kamrani, M., B. Wali, and A. J. Khattak. Can Data Generated by Connected Vehicles Enhance Safety? Proactive Approach to Intersection Safety Management. *Transportation Research Record: Journal of the Transportation Research*, 2017. 2659: 80–90.
16. Kamrani, M., R. Arvin, and A. J. Khattak. Extracting Useful Information from Basic Safety Message Data: An Empirical Study of Driving Volatility Measures and Crash Frequency at Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672: 290–301.
17. Wali, B., A. J. Khattak, H. Bozdogan, and M. Kamrani. How Is Driving Volatility Related to Intersection Safety?

A Bayesian Heterogeneity-Based Analysis of Instrumented Vehicles Data. *Transportation Research Part C: Emerging Technologies*, Vol. 92, 2018, pp. 504–524. https://doi.org/10.1016/J.TRC.2018.05.017.

18. Paleti, R., N. Eluru, and C. R. Bhat. Examining the Influence of Aggressive Driving Behavior on Driver Injury Severity in Traffic Crashes. *Accident Analysis & Prevention*, Vol. 42, 2010, pp. 1839–1854. https://doi.org/10.1016/j.aap.2010.05.005.
19. Jun, G., J. Ogle, and R. Guensler. Relationships Between Crash Involvement and Temporal-Spatial Driving Behavior Activity Patterns: Use of Data for Vehicles with Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*, 2007. 2019: 246–255.
20. Haque, M. M., and S. Washington. The Impact of Mobile Phone Distraction on the Braking Behaviour of Young Drivers: A Hazard-Based Duration Model. *Transportation Research Part C: Emerging Technologies*, Vol. 50, 2015, pp. 13–27. https://doi.org/10.1016/J.TRC.2014.07.011.
21. Mollicone, D., K. Kan, C. Mott, R. Bartels, S. Bruneau, M. van Wollen, A. R. Sparrow, and H. P. A. Van Dongen. Predicting Performance and Safety Based on Driver Fatigue. *Accident Analysis & Prevention*, Vol. 126, 2019, pp. 142–145. https://doi.org/10.1016/J.AAP.2018.03.004.
22. Kamla, J., T. Parry, and A. Dawson. Analysing Truck Harsh Braking Incidents to Study Roundabout Accident Risk. *Accident Analysis & Prevention*, Vol. 122, 2019, pp. 365–377. https://doi.org/10.1016/J.AAP.2018.04.031.
23. Desai, J., H. Li, J. K. Mathew, Y.-T. Cheng, A. Habib, and D. M. Bullock. Correlating Hard-Braking Activity with Crash Occurrences on Interstate Construction Projects in Indiana. *Journal of Big Data Analytics in Transportation*, Vol. 3, No. 1, 2021, pp. 27–41. https://doi.org/10.1007/s42421-020-00024-x.
24. Office of the Governor State of Michigan. Governor Whitmer Signs "Stay Home, Stay Safe" Executive Order. https://www.michigan.gov/whitmer/news/press-releases/2020/03/23/governor-whitmer-signs-stay-home-stay-safe-executive-order. Accessed July 11, 2022.
25. American Association of State Highway and Transportation Officials (AASHTO). *Highway Safety Manual*. American Association of State Highway and Transportation Officials, Washington, D.C., 2010.
26. Washington, S., M. Karlaftis, and F. L. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2011.
27. Donnell, E. T., and J. M. Mason. Methodology to Develop Median Barrier Warrant Criteria. *Journal of Transportation Engineering*, Vol. 132, No. 4, 2006, pp. 269–281. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:4(269).
28. Hughes, J. E., D. Kaffine, and L. Kaffine. Decline in Traffic Congestion Increased Crash Severity in the Wake of COVID-19. *Transportation Research Record: Journal of the Transportation Research Board*, 2023. 2677: 892–903.
29. Gupta, N., A. Bamney, A. Rostami, E. Kamjoo, and P. T. Savolainen. How Did the COVID-19 Pandemic Affect Driver Speed Selection and Crash Risk on Rural Freeways? *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 97, 2023, pp. 181–206. https://doi.org/10.1016/j.trf.2023.07.008.