Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 6562649, 19 pages
https://doi.org/10.1155/2022/6562649

Research Article

Crash/Near-Crash Analysis of Naturalistic Driving Data Using Association Rule Mining

Yansong Qu, Zhenlong Li, Qin Liu, Mengniu Pan, and Zihao Zhang

The College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China

Correspondence should be addressed to Zhenlong Li; lzl@bjut.edu.cn

Received 24 May 2022; Revised 19 August 2022; Accepted 31 August 2022; Published 5 October 2022

Academic Editor: Yajie Zou

This study explores the associations between crash/near-crash (C/NC) events and roadway, driver-related, and environmental factors in naturalistic driving studies (NDS). We used the Naturalistic Engagement in Secondary Tasks (NEST) dataset, which is massive and detailed and contains 50 million miles of naturalistic driving data resulting from the Strategic Highway Research Program 2 (SHRP2). Association rule mining (ARM) is applied to extract the rules for frequently occurring events. The generated association rules are filtered by four metrics (support, confidence, lift, and conviction) and validated by the lift increase criterion. A three-step analysis is performed to obtain a comprehensive understanding of the rules of C/NC events. The 20 most frequent items are first selected to investigate their relationship with the C/NC events. Subsequently, the association rules are used to identify the factors contributing to C/NC events. Finally, correlations between contributing factors and different severities of crashes (I—most severe, II—police-reportable, III—minor crash, and IV—low-risk tire strike) are analyzed by ARM. The results demonstrate that C/NC events occur most frequently on straight and level road segments with no controlled intersections or traffic control devices when drivers are performing secondary tasks. Thus, the reasons for these crashes are carelessness and overconfidence. In addition, a median strip or barrier and a wider road can significantly reduce the frequency and severity of crash events. Moreover, gender, age, average annual mileage, and secondary tasks are highly correlated with the frequency and severity of C/NC events. Drivers with visual-spatial disabilities or crash records are more likely to be involved in the most severe crash events. Near-crash events occur more frequently at higher traffic density and on roads with traffic control devices and controlled intersections. These conditions may keep drivers alert, preventing crashes.

1. Introduction

The National Highway Traffic Safety Administration (NHTSA) data [1] show that approximately 38,680 people died in traffic crashes in the United States in 2020, representing an increase of almost 7.2% compared to the 36,096 fatalities reported in 2019 and the largest number of fatalities since 2007. The increase in traffic crashes has harmed many families, although most of the injuries and deaths could have been averted. Thus, it is essential to determine the correlations between the contributing factors and crash/near-crash (C/NC) events to minimize their occurrence. However, many factors contribute to C/NC events, with latent correlations hidden in the C/NC data. Thus, it is challenging to extract the correlations between the contributing factors and the causes of C/NC events to prevent them. Consequently, traffic safety has become an urgent and crucial topic in transportation research.

Data acquisition is a critical prerequisite for traffic safety studies. Many safety studies [2–4] have focused on extracting associations between C/NC events and roadway features using police report data due to easy accessibility. However, the lack of available factors, such as driving behavior and driver characteristics, has limited the comprehensiveness of these studies. Therefore, several experimental studies [5–7] have analyzed the impacts of different driving behaviors on C/NC events in a simulated environment. In experimental studies, dozens of drivers were recruited for experiments. For example, in secondary task engagement experiments, participants are asked to perform certain secondary tasks under specific C/NC conditions. Eye movement, heart rate, and vehicle kinetic data are simultaneously recorded during

the experiments [8]. Although experimental studies can extract valuable information because of their ability to simulate C/NC conditions, they may not be able to mine the latent rules of C/NC events for two main reasons [9–12]: (1) The participants are equipped with eye-tracking glasses, galvanic skin resistance (GSR) electrodes, wearable sensors, optical probes, and photoplethysmography (PPG) sensors to obtain data from multiple sources. The participants may not feel comfortable in the simulated driving environment due to the equipment. Therefore, the applicability of the experiment's results is questionable. (2) Obtaining instructions from a computer screen rather than responding to traffic conditions is common in driving simulations. This situation does not accurately represent the real-world driving experience.

Many studies used observational data to ensure the transferability of the results to real-life conditions [13–16]. Observational studies or naturalistic driving studies (NDS) [17] provide realistic conditions to gather C/NC data for accident analysis and prevention. Multichannel video, sensor, kinematic, and vehicle network data can be obtained from vehicles equipped with a data acquisition system (DAS) in a naturalistic driving setting. The highly detailed and comprehensive dataset is suitable for traffic safety studies and many other research fields.

Detailed and comprehensive datasets have been obtained, representing a solid foundation for traffic safety analysis. Researchers used these datasets and different methods to analyze different aspects of traffic safety. Some researchers used statistical models to reveal the correlations between variables and the occurrence of C/NC events using NDS. For instance, Papazikou et al. [18] investigated vehicle kinematics during crashes to obtain reliable indicators of the time to collision (TTC). Kreusslein et al. [19] focused on the characteristics of mobile phone calls, including the call duration, glance behavior, call type, and mobile phone location, to determine the influence of making mobile phone calls. Schlick et al. [20] used hierarchical regression models to determine the associations between motor vehicle crashes and different contributing factors.

Driving behavior analysis and machine learning methods have been used to identify the cause of C/NC events. Zou et al. [21] predicted vehicle acceleration using behavioral semantic analysis to prevent accidents caused by rapid acceleration. Guo et al. [22] utilized SHapley Additive exPlanation (SHAP) to analyze the importance of features related to crash events; sharp deceleration was the most important feature.

Association rule mining (ARM) has been proposed for crash analysis [23, 24]. ARM is widely used in the traffic safety field because it can reveal the intrinsic relationships between the contributing factors and the accidents without assumptions and significantly outperforms traditional modelling techniques. A summary of the applications of ARM for crash analysis is presented in Table 1.

Several studies [33–37] used ARM for crash analysis under different conditions, such as truck crashes or near crashes. Unlike these studies, we propose a three-step method using the frequent pattern (FP) growth algorithm [38] to mine the correlations between different categorical variables and C/NC events using the Naturalistic Engagement in Secondary Tasks (NEST) dataset [39]. The 20 most frequent items are first selected to determine which features are associated with C/NC events. The association rules describing the factors contributing to C/NC crash events are then identified. Finally, association rules are used to analyze crash events of different severities. Suggestions for practical applications are provided. The flowchart of the proposed approach is illustrated in Figure 1.

The remainder of this paper is organized as follows. Section 2.1 presents the dataset and preprocessing steps. The methodology is described in Section 2.2, focusing on the principles of the FP growth algorithm and the formulations of four metrics: support, confidence, lift, and conviction. The results are presented and discussed in Section 3, the findings and discussions are drawn in Section 4, and conclusions are summarized in Section 5.

2. Materials and Methods

2.1. Data Description

2.1.1. Dataset Overview. We used C/NC data from the NEST dataset [39], which is a subset of the Strategic Highway Research Program 2 (SHRP2) database produced under the collaboration between the Virginia Tech Transportation Institute (VTTI) and the Toyota Collaborative Safety Research Center (Toyota CSRC). This dataset contains high-level data and detailed time-series data on secondary task engagement and distraction-related safety-critical events (SCEs) during real-world driving. The summary data provide information at the event level, and the time-series data provide frame-by-frame detailed information at the millisecond level. We only used the summary data in this study.

The summary data contain information on the event severity of baseline, crash, and near-crash events, with a total of 1080 samples. We did not consider the baseline data because they contain no C/NC events. The duration of the C/NC events was 30 s, including 20 s prior to the event and 10 s following it. The summary data comprised 36 items. The subtasks and environmental conditions were split into three fractions for each 10 s duration, while the driver information and other information were not. After deleting samples with too many missing values, we obtained 699 C/NC event samples.

2.1.2. Variables. The raw summary data of the C/NC events contains 36 categorical variables. Twenty of them were chosen to analyze the patterns of the C/NC events. The remaining 16 variables were not chosen for the following three reasons: (1) a large percentage of missing values, (2) heavily skewed distribution, and (3) overlap in meaning. For example, the stop sign, merge sign, yield sign, slow or other warning signs, and railroad crossing sign variables are included in the raw summary data. However, most of the values are blank because these signs do not occur frequently; thus, the distribution is skewed. In addition, the traffic control

Table 1: Summary of ARM applications for crash analysis.

Authors	Publication year	Methods	Datasets
Wu et al. [3]	2019	Fault tree analysis (FTA) & Apriori	STATS19
Yu et al. [25]	2019	Apriori	Reported crashes in Wisconsin in 2016
Hong et al. [26]	2020	Apriori	Truck-involved crashes data in Korean Expressway Corporation
Hong et al. [27]	2020	Apriori	HAZMAT vehicle-involved crash in South Korea from 2008 to 2017
Das et al. [28]	2021	Apriori	Police-reported crashes in Louisiana from 2010 to 2015
Kong et al. [29]	2021	Apriori	HPMS & SPMD
Kong et al. [30]	2021	Apriori	VCC50 Elite dataset
Montella et al. [31]	2021	Apriori	Police reports in Italy from 2001 to 2011 combined with site inspections' information
Tamakloe et al. [32]	2022	Binary logit regression (BLR) & Apriori	Motorcycle-involved collisions in Greater Accra Region

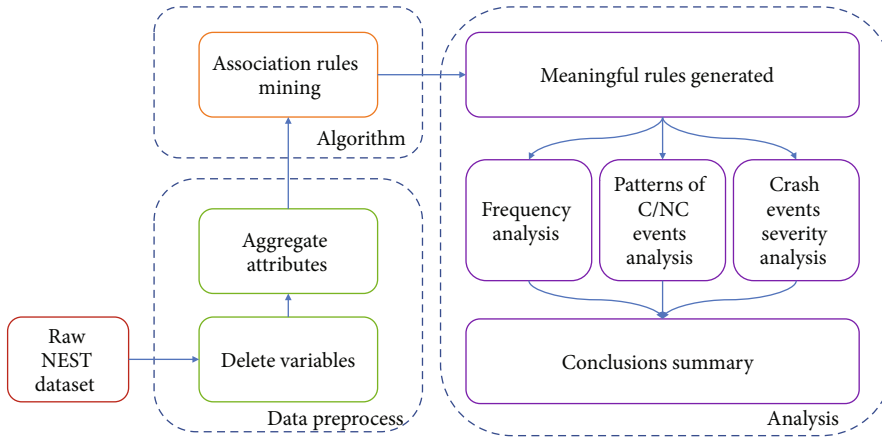

Figure 1: Flowchart of the proposed approach.

variable represents these signs at a higher level. Therefore, these variables were deleted, and only the traffic control variable was used. Note that crucial variables were retained even if they had a skewed distribution or an overlap in meaning.

Some of the chosen variables required aggregation because they contained many attributes, skewing the distribution. Therefore, the attributes of these variables were categorized into a higher level, such as secondary task, traffic density, locality, age group, and annual miles. For example, different secondary tasks (including no secondary task) were aggregated into secondary tasks (yes) and no secondary tasks (no). This approach was different from a previous study [40] because all C/NC events were analyzed comprehensively in this paper rather than focusing on one aspect. More details on the variables are presented in Table 2.

2.1.3. Distribution of Attributes. The distribution of attributes is significant for hyperparameter selection, such as the support value, and influences the association rules generated by ARM. For example, some attributes of a variable occurred infrequently and might not been considered because of a high support value; thus, they might be filtered out by ARM and excluded from the association rules, resulting in errors in evaluating the attribute's contribution to C/NC events.

Figure 2 describes the distribution of attributes for the crash and near-crash events. There were 447 crash events and 252 near-crash events.

Figure 2 shows that (1) most percentages are greater than 0.05, indicating that 0.05 might be a suitable initial support value; (2) some attributes are associated with a higher proportion of crash events than near-crash events, such as no lanes, lane number ≤ 2, improper driver behavior, and teenager driving. This implies a correlation between the severity of events and these attributes.

2.2. Methodology. Recent studies used various techniques to conduct pattern mining using large amounts of crash data, such as ARM [36], Bayesian networks [41], neural networks [42], linear regression networks [43], cluster analysis [44], random forests [45], and support vector machine [46]. ARM has the advantage of finding meaningful associations and providing valuable insights into the interdependence between roadway, environmental, and driver-related factors and the frequency and severity of crashes [29]. Besides, ARM is more suitable for discovering patterns in large data

Table 2: Description of the categorical variables and their attributes.

	Variable	Attribute	Description/definition
Severity	Event severity	Crash	The outcome of event, only considering the crash event and near-crash event
		Near crash	
	Crash severity	I—most severe	The ranking of crash severity. This variable is coded only for events that include a crash
		II—police-reportable crash	
		III—minor crash	
		IV—low-risk tire strike	
Road	Traffic flow	Divided (median strip or barrier)	The roadway design
		No lanes	
		Not divided	
		One-way traffic	
	Travel lanes	Lanes ≤ 2	The number of lanes
		2 < lanes ≤ 7	
	Alignment	Curve	Roadway alignment
		Straight	
	Road grade	Nonlevel	Roadway grade
		Level	
Driver	Driver behavior	Yes	Whether improper behaviors such as aggressive driving appear or not
		No	
	Driver impairments	Yes	Whether any apparent impairments such as intoxication or fatigue appear or not
		No	
	Secondary task	Secondary task observed	Whether the driver engages in any secondary tasks or not
		No secondary task observed	
	Age group	16-19	The age group of drivers
		20-24	
		25-34	
		35-64	
		65-99	
	Gender	M	Gender of drivers
		F	
	Annual miles	Less than 10,000 miles	Estimated average annual mileage over five years
		10,000-15,000 miles	
		More than 15,000 miles	
	NUMVIOL	0	The number of moving or traffic violations the participant has had in the three years
		1	
		2 or more	
	NUMcrash	0	The number of crashes the participant has been involved in during the three years
		1	
		2 or more	
	Score full text	1—perfect	The clock drawing performed by the participant was scored based on a six-point scoring system
		2—minor visuospatial errors	
		3—inaccurate time, minor visuospatial errors	
		4—moderate visuospatial errors	

Table 2: Continued.

	Variable	Attribute	Description/definition
Environment	Traffic density	Free flow Stable flow Unstable/forced flow	The density of traffic flow
	Intersections entered	Yes No	Whether interrupted by controlled intersections or not
	Traffic control	Yes No	Whether influenced by any traffic controls or not
	Surprised	Yes No	Whether anything might be considered surprising to an average driver or to this particular driver occurs or not
	Locality	Residential Business/industrial Others	The surrounding area

volumes than confirming hypotheses [36] and is not influenced by missing values. Thus, it is preferable to machine learning and linear regression methods. Therefore, ARM was chosen to analyze C/NC data.

The Apriori algorithm [23] is considered the most popular and efficient ARM method compared to the weighted classification based on association rule (WCBA) method [47], fast classification based on association rule (FCBA) method [48], and the maximal frequent itemset algorithm (MAFIA) [49]. However, it scans the entire dataset for frequent items, resulting in high computational complexity, especially for a large dataset. The FP growth algorithm [50] is an improvement of the Apriori algorithm that requires only two scans of the database to develop the FP tree. Thus, it can identify frequent items in a large database with a low execution time. Due to the advantages of the FP growth algorithm, it is used here to extract frequent items.

In this study, the association rules are mined in two steps: (1) the FP growth algorithm is used to detect frequent item sets and (2) association rules are mined from the frequent item sets.

It is assumed that $I = \{i1, i2, \cdots, im\}$ is a collection of categorical variables (item sets), and $T = \{t1, t2, \cdots, tn\}$ is a collection of C/NC events (transactions), where m is the number of item sets that is much greater than n, which is the number of transactions. All association rules are generated based on I and T. However, not all the association rules are needed. For example, $\{\text{trafficflow} = \text{no lanes}\} \longrightarrow \{\text{trafficdensity} = \text{free flow}\}$ may be an association rule with a high support value, but it may not provide any new or meaningful information because a road with no lanes implies a low-grade road unsuitable for high traffic density. Thus, these types of rules should be discarded. X is defined as the antecedent (e.g., $\{\text{trafficflow} = \text{no lanes}\}$), and Y is defined as the consequent (e.g., $\{\text{event} = \text{near} - \text{crash event}\}$). The antecedent and consequent are used to discard meaningless association rules. However, this does not indicate that X is the cause of Y, Y is the result of X, or X and Y have a causal relationship. Four performance metrics are typically used to test the model performance and validity: support, confidence, lift, and conviction. The support indicates how frequently the itemset appears in the dataset; it is the ratio of the number of transactions containing the item set to the total number of transactions. The confidence is the percentage of all transactions satisfying X that also satisfy Y. It is the ratio of the number of transactions including items X and Y to the number of transactions including item X. The lift of a rule refers to the frequency of items X and Y in a transaction. However, the frequency of item X or item Y should be simultaneously considered. The lift value reflects the correlation between X and Y in the association rules. When the lift value is greater than 1, the higher the value, the higher the positive correlation between X and Y is. When the lift value is less than 1, the lower the value, the higher the negative correlation between X and Y is. When the lift value is equal to 1, there is no correlation between X and Y. A rule with a single antecedent and a single consequent is referred to as a 2-item rule. Similarly, a rule with k-1 antecedents and a single consequent is denoted as a k-item rule, where k is the sum of the number of antecedents and the number of consequents. The support, confidence, lift, and conviction are computed as follows:

$$\begin{aligned}
\text{Support}(X \longrightarrow Y) &= P(X \cup Y),\\
\text{Confidence}(X \longrightarrow Y) &= P(X|Y) = \frac{P(X \cup Y)}{P(X)},\\
\text{Lift}(X \longrightarrow Y) &= \frac{\text{confidence}(X \longrightarrow Y)}{P(Y)} = \frac{P(X \cup Y)}{P(X)P(Y)},\\
\text{Conviction}(X \longrightarrow Y) &= \frac{\text{confidence}(X \longrightarrow Y)}{P(Y)} = \frac{P(X \cup Y)}{P(X)P(Y)},
\end{aligned} \tag{1}$$

where X is the antecedent, Y is the consequent, $P(X)$ is the percentage or probability of a transaction containing item X, $\text{support}(X \longrightarrow Y)$ is the support value of the association rule $X \longrightarrow Y$, $\text{confidence}(X \longrightarrow Y)$ is the confidence value of the association rule $X \longrightarrow Y$, $\text{lift}(X \longrightarrow Y)$ is the lift value of the association rule $X \longrightarrow Y$, and $\text{conviction}(X \longrightarrow Y)$ is the conviction value of the association rule $X \longrightarrow Y$.

Variable	Attributes	Crash		Near-crash	
		Count	Percentage	Count	Percentage
TRAFFIC FLOW	Divided (median strip or barrier)	99	0.22	110	0.44
	No lanes	90	0.20	0	0.00
	Not divided	235	0.53	129	0.51
	One-way traffic	23	0.05	13	0.05
TRAVEL LANES	lanes<=2	280	0.63	72	0.29
	2<lanes<=7	166	0.37	180	0.71
ALIGNMENT	Curve	57	0.13	19	0.08
	Straight	389	0.87	233	0.92
ROADGRADE	Non-level	68	0.15	20	0.08
	Level	379	0.85	232	0.92
DRIVERBEHAVIOR	Yes	240	0.54	63	0.25
	No	207	0.46	189	0.75
DRIVERIMPAIRMENTS	Yes	36	0.08	3	0.01
	No	411	0.92	249	0.99
SECONDARYTASK	Secondary task observed	424	0.95	228	0.90
	No secondary task observed	23	0.05	24	0.10
AGEGROUP	16-19	141	0.32	18	0.07
	20-24	141	0.32	99	0.39
	25-34	48	0.11	57	0.23
	35-64	54	0.12	54	0.21
	65-99	57	0.13	21	0.08
GENDER	M	192	0.43	132	0.52
	F	255	0.57	120	0.48
ANNUAL MILES	Less than 10,000 miles	180	0.40	54	0.21
	10,000-15,000 miles	129	0.29	87	0.35
	more than 15,000 miles	126	0.28	108	0.43
NUMVIOL	0	267	0.60	114	0.45
	1	96	0.21	75	0.30
	2 or More	81	0.18	63	0.25
NUMCRASH	0	228	0.51	153	0.61
	1	141	0.32	63	0.25
	2 or More	69	0.15	33	0.13
SCOREFULLTEXT	1-Perfect	102	0.23	66	0.26
	2-Minor visuospatial errors	279	0.62	171	0.68
	3-Inaccurate time, minor visuospatial errors	30	0.07	12	0.05
	4-Moderate visuospatial errors	24	0.05	3	0.01
TRAFFIC DENSITY	Free flow	242	0.54	168	0.67
	Stable flow	169	0.38	36	0.14
	Unstable/Forced flow	28	0.06	47	0.19
INTERSECTIONSENTERED	Yes	68	0.15	47	0.19
	No	379	0.85	205	0.81
TRAFFIC CONTROL	Yes	148	0.33	98	0.39
	No	299	0.67	154	0.61
SURPRISED	Yes	249	0.56	171	0.68
	No	198	0.44	81	0.32
LOCALITY	Residential	145	0.32	37	0.15
	Business/industrial	207	0.46	131	0.52
	Others	95	0.21	84	0.33

Figure 2: Distribution of attributes.

The "mlxtend" package in Python 3.7 is used to implement the FP growth algorithm for frequent items and mine the association rules with a minimum support value of 0.05 and a minimum confidence value of 0.05 as hyperparameters.

3. Results

3.1. Frequency Analysis. The 20 most frequent items were selected to determine which features the C/NC events are associated with. As shown in Figure 3, the most frequent item is no driver impairment, and the second most frequent item is secondary tasks, indicating that most drivers are driving normally, and secondary tasks are highly associated with crash events. In addition, the most frequent items related to the road are a straight road, level road, and no controlled intersections. It can also be deduced from Figure 3 that the C/NC events are highly associated with driving normally and are associated with performing secondary tasks on straight and level road segments with no controlled intersections. These conditions are common in real life and have the highest probability of crashes.

Figures 4(a) and 4(b) show the frequency plots for crash events and near-crash events, respectively. Several differences are observed in these two plots: (1) the secondary task is the most frequent item contributing to crash events with a frequency of 94.85%, whereas this item ranks fourth for near-crash events with a frequency of 90.47%, indicating that secondary tasks are frequently associated with crash events. (2) The number of travel lanes less than or equal to 2 ranks eighth for crash events (frequency of 62.64%), and the number of travel lanes between 2 and 7 ranks seventh for near crashes, with a frequency of 71.43%, indicating that the probability of a crash is higher for fewer lanes. (3) Free flow ranks 12^{th} for crash events, with a frequency of 66.67%. This result suggests that a free traffic flow may keep the drivers over-confident, causing crashes. (4) Improper behavior ranks 13^{th} for crash events and is not correlated with near-crash events. Thus, improper behavior occurs more frequently in crash events. (5) An annual mileage of less than 10000 miles is associated with crash events, and an annual mileage greater than 15000 miles is more frequently associated with near-crash events, indicating that drivers with more driving experience are less likely to be involved in crashes.

3.2. Model Performance and Descriptive Statistics of the Parameters. We created two-key plots [30] to visualize the patterns extracted from the association rules of the C/NC events. There are 142794 rules for crash events and 18759 rules for near-crash events generated by the FP growth algorithm, with a minimum support value of 0.05 and a minimum confidence value of 0.05. Because there are numerous association rules, we randomly selected some to show the pattern. We merged the 3-item rules and 4-item rules as well as the 5-item rules and 6-item rules. In Figure 5, the range of support values for the 2-item rules is 0.05 to 0.6, and the confidence values of these rules exceed 0.4. For the 3-4-item rules, the range of support values is 0.05 to 0.5, and the confidence values also exceed 0.4. The 5-6-item rules have a similar trend, but the maximum value of support values is less than 0.25.

Figure 6 shows the two-key plots for the rules of the near-crash events. The range of the support values is 20% smaller, and the confidence value range for the majority of rules of the near-crash events is 80% lower than in Figure 5.

3.3. Obtaining the Patterns from the Association Rules of the C/NC Events

3.3.1. Crash Event Patterns. Table 3 presents the 25 top rules selected from 142,794 rules according to the lift value (from high to low) for crash events. The 6-item rule {trafficflow = not divided + travellanes = lanes ≤ 2 + NUMVIOL = 0 + gender = F + driverbehavior = improper behavior} is used as an example. A male person driving on an undivided road with less than 2 lanes is more likely to be involved in a crash when performing improper behavior, such as aggressive driving, even if he has no violations. The corresponding metrics are support = 0.053, confidence = 1, lift = 1.564, and conviction = inf. This can be interpreted as follows: the support value indicates that only 5.3% of crash events contain these five items. The confidence value indicates that if an event contains the five items, it is a crash event. The lift value shows that the percentage of crash events with these five items is 1.564 times higher than that of other crash events in the dataset. The conviction indicates the relationship between antecedents and consequents; the higher the conviction, the stronger the relationship is.

The rules for crash events are summarized from three aspects: (1) road: roadways with no lanes or undivided roads (rules 1, 6, 7, 8, 9, 10, and 21) or roads with less than two lanes (6, 16, 17, 20, 21, 22, 23, 24, and 25), and level roads (rule 7, 24) are more likely to be associated with crash events. (2) Driver: young (rule 3, 15) female (rule 21) participants with minor visual-spatial disabilities (rule 18) and an estimated average annual mileage over five years of less than 10,000 miles (rules 11, 13, 14, and 15) are more likely to be associated with crash events when performing secondary tasks (rule 23), improper behavior (rules 12, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, and 25) or impairments (rule 8) observed. Note that the number of traffic violations or being involved in a crash are not significantly correlated with crash events (rules 16, 17, 18, 21, 22, 23, 24, and 25). (3) Environment: crash events occur more frequently when the traffic density is free flow (rules 4, 11, 13, 19, and 20), there is no traffic control (rules 9, 19) or controlled intersections (rules 10, 25), and the area is residential (rules 5, 12, 13, 14, and 15) or business/industrial (rules 16, 22, 23, 24, and 25). Note that sudden unexpected events, such as breaking of a lead vehicle, animals, or pedestrians entering the roadway at a non-marked location or vehicle swerving in front of the driver, do not contribute significantly to crash events (rules 12, 17, 18, 19, and 20).

The likely reasons for these results are as follows. Undivided roads or roads with fewer than two lanes are typically low-grade roads. Young drivers have less driving experience and are more likely to underestimate the danger of driving on these road segments, especially when there are no

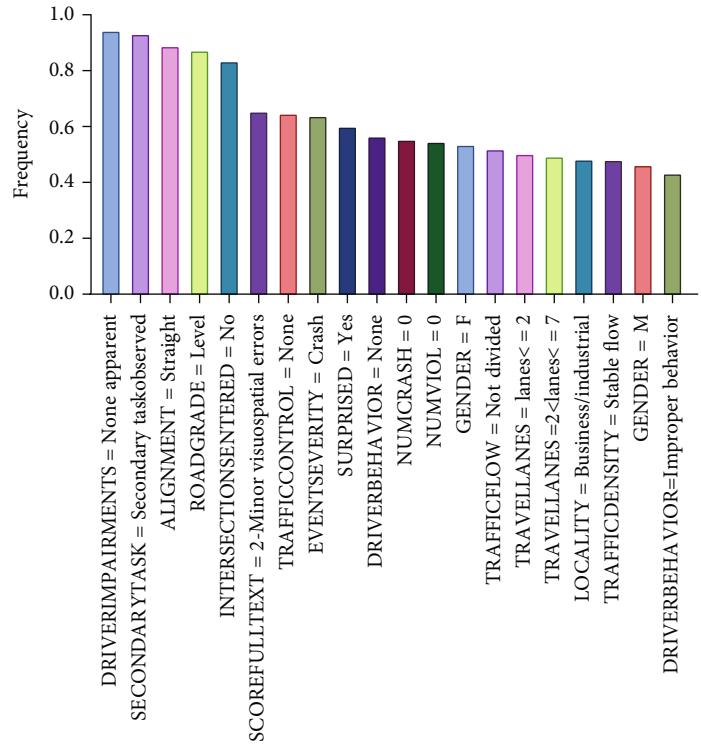

Figure 3: Item frequency in C/NC events.

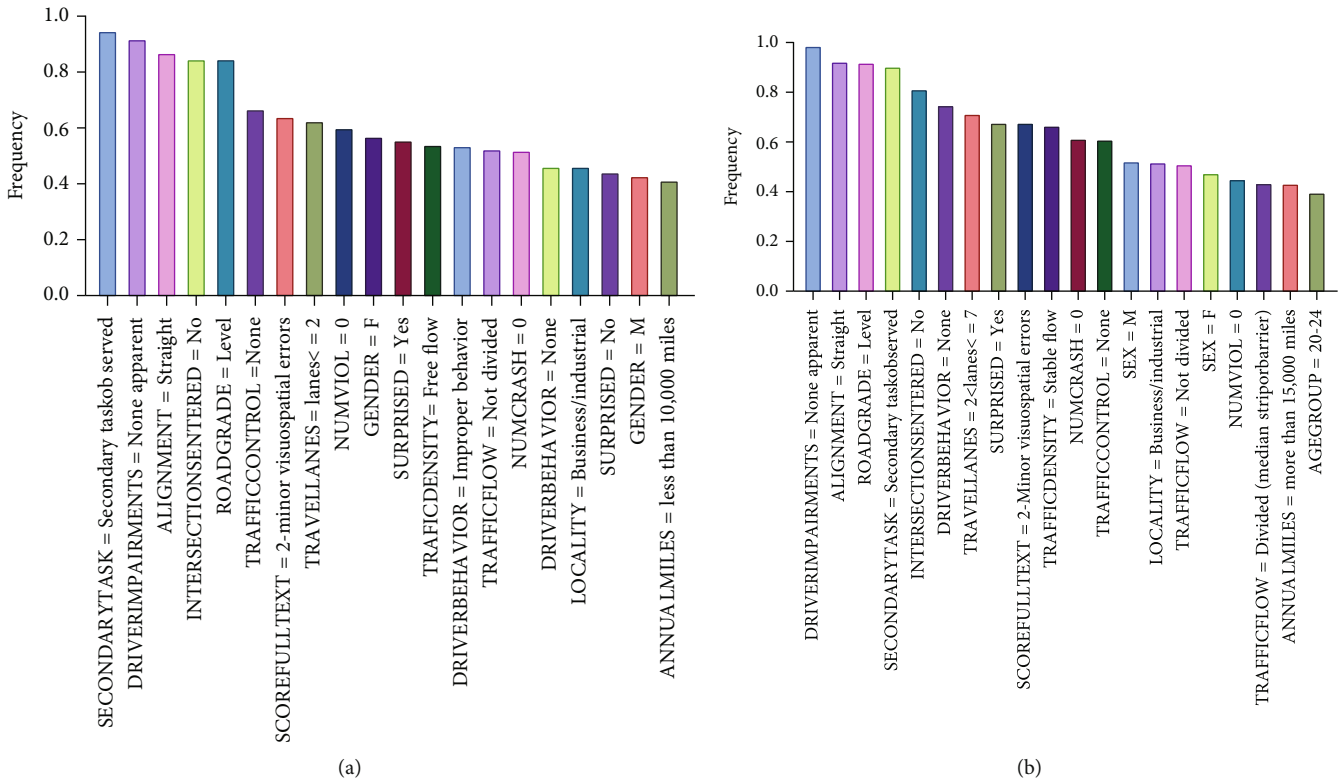

Figure 4: Item frequency in (a) crash events and (b) near-crash events.

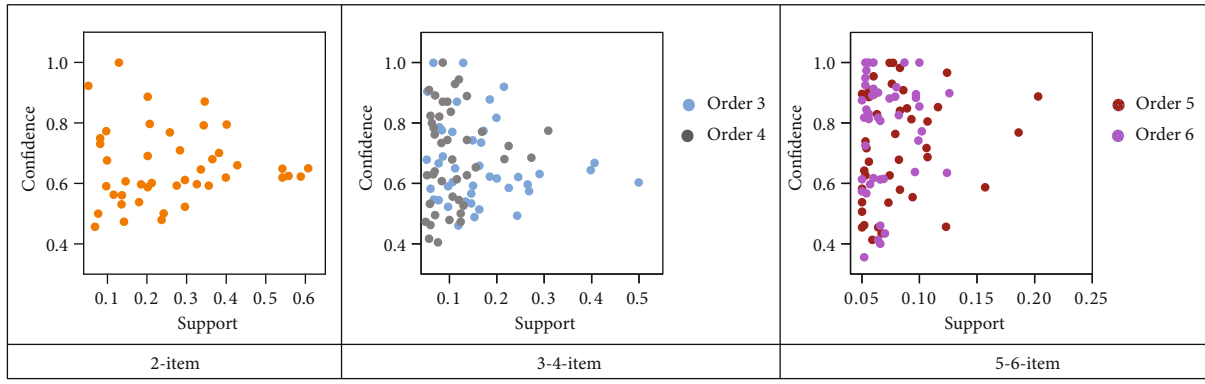

Figure 5: Two-key plots for crash events.

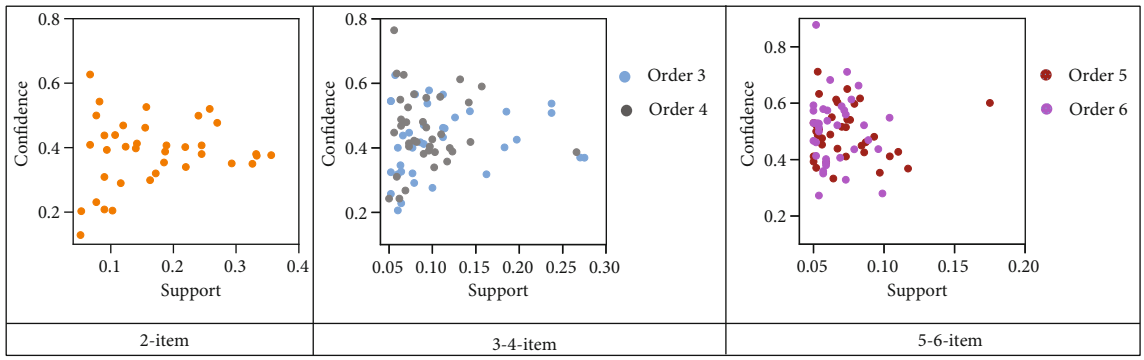

Figure 6: Two-key plots for near-crash events.

vehicles, traffic control, or intersections to interrupt driving. Under these conditions, drivers can be involved in crashes when they suffer from fatigue or perform secondary tasks or improper behavior.

3.3.2. Near-Crash Event Patterns. Table 4 presents the 25 top rules selected from 142,794 rules according to the lift value (from high to low) for near-crash events. The first 6-item rule {age group = 20 − 24 + locality = business/industrial + traffic density = stable flow + travel lanes = 2 < lanes ≤ 7 + secondary task = secondary task observed} is used as example. When a driver is affected by the interactions with others in traffic, the driver's speed is influenced. In addition, maneuvering in stable flow requires substantial vigilance by the driver, and the general comfort level declines. A young man driving on a wide road in a business/industrial area is more likely to be involved in a near-crash event when he is performing secondary tasks. The corresponding metrics are support = 0.05, confidence = 0.946, lift = 2.264, and conviction = 11.83. This can be interpreted as follows: the support value indicates that only 5% of near-crash events contain these five items. The confidence value shows that an event containing the five items has a 94.6% probability of being a near-crash event. The lift value demonstrates that the percentage of near-crash events with these five items is 2.264 times higher than that of other near-crash events in the dataset. The consequent depends significantly on the antecedent because the conviction value is higher (11.83) than the others.

The rules for near-crash events are summarized from three aspects: (1) road: level roads (rules 9, 19, 22, 23, and 24), divided roads (median strip or barrier) (rule 3), roads with 2 to 7 lanes (rules 4, 11, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, and 25), and straight roads (rules 10, 20) are more likely to be associated with near-crash events. (2) Driver: middle-aged and older (rules 2, 5, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, and 25) male (rule 13) participants with an estimated average annual mileage over five years of more than 15,000 miles (rules 6, 7, 8, 9, 10, 12, and 15) are more likely to be associated with near-crash events when they are performing secondary tasks (rules 8, 18, 21, and 25). Note that driver impairments (rules 7, 17, 23, 24, and 25), driver behavior (rules 12, 23), or unexpected events (rule 6) are not correlated with near-crash events. (3) Environment: near-crash events occur more frequently when the traffic flow is stable (rules 13, 14, 16, 21, and 22) or unstable/forced (rule 1), and the area is business/industrial (rules 11, 16, 17, 18, 19, 20, 21, 22, 24, and 25).

The likely reasons for these results are as follows. Divided roads and more lanes have fewer crashes. However, the high traffic density limits the drivers' freedom to maneuver, making them irritable in a stable or unstable/forced traffic flow. The drivers are inclined to overtake and accelerate frequently under these conditions and underestimate the danger, especially older drivers with higher confidence in their driving experience. If they perform secondary tasks and their attention is distracted, near-crash events are likely to occur.

Table 3: Selected rules for crash events.

No.	Antecedents	S	Conf	L	Conv
	2-item rules				
1	Traffic flow = no lanes	0.129	1.000	1.564	inf
2	Driver impairments = impairments appear	0.052	0.923	1.443	4.687
3	Age group = 16 − 19	0.202	0.887	1.387	3.185
4	Traffic density = free flow	0.346	0.871	1.361	2.784
5	Locality = residential	0.207	0.797	1.246	1.773
	3-item rules				
6	Traffic flow = no lanes + travel lanes = lanes ≤ 2	0.129	1.000	1.564	inf
7	Traffic flow = no lanes + road grade = level	0.123	1.000	1.564	inf
8	Traffic flow = no lanes + driver impairments = none apparent	0.120	1.000	1.564	inf
9	Traffic flow = no lanes + traffic control = none	0.119	1.000	1.564	inf
10	Traffic flow = no lanes + intersections entered = no	0.119	1.000	1.564	inf
	4-item rules				
11	NUMcrash = 1 + traffic density = free flow + annual miles = less than 10,000 miles	0.057	1.000	1.564	inf
12	Driver behavior = improper behavior + locality = residential + surprised = no	0.074	1.000	1.564	inf
13	Annual miles = less than 10,000 miles + traffic density = free flow + locality = residential	0.082	1.000	1.564	inf
14	Annual miles = less than 10,000 miles + driver behavior = improper behavior + locality = residential	0.064	1.000	1.564	inf
15	Annual miles = less than 10,000 miles + age group = 20 − 24 + locality = residential	0.052	1.000	1.564	inf
	5-item rules				
16	NUMVIOL = 0 + travel lanes = lanes ≤ 2 + driver behavior = improper behavior + locality = business/industrial	0.062	1.000	1.564	inf
17	Travel lanes = lanes ≤ 2 + driver behavior = improper behavior + surprised = no + NUMcrash = 0	0.062	1.000	1.564	inf
18	Driver behavior = improper behavior + surprised = no + score full text = 2 − minor visuospatial errors + NUMcrash = 0	0.056	1.000	1.564	inf
19	Driver behavior = improper behavior + traffic density = free flow + surprised = no + traffic control = none	0.092	1.000	1.564	inf
20	Travel lanes = lanes ≤ 2 + driver behavior = improper behavior + traffic density = free flow + surprised = no	0.090	1.000	1.564	inf
	6-item rules				
21	Traffic flow = not divided + travel lanes = lanes ≤ 2 + NUMVIOL = 0 + gender = F + driver behavior = improper behavior	0.053	1.000	1.564	inf
22	Travel lanes = lanes ≤ 2 + driver impairments = none apparent + NUMVIOL = 0 + locality = business/industrial + driver behavior = improper behavior	0.062	1.000	1.564	inf
23	Travel lanes = lanes ≤ 2 + NUMVIOL = 0 + locality = business/industrial + driver behavior = improper behavior + secondary task = secondary task observed	0.059	1.000	1.564	inf
24	Travel lanes = lanes ≤ 2 + NUMVIOL = 0 + road grade = level + locality = business/industrial + driver behavior = improper behavior	0.056	1.000	1.564	inf
25	Travel lanes = lanes ≤ 2 + NUMVIOL = 0 + intersections entered = no + locality = business/industrial + driver behavior = improper behavior	0.053	1.000	1.564	inf

Table 4: Selected rules for near-crash events.

No.	Antecedents	S	Conf	L	Conv
	2-item rules				
1	Traffic density = unstable/forced flow	0.067	0.627	1.738	1.713
2	Age group = 25 − 34	0.082	0.543	1.506	1.399
3	Traffic flow = divided (median strip or barrier)	0.157	0.526	1.460	1.350
4	Travel lanes = 2 < lanes ≤ 7	0.258	0.520	1.443	1.333
5	Age group = 35 − 64	0.077	0.500	1.387	1.279
	3-item rules				
6	Surprised = no + annual miles = more than 15,000 miles	0.064	0.429	1.189	1.119
7	Annual miles = more than 15,000 miles + driver impairments = none apparent	0.155	0.486	1.349	1.245
8	Secondary task = secondary task observed + annual miles = more than 15,000 miles	0.144	0.455	1.262	1.173
9	Road grade = level + annual miles = more than 15,000 miles	0.143	0.483	1.340	1.237
10	Alignment = straight + annual miles = more than 15,000 miles	0.143	0.476	1.321	1.221
	4-item rules				
11	Locality = business/industrial + age group = 20 − 24 + travel lanes = 2 < lanes ≤ 7	0.063	0.863	2.393	4.659
12	Driver behavior = none + age group = 35 − 64 + annual miles = more than 15,000 miles	0.056	0.813	2.254	3.411
13	Traffic density = stable flow + age group = 20 − 24 + gender = M	0.054	0.809	2.243	3.340
14	Traffic density = stable flow + age group = 20 − 24 + travel lanes = 2 < lanes ≤ 7	0.072	0.806	2.237	3.304
15	Age group = 35 − 64 + annual miles = more than 15,000 miles + surprised = yes	0.052	0.800	2.219	3.197
	5-item rules				
16	Travel lanes = 2 < lanes ≤ 7 + traffic density = stable flow + age group = 20 − 24 + locality = business/industrial	0.050	0.921	2.555	8.100
17	Travel lanes = 2 < lanes ≤ 7 + age group = 20 − 24 + driver impairments = none apparent + locality = business/industrial	0.059	0.891	2.472	5.883
18	Travel lanes = 2 < lanes ≤ 7 + age group = 20 − 24 + secondary task = secondary task observed + locality = business/industrial	0.063	0.880	2.441	5.329

Table 4: Continued.

No.	Antecedents	S	Conf	L	Conv
19	Travel lanes = 2 < lanes ≤ 7 + age group = 20 − 24 + road grade = level + locality = business/industrial	0.063	0.880	2.441	5.329
20	Travel lanes = 2 < lanes ≤ 7 + age group = 20 − 24 + alignment = straight + locality = business/industrial	0.062	0.860	2.385	4.568
	6-item rules				
21	Age group = 20 − 24 + locality = business/industrial + traffic density = stable flow + travel lanes = 2 < lanes ≤ 7 + secondary task = secondary task observed	0.050	0.946	2.624	11.830
22	Age group = 20 − 24 + road grade = level + locality = business/industrial + traffic density = stable flow + travel lanes = 2 < lanes ≤ 7	0.050	0.946	2.624	11.830
23	Driver behavior = none + driver impairments = none apparent + age group = 20 − 24 + road grade = level + travel lanes = 2 < lanes ≤ 7	0.067	0.922	2.556	8.153
24	Driver impairments = none apparent + age group = 20 − 24 + road grade = level + locality = business/industrial + travel lanes = 2 < lanes ≤ 7	0.059	0.911	2.527	7.194
25	Driver impairments = none apparent + age group = 20 − 24 + locality = business/industrial + travel lanes = 2 < lanes ≤ 7 + secondary task = secondary task observed	0.059	0.911	2.527	7.194

Table 5: Patterns of four types of crash events.

No.	Severity	Antecedents	S	Conf	L	Conv
		2-item rules				
1	I	NUMcrash = 1	0.053	0.181	1.321	1.054
2		Gender = M	0.073	0.157	1.146	1.024
3	II	Locality = residential	0.052	0.198	1.213	1.043
4		Age group = 20 − 24	0.064	0.188	1.150	1.030
5	III	Annual miles = more than 15,000 miles	0.070	0.209	1.162	1.037
6		Locality = others	0.053	0.207	1.147	1.033
7	IV	Annual miles = less than 10,000 miles	0.067	0.201	1.265	1.053
8		Traffic density = stable flow	0.096	0.199	1.252	1.050
		3-item rules				
9	I	NUMcrash = 1 + driver impairments = none apparent	0.052	0.185	1.344	1.058
10		Score full text = 2 − minor visuospatial errors + gender = M	0.052	0.182	1.324	1.054
11	II	NUMVIOL = 0 + gender = M	0.053	0.209	1.282	1.058
12		Travel lanes = lanes ≤ 2 + traffic density = free flow	0.060	0.200	1.226	1.046
13	III	Traffic control = yes + travel lanes = 2 < lanes ≤ 7	0.050	0.226	1.253	1.059
14		Annual miles = more than 15,000 miles + alignment = straight	0.066	0.219	1.215	1.050
15	IV	Locality = business/industrial + traffic flow = not divided	0.060	0.233	1.469	1.097
16		Traffic density = stable flow + NUMcrash = 0	0.064	0.228	1.438	1.090
		4-item rules				
17	I	Road grade = level + score full text = 2 − minor visuospatial errors + traffic density = stable flow	0.052	0.186	1.351	1.059
18		Score full text = 2 − minor visuospatial errors + traffic density = stable flow + driver impairments = none apparent	0.054	0.176	1.281	1.047
19	II	Travel lanes = lanes ≤ 2 + traffic density = free flow + intersections entered = no	0.056	0.210	1.286	1.059
20		NUMVIOL = 0 + driver impairments = none apparent + gender = M	0.052	0.207	1.269	1.055
21	III	Annual miles = more than 15,000 miles + road grade = level + alignment = straight	0.064	0.232	1.287	1.067
22		Annual miles = more than 15,000 miles + road grade = level + intersections entered = no	0.057	0.226	1.254	1.059
23	IV	Road grade = level + locality = business/industrial + traffic flow = not divided	0.059	0.244	1.537	1.113
24		Road grade = level + traffic flow = not divided + traffic control = none	0.059	0.241	1.519	1.109
		5-item rules				
25	I	Road grade = level + score full text = 2 − minor visuospatial errors + traffic density = stable flow + driver impairments = none apparent	0.050	0.183	1.334	1.056
26		Driver behavior = none + score full text = 2 − minor visuospatial errors + road grade = level + driver impairments = none apparent	0.056	0.176	1.279	1.047
27	II	Driver impairments = none apparent + travel lanes = lanes ≤ 2 + traffic density = free flow + intersections entered = no	0.050	0.211	1.293	1.061
28		Travel lanes = lanes ≤ 2 + traffic density = free flow + secondary task = secondary task observed + intersections entered = no	0.052	0.202	1.240	1.049
29	III	Annual miles = more than 15,000 miles + road grade = level + intersections entered = no + alignment = straight	0.057	0.240	1.329	1.078
30		Annual miles = more than 15,000 miles + road grade = level + secondary task = secondary task observed + alignment = straight	0.063	0.239	1.327	1.077
31	IV		0.056	0.248	1.564	1.119

Table 5: Continued.

No.	Severity	Antecedents	*S*	Conf	*L*	Conv
		Road grade = level + locality = business/industrial + traffic flow = not divided + alignment = straight				
32		Traffic control = none + road grade = level + traffic flow = not divided + driver impairments = none apparent	0.057	0.247	1.555	1.117
		6-item rules				
33	I	Score full text = 2 − minor visuospatial errors + alignment = straight + driver behavior = none + road grade = level + driver impairments = none apparent	0.050	0.173	1.262	1.043
34		Score full text = 2 − minor visuospatial errors + driver impairments = none apparent + intersections entered = no + road grade = level + traffic control = none	0.059	0.160	1.162	1.026
35	II	Secondary task = secondary task observed + driver impairments = none apparent + alignment = straight + road grade = level + driver behavior = improper behavior	0.054	0.179	1.099	1.020
36		Driver impairments = none apparent + alignment = straight + NUMVIOL = 0 + intersections entered = no + traffic control = none	0.053	0.179	1.096	1.019
37	III	Secondary task = secondary task observed + alignment = straight + annual miles = more than 15,000 miles + intersections entered = no + road grade = level	0.057	0.252	1.396	1.095
38		Secondary task = secondary task observed + alignment = straight + annual miles = more than 15,000 miles + road grade = level + driver impairments = none apparent	0.060	0.240	1.331	1.079
39	IV	Traffic flow = not divided + driver impairments = none apparent + intersections entered = no + road grade = level + traffic control = none	0.057	0.247	1.555	1.117
40		Traffic flow = not divided + alignment = straight + locality = business/industrial + road grade = level + driver impairments = none apparent	0.054	0.247	1.554	1.117

Table 6: Key findings.

Researches	Road	Driver	Environment
Our study	(1) Wider and median strip can reduce the frequency and severity of crashes (2) Only combined with other factors, level roadway and straight alignment are related to C/NC events	(1) The females are likely to be linked with lower-severe crashes, while males are likely to be linked with severe crashes and near crashes (2) Young age and less annual miles are more linked to crashes. However, the age does not show a strong correlation with the severity of C/NC events (3) Improper behavior and secondary tasks are correlated with crashes (4) Crash records and minor visual spatial disabilities are associated with the most severe events, while age, driver impairments and improper behaviors do not strongly correlate with the severity of crashes	(1) In free flow, crashes are more likely to occur (2) Traffic control and intersections are associated with C/NC events; this is more common in residential or business/industrial areas
Kong et al. [31]	(1) Small radius curves are linked with run-off-the-road (ROR) crashes (2) Large radius curves and favourable pavement conditions are associated with severe and fatal injury (KSI) crashes	(1) Female gender and young age are linked with ROR crashes, while male gender and older age are associated with KSI crashes (2) Driver impairments and improper behaviors are two main contributing factors of KSI crashes	(1) Bad weathers are linked with ROR crashes (2) Clear weathers are associated with KSI crashes
Kong et al. [30]	(1) Interstate highway or divided highway are highly associated with near crashes because of the overconfidence and secondary tasks	(1) When drivers perform secondary tasks, the main cause of near-crash events is the leading vehicle suddenly slowed or stopped. When not performing secondary tasks, lane-changing behavior is the main cause (2) When drivers perform secondary tasks, the most common evasive maneuver of avoiding the near crash is braked only. When not performing secondary tasks, the evasive maneuver is either steered or braked and steered	(1) Drivers are more concentrated in bad environmental conditions
Yu et al. [25]	(1) Crashes mostly occurred in urban areas with no physical separation (2) Crashes are more likely to occur on straight roads	(1) Male drivers are more prone to be associated with property damage than female drivers (2) Drivers aged 16–25 are most likely to be involved in crashes (3) Male drivers are more prone to fail to keep the vehicle under control	(1) Crashes are more likely to occur at an intersection
Hong et al. [27]	(1) Single-vehicle crashes are more likely induced by straight alignment	(1) Male and older drivers are highly linked to hazardous material vehicle involved crashes	(1) Dark conditions and poor visibility are two main contributing factors

3.3.3. Comparison of the C/NC Patterns. A comparison of the C/NC patterns is performed from three aspects: (1) road: divided roads, roads with no lanes, and the number of lanes are the main differences between the C/NC patterns. Crash events are more unlikely to occur on divided roads with more than 2 lanes. (2) Driver: the age group and annual miles are two significant factors in C/NC events. Drivers associated with crash events are predominantly 16-24-year-old teenagers with relatively little driving experience, whereas drivers involved in near-crash events are more likely older people (20-64 year old) with more driving experience. In addition, drivers are more likely to be associated with crash events when performing improper behaviors, such as aggressive driving and drunk driving, whereas secondary tasks are more influential in near-crash events. (3) Environment: crash events occur more likely in free flow, when the comfort level of drivers is high, in areas without traffic control or controlled intersections, and in residential or business/industrial areas. Near-crash events are more common in stable traffic flow or unstable/forced flow in business/industrial areas. The likely reason is that high traffic density keeps drivers alert, preventing crashes.

Near-crash events occur due to a combination of factors (i.e., traffic density levels, secondary tasks, and improper

driving behavior). Although near-crash events do not result in economic loss or casualties, some risk factors can turn near-crash events into crash events. Thus, it is necessary to discuss the relationship between crash and near-crash events and determine which conditions change near-crash events to crash events: (1) road: crash events are more likely to occur on narrow roads, whereas near-crash events are more likely to occur on wide roads. Thus, we assume near-crash events may change into crash events because of changes in the road features from urban to rural area roads or from main roads to bypasses. (2) Driver: older drivers are more likely to be involved in near-crash events rather than crash events; however, if they perform improper driving behavior, a near-crash event may become a crash event. (3) Environment: Bernat et al. [51] found that night-time single vehicle crashes (SVCs) were strongly related to drunk driving, and improper driving behavior was more likely when there were no vehicles nearby. Thus, improper driving behavior might increase the probability of turning near-crash events into crash events in free flow.

3.3.4. Patterns of Four Types of Crash Events. The association rules between different categorical variables and the severity of crash events are analyzed, and crash events are categorized into severity levels: I—most severe, II—police-reportable, III—minor crash, and IV—low-risk tire strike. Note that the definition of the four severity levels of crash events is derived from the NEST [39] dataset. Forty association rules are considered according to the lift value (Table 5).

Undivided roadways (rules 15, 23, 24, 31, 32, 39, and 40) are strongly associated with IV—low-risk tire strike events. However, this does not indicate that a low-risk tire strike causes severe crash events. Straight roads (rules 14, 21, 29, 30, 31, 33, 35, 36, 37, 38, and 40) are rarely associated with 2-item, 3-item, or 4-item rules but are more commonly with 5-item and 6-item rules. It is assumed that crashes rarely occur on straight road segments. However, crash events are more likely when a straight road is combined with other antecedents. Similar to the straight road segment, level road segments (rules 17, 21, 22, 23, 24, 25, 26, 29, 30, 31, 32, 33, 34, 35, 37, 38, 39, and 40) combined with other factors have an increased likelihood of crash events. Police-reportable events (II) are more likely on roads with less than two lanes (rules 12, 19, 27, and 28). Minor crash events (rule 13) (III) are more likely on roads with more than two lanes, indicating that widening the roadway can reduce the frequency and severity of crash events.

Male (rules 2, 10, 11, and 20) drivers are more likely to be associated with I—most severe events and II—police-reportable events. Drivers with one crash record during the past five years (rules 1, 9) are more likely to be associated with I—most severe events. The age group (rule 4) does not show a strong correlation with the crash severity. Drivers with annual miles greater than 15000 miles (rules 5, 14, 21, 22, 29, 30, 37, and 38) have a low correlation with severe crash events, indicating that drivers with more driving experience drive more safely. Minor visual-spatial disabilities do not show a strong correlation with crash events. However, they are strongly associated with I—most severe events. We speculate that minor visual-spatial disabilities do not affect driving significantly. However, if crash events are about to occur, the visual-spatial disabled drivers (rules 10, 17, 18, 25, 26, 33, and 34) may have more problems if a crash occurs. Thus, the crash events are typically more severe. Driver impairments (rules 9, 18, 20, 25, 26, 27, 32, 33, 34, 35, 36, 38, 39, and 40) and improper behavior (rules 26, and 33) are not strongly correlated with the severity of crash events, whereas performing secondary tasks (rules 28, 30, 35, 37, and 38) results in more frequent II—police-reportable crash events and III—minor crash events.

Driving in residential areas and other areas (rules 3, 6) is more likely associated with level II or III crash events. However, driving in business/industrial areas (rules 15, 23, 31, and 40) is more likely associated with IV—low-risk tire strike crash events. I—most severe events (rules 17, 18, and 25) and IV—low-risk tire strike events (rules 8, and 16) occur more likely when the traffic flow is stable. II—police-reportable crash events occur more likely in free flow (rules 12, 19, 27, and 28). Interruptions due to traffic control (rules 13, 24, 32, 34, 36, and 39) or controlled intersections (rules 19, 22, 27, 28, 29, 34, 36, 37, and 39) do not affect the severity of crash events.

4. Findings and Discussion

The key findings are summarized as follows:

(1) Road

(a) Undivided roadways are more likely associated with crash events, especially IV—low-risk tire strike events. In contrast, divided roadways are more likely associated with near-crash events. It is assumed that a median strip or barrier could prevent crashes

(b) Roads with less than 2 lanes are highly correlated with crash events, especially II—police-reportable events. Roads with 2-7 lanes are highly correlated with near-crash events or lower-severity crash events. Wider roadways are recommended to reduce the frequency and severity of crash events

(c) Crash events mainly occur on level roads, whereas near-crash events mainly occur on straight roads. However, this factor is only related to C/NC events in combination with other factors

(2) Driver

(a) Female drivers have a low correlation with low-severity crash events, whereas male drivers have a high correlation with severe crash and near-crash events

(b) Young drivers have a higher likelihood of being involved in crash events, whereas middle-aged

and older drivers show a stronger association with near-crash events. However, the driver's age is not highly correlated with the severity of crash events

(c) Crash events occur more likely when the drivers' estimated average annual mileage during the past five years is less than 10,000 miles. Near-crash events are more likely to occur when the drivers' average annual mileage during the past five years is greater than 15,000 miles. It is assumed that drivers with more driving experience have a safer driving style

(d) Performing secondary tasks is highly correlated with crash events (especially the II—police-reportable crash events and III—minor crash events) and near-crash events

(e) Improper behavior is linked to crash events, whereas driver impairments are not. Both factors are not strongly correlated with the severity of crash events

(f) The number of traffic violations or crash records is not strongly correlated to the frequency of C/NC events. However, drivers with one crash record during the past five years are more likely to be associated with I—most severe events

(g) Minor visual-spatial disabilities are not strongly correlated with crash events but are strongly correlated with I—most severe events. It is assumed that minor visual-spatial disabilities do not affect driving significantly. However, during a crash event, visual-spatial disabled drivers may have problems handling the situation; thus, the crash event is typically more severe

(3) Environment

(a) Crash events occur more likely in free flow traffic, and near-crash events are more likely in stable or unstable/forced flow. The results suggest that a higher traffic density keeps drivers alert, preventing crashes

(b) Crash events are more likely in sections with no traffic control or controlled intersections. However, these factors do not affect the severity of crash events

(c) Residential or business/industrial areas have a higher correlation with C/NC events than other areas. More traffic safety precautions should be considered in these areas

The key findings of a comparison of our results and three similar studies are summarized in Table 6.

We analyzed the associations between various factors and C/NC events and the crash severity. The following was observed: (1) road: Kong et al. [30] found associations between near-crash events and roads with median strips. Yu et al. [25] observed that most crashes occurred in urban areas on undivided roads. We also found that a median strip reduced the frequency and severity of crash events. Yu et al. [25] reported that crashes were more likely on straight road sections, similar to our study. However, we found that crashes were associated with straight road sections in combination with other factors. (2) Driver: similar to most other studies, we also found that gender, age, improper driving behavior, and secondary tasks were correlated with C/NC events. In contrast to other studies, we observed that only severe crashes were correlated with minor visual-spatial disabilities. Thus, we speculate that minor visual-spatial disabilities do not affect driving. However, in a serious crash, the visual-spatial disabled drivers may be more likely to lose control. (3) Environment: Kong et al. [30] found that drivers had shorter reaction times in inclement weather, and clear weather was associated with KSI crashes. Similarly, we observed that crash events occurred more likely in road sections without traffic control and intersections in residential or business/industrial areas, suggesting that accidents often occur under the most common road conditions.

5. Conclusions

This study investigated the correlations between C/NC events and driver, road, and environment-related categorical variables, such as secondary tasks, road conditions, and traffic density. We used the FP growth ARM algorithm to obtain new insights into C/NC events. The patterns of C/NC events were analyzed to determine which variables were associated with C/NC events. This paper provides two major contributions. First, we used a large dataset containing categorical variables collected from naturalistic driving studies, including driver, vehicle, and environment-related data. Therefore, it is believed that our results are robust and unbiased. Second, a framework was developed to mine the association rules of the C/NC events and crash events with different severities. In many cases, multiple variables were associated with C/NC events. We used the support, confidence, lift, and conviction metrics to measure the strength of association between the rules and outcomes.

Interesting correlations were observed between the categorical variables and C/NC events, and differences were revealed between crash and near-crash events. The top 5-item rules for crash events $\{\text{NUMVIOL} = 0 + \text{travel lanes} = \text{lanes} \leq 2 + \text{driver behavior} = \text{improper behavior} + \text{locality} = \text{business/industrial}\}$ and near-crash events $\{\text{travel lanes} = 2 < \text{lanes} \leq 7 + \text{traffic density} = \text{stable flow} + \text{age group} = 20 - 24 + \text{locality} = \text{business/industrial}\}$ are used as examples. In these two association rules, travel lanes and locality were significantly correlated with the occurrence of C/NC events. However, the correlation strength differed for different categorical variables. Drivers with an aggressive driving style were more likely to be involved in a crash when driving on roads with less than two lanes in a business/industrial area. Drivers driving in a business/industrial area on roads with

more than 2 lanes in stable traffic were more likely to be involved in near-crash events.

This study is expected to provide useful information for future research on C/NC events using ARM methods and suggestions for traffic engineers to improve road safety and prevent accidents. However, this study has three limitations. First, we did not include all rules in the analysis due to the large number of generated rules. Second, although we included a large range of categorical variables and extracted the association rules between the variables and C/NC events, we did not evaluate the correlations between the categorical variables. For example, many researchers have found that performing secondary tasks, such as using a phone or talking to passengers while driving, significantly increased driving risks. However, we aggregated all secondary tasks into one category. Third, some important categorical variables were discarded for the reasons described in Section 2.2, although they may have influenced the C/NC events. These limitations will be addressed in future studies.

Data Availability

The Naturalistic Engagement in Secondary Tasks (NEST) data used to support the findings of this study have been deposited in the SHRP2 Naturalistic Driving Study repository (doi:10.15787/VTT1/OZQ6BL).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

Thanks are due to SHRP2 Naturalistic Driving Study for collecting and providing the detailed dataset.

References

[1] https://www.nhtsa.gov/.

[2] G. Khan, X. Qin, and D. A. Noyce, "Spatial analysis of weather crash patterns," *Journal of Transportation Engineering*, vol. 134, no. 5, pp. 191–202, 2008.

[3] P. Wu, X. Meng, L. Song, and W. Zuo, "Crash risk evaluation and crash severity pattern analysis for different types of urban junctions: fault tree analysis and association rules approaches," *Transportation Research Record*, vol. 2673, no. 1, pp. 403–416, 2019.

[4] X. Wang, X. Wu, M. Abdel-Aty, and P. J. Tremont, "Investigation of road network features and safety performance," *Accident; Analysis and Prevention*, vol. 56, pp. 22–31, 2013.

[5] A. Bélanger, S. Gagnon, and A. Stinchcombe, "Crash avoidance in response to challenging driving events: the roles of age, serialization, and driving simulator platform," *Accident; Analysis and Prevention*, vol. 82, pp. 199–212, 2015.

[6] J. Bärgman, V. Lisovskaja, T. Victor, C. Flannagan, and M. Dozza, "How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from SHRP2," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 35, pp. 152–169, 2015.

[7] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor, "Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain," *Accident; Analysis and Prevention*, vol. 163, article 106454, 2021.

[8] A. M. Noble, M. Miles, M. A. Perez, F. Guo, and S. G. Klauer, "Evaluating driver eye glance behavior and secondary task engagement while using driving automation systems," *Accident; Analysis and Prevention*, vol. 151, article 105959, 2021.

[9] Y. Forster, V. Geisel, S. Hergeth, F. Naujoks, and A. Keinath, "Engagement in non-driving related tasks as a non-intrusive measure for mode awareness: a simulator study," *Information*, vol. 11, no. 5, p. 239, 2020.

[10] T. Morgenstern, E. M. Wögerbauer, F. Naujoks, J. F. Krems, and A. Keinath, "Measuring driver distraction - evaluation of the box task method as a tool for assessing in-vehicle system demand," *Applied Ergonomics*, vol. 88, article 103181, 2020.

[11] A.-C. Hensch, N. Rauh, C. Schmidt et al., "Effects of secondary tasks and display position on glance behavior during partially automated driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 68, pp. 23–32, 2020.

[12] D. Onate-Vega, O. Oviedo-Trespalacios, and M. J. King, "How drivers adapt their behaviour to changes in task complexity: the role of secondary task demands and road environment factors," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 71, pp. 145–156, 2020.

[13] P. Bakhit, B. Guo, and S. Ishak, "Crash and near-crash risk assessment of distracted driving and engagement in secondary tasks: a naturalistic driving study," *Transportation Research Record*, vol. 2672, no. 38, pp. 245–254, 2018.

[14] M. Bakhtiyari, A. Delpisheh, A. B. Monfared et al., "The road traffic crashes as a neglected public health concern; an observational study from Iranian population," *Traffic Injury Prevention*, vol. 16, no. 1, pp. 36–41, 2015.

[15] M. Haque, S. Washington, and B. Watson, "A methodology for estimating exposure-controlled crash risk using traffic police Crash Data," *Procedia-Social and Behavioral Sciences*, vol. 104, pp. 972–981, 2013.

[16] N. Sze, S. Wong, X. Pei, P. Choi, and Y. Lo, "Effective measures in combating red light violation: an observational study in Hong Kong," in *International Conference of Hong Kong Society for Transportation Studies*, p. 65, Hong Kong Society for Transportation Studies (HKSTS), 2010.

[17] J. M. Hankey, M. A. Perez, and J. A. McClafferty, *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets*, Virginia Tech Transportation Institute, 2016.

[18] E. Papazikou, M. Quddus, P. Thomas, and D. Kidd, "What came before the crash? An investigation through SHRP2 NDS data," *Safety Science*, vol. 119, pp. 150–161, 2019.

[19] M. Kreusslein, T. Morgenstern, T. Petzoldt, A. Keinath, and J. F. Krems, "Characterising mobile phone calls while driving on limited-access roads based on SHRP 2 naturalistic driving data," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 70, pp. 208–222, 2020.

[20] C. J. R. Schlick, D. B. Hewitt, C. M. Quinn et al., "A national survey of motor vehicle crashes among general surgery residents," *Annals of Surgery*, vol. 274, no. 6, pp. 1001–1008, 2021.

[21] Y. Zou, L. Ding, H. Zhang, T. Zhu, and L. Wu, "Vehicle acceleration prediction based on machine learning models and

driving behavior analysis," *Applied Sciences*, vol. 12, no. 10, p. 5259, 2022.

[22] M. Guo, X. Zhao, Y. Yao, C. Bi, and Y. Su, "Application of risky driving behavior in crash detection and analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 591, article 126808, 2022.

[23] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, pp. 487–499, 1994.

[24] R. Srikant and R. Agrawal, *Mining generalized association rules*, IBM Research Division Zurich, 1995.

[25] S. Yu, Y. Jia, and D. Sun, "Identifying factors that influence the patterns of road crashes using association rules: a case study from Wisconsin, United States," *Sustainability*, vol. 11, no. 7, 2019.

[26] J. Hong, R. Tamakloe, and D. Park, "Discovering insightful rules among truck crash characteristics using Apriori algorithm," *Journal of Advanced Transportation*, vol. 2020, Article ID 4323816, 16 pages, 2020.

[27] J. Hong, R. Tamakloe, and D. Park, "Application of association rules mining algorithm for hazardous materials transportation crashes on expressway," *Accident; Analysis and Prevention*, vol. 142, article 105497, 2020.

[28] S. Das, X. Kong, and I. Tsapakis, "Hit and run crash analysis using association rules mining," *Journal of Transportation Safety & Security*, vol. 13, no. 2, pp. 123–142, 2021.

[29] A. Montella, F. Mauriello, M. Pernetti, and M. Rella Riccardi, "Rule discovery to identify patterns contributing to overrepresentation and severity of run-off-the-road crashes," *Accident; Analysis and Prevention*, vol. 155, p. 106119, 2021.

[30] X. Kong, S. Das, and Y. Zhang, "Mining patterns of near-crash events with and without secondary tasks," *Accident; Analysis and Prevention*, vol. 157, article 106162, 2021.

[31] X. Kong, S. Das, and Y. Zhang, "Patterns of near-crash events in a naturalistic driving dataset: applying rules mining," *Accident; Analysis and Prevention*, vol. 161, article 106346, 2021.

[32] R. Tamakloe, S. Das, E. Nimako Aidoo, and D. Park, "Factors affecting motorcycle crash casualty severity at signalized and non- signalized intersections in Ghana: insights from a data mining and binary logit regression approach," *Accident; Analysis and Prevention*, vol. 165, p. 106517, 2022.

[33] S. Das and X. Sun, *Investigating the pattern of traffic crashes under rainy weather by association rules in data mining*, Transportation Research Board 93rd Annual Meeting, 2014.

[34] S. Kumar and D. Toshniwal, "Analysing road accident data using association rule mining," in *2015 International Conference on Computing, Communication and Security (ICCCS)*, pp. 1–6, Pointe aux Piments, Mauritius, 2015, January.

[35] C. Xu, J. Bao, C. Wang, and P. Liu, "Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China," *Journal of Safety Research*, vol. 67, pp. 65–75, 2018.

[36] S. Das, A. Dutta, R. Avelar, K. Dixon, X. Sun, and M. Jalayer, "Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures," *International Journal of Urban Sciences*, vol. 23, no. 1, pp. 30–48, 2019.

[37] S. Das, R. Tamakloe, H. Zubaidi, I. Obaid, and A. Alnedawi, "Fatal pedestrian crashes at intersections: trend mining using association rules," *Accident; Analysis and Prevention*, vol. 160, article 106306, 2021.

[38] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1–12, 2000.

[39] J. Owens, L. Angell, J. M. Hankey, J. Foley, and K. Ebe, "Creation of the Naturalistic Engagement in Secondary Tasks (NEST) distracted driving dataset," *Journal of safety research*, vol. 54, pp. 33. e29–36, 2015.

[40] M. Risteska, D. Kanaan, B. Donmez, and H.-Y. W. Chen, "The effect of driving demands on distraction engagement and glance behaviors: results from naturalistic data," *Safety Science*, vol. 136, article 105123, 2021.

[41] G. Prati, L. Pietrantoni, and F. Fraboni, "Using data mining techniques to predict the severity of bicycle crashes," *Accident; Analysis and Prevention*, vol. 101, pp. 44–54, 2017.

[42] Q. Zeng and H. Huang, "A stable and optimized neural network model for crash injury severity prediction," *Accident; Analysis and Prevention*, vol. 73, pp. 351–358, 2014.

[43] B. J. Russo and P. T. Savolainen, "A comparison of freeway median crash frequency, severity, and barrier strike outcomes by median barrier type," *Accident; Analysis and Prevention*, vol. 117, pp. 216–224, 2018.

[44] F. Chang, P. Xu, H. Zhou, A. H. Chan, and H. Huang, "Investigating injury severities of motorcycle riders: a two-step method integrating latent class cluster analysis and random parameters logit model," *Accident; Analysis and Prevention*, vol. 131, pp. 316–326, 2019.

[45] M.-M. Chen and M.-C. Chen, "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest," *Information*, vol. 11, no. 5, p. 270, 2020.

[46] C. Chen, G. Zhang, Z. Qian, R. A. Tarefder, and Z. Tian, "Investigating driver injury severity patterns in rollover crashes using support vector machine models," *Accident; Analysis and Prevention*, vol. 90, pp. 128–139, 2016.

[47] J. Alwidian, B. H. Hammo, and N. Obeid, "WCBA: weighted classification based on association rules algorithm for breast cancer disease," *Applied Soft Computing*, vol. 62, pp. 536–549, 2018.

[48] J. Alwidian, B. Hammo, and N. Obeid, "FCBA: fast classification based on association rules algorithm," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 16, no. 12, p. 117, 2016.

[49] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, "MAFIA: a maximal frequent itemset algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1490–1504, 2005.

[50] P. Harrington, *Machine Learning in Action*, Simon and Schuster, 2012.

[51] D. H. Bernat, W. T. Dunsmuir, and A. C. Wagenaar, "Effects of lowering the legal BAC to 0.08 on single-vehicle-nighttime fatal traffic crashes in 19 jurisdictions," *Accident; Analysis and Prevention*, vol. 36, no. 6, pp. 1089–1097, 2004.