Understanding Crash Risk Using a Multi-Level Random Parameter Binary Logit Model: Application to Naturalistic Driving Study Data

Lauren Hoover¹, Tanmoy Bhowmik¹, Shamsunnahar Yasmin², and Naveen Eluru¹

Abstract
This study presents a framework to employ naturalistic driving study (NDS) data to understand and predict crash risk at a disaggregate trip level accommodating for the influence of trip characteristics (such as trip distance, trip proportion by speed limit, trip proportion on urban/rural facilities) in addition to the traditional crash factors. Recognizing the rarity of crash occurrence in NDS data, the research employs a matched case-control approach for preparing the estimation sample. The study also conducts an extensive comparison of different case-to-control ratios including 1:4, 1:9, 1:14, 1:19, and 1:29. The model parameters estimated with these control ratios are reasonably similar (except for the constant). Employing the 1:9 sample, a multi-level random parameters binary logit model is estimated where multiple forms of unobserved variables are tested including (a) common unobserved effects for each case-control panel, (b) common unobserved factors affecting the error margin in the trip distance variable, and (c) random effects for all independent variables. The estimated model is calibrated by modifying the constant parameter to generate a population conforming crash risk model. The calibrated model is employed to predict crash risk of trips not considered in model estimation. This study is a proof of concept that NDS data can be used to predict trip-level crash risk and can be used by future researchers to develop crash risk models.

Keywords
safety, crash analysis, crash data, crash prediction models

Given the significant emotional, economic, and social costs of traffic crashes, “Vision Zero,” a movement in which communities set a goal to eliminate traffic fatalities and severe injuries within a specified timeframe, has been conceptualized (1). Several urban regions—including Orlando, Tampa, New York City, Chicago, Austin, Denver, and Los Angeles—have committed to meeting the goals of the Vision Zero movement (1). A major component of achieving Vision Zero goals includes developing statistical and econometric models to understand the underlying causes of crashes and to identify strategies for crash prevention and crash consequence mitigation.

Traditional safety research can be broadly classified along two directions—crash frequency and severity analysis. The first direction of research focuses on understanding the factors contributing to the number of crashes on a facility type in a specific time-period (2–4). The second direction of research examines factors affecting crash consequence (usual injury severity) conditional on the occurrence of a crash (5–7). The evolution of the safety field along these two primary research directions is based on how crash data is typically recorded—compiled by police or medical professionals. Traditional crash data has been instrumental in understanding the influence of various factors drawn from driver
demographics, vehicle characteristics, roadway characteristics, crash characteristics, and environmental factors on crash frequency and severity. However, the data does not allow us to examine the underlying cause of a crash. Crash-frequency models simply aggregate the crashes on a facility and are useful to examine the role of roadway environment in affecting crashes. On the other hand, the crash severity models focus on the crash consequence without having any information on the trip that resulted in the crash. As previously stated, this limitation is mainly a consequence of the absence of such detailed trip data.

The paradigm of crash data collection, however, can potentially undergo a significant change with the advent of naturalistic driving studies (NDS). Naturalistic driving data is obtained from drivers willing to participate in a data collection exercise through a host of sensors that are placed in vehicles recording driver behavior (such as on-task behavior, eye movement) and their actions (such as speed, acceleration) in real time. The first large-scale NDS was conducted in the Northern Virginia and Washington DC area monitoring 100 cars for about a year (8). More recently, another NDS titled the Second Strategic Highway Research Program (SHRP2) was conducted, with over 3,500 participants from six data collection sites across the United States, recording 1,951 crashes and 6,956 near-crashes (9). The ability to record trips involving crashes alongside those that do not include crashes allows researchers to compare driver behaviors and environmental factors in crash and non-crash trips and identify those factors that are more frequent in crash trips. In this study a trip starts when the car is turned on and ends when the car turns off. The NDS data allows for understanding the underlying timeline of the crash and accounts for driver behavior (as opposed to simply focusing on driver demographics). Thus, using NDS data, in theory, analysts can understand crash occurrence (yes/no at a trip level) and crash consequence (for trips involved in a crash) as a disaggregate event.

In this context, the current study makes two important contributions to safety literature. First, we present a framework to employ NDS data to understand and predict crash risk at a disaggregate trip level accommodating for the influence of trip characteristics (such as trip distance, trip proportion by speed limit, trip proportion on urban/rural facilities) in addition to the traditional crash factors. Second, we employ a rigorous case-control study design for understanding trip-level crash risk. NDS data collection is not primarily geared toward understanding potential crash occurrence and severity. Given the rarity of crashes, even such an exhaustive exercise as SHRP2 produced only 1,951 crash events from 5,512,900 trips (10). Therefore, trips with crashes represent only a small sample of the trips database. A binary outcome model of crash risk—whether a trip will result in a crash or not—will be extremely challenging to estimate with the small sample share. The sample share challenge observed in the trip-level crash risk has been documented in transportation safety literature in the context of crash/near-crash events in NDS (See Guo [11] for a detailed review) and real-time crash risk models developed in safety literature (12, 13). The current research will draw on earlier case-control literature in transportation safety to customize the case-control study design for our analysis.

**Earlier Research**

Our review of earlier research focused on two dimensions: (1) studies employing naturalistic driving data to draw insights on factors affecting crash occurrence and (2) research methods employed for analysis.

Several studies have employed naturalistic data for safety analysis. The most commonly employed NDS data sets include the 100-Car NDS (14, 15) or the SHRP2 NDS (16–18). The dimensions affecting crash/near-crash risk examined in these NDS studies include various driver behaviors such as driver inattention (14, 16), glance behavior (19), aggressive/risky driving and speeding (15, 20–22), and secondary task involvement (18, 23). Studies using NDS data have also examined crash/near-crash risk based on driver characteristics such as age (22, 24) and history of sleep disorders (25). Studies have also considered non-driver-related factors such as lighting conditions (23), pavement surface condition (23), and vehicle kinematics (26). Apart from the two major NDS studies, a small number of studies examined the role of driver actions in crash/near-crash events for commercial drivers (27), and influence of behavioral and environmental factors present before a crash for teenage drivers (28).

Analysis of NDS data is conducted using two main types of case-control study designs: (a) case-cohort design and (b) case-crossover design (11). In the case-cohort design, control periods are randomly selected for each driver proportional to their driving time or mileage. In the case-crossover design, controls for an event are selected using the same subject to account for subject-specific confounding factors. The analysis framework for crash/near-crash event is the logistic regression model. However, to accommodate for the unobserved factors associated with the same driver or other common elements, multi-level random parameter logit regression approaches are employed. An important element of discussion in case-control study design is the ratio of cases and controls. Mittleman et al. (29) suggested a 1:4 ratio.
for case-crossover studies. Most of the existing literature in safety employs a ratio ranging from 1:1 to 1:10. However, it is important that an examination of a stable ratio of cases and controls is conducted for each empirical context. Furthermore, even if the parameters are unbiased, model estimates from case-control studies cannot be used to calculate risk directly without employing corrections for the constant (see Zhang and Kai [30] for a detailed discussion). The case-control model outputs can only be used to calculate the odds ratio (31). The application of case-control model outputs is limited without the constant correction. In summary, the current study develops a case-cohort study design for trip-level crash risk analysis. We will rigorously examine the impact of control group sample size on the variable parameters and identify an appropriate case-to-control ratio for our analysis. The proposed model for the estimation will also accommodate for the presence of any unobserved factors in trip-level crash risk. It is possible that all the control group records matched with the case might have some common unobserved factors influencing crash risk. To accommodate for this potential unobserved heterogeneity, a multi-level random parameters binary logit model structure is employed in our analysis. The estimated model system is used to generate crash risk for a hold-out sample of data records by correcting the estimated case-cohort model for the general trip population.

**Data Preparation**

The data for our analysis is drawn from the SHRP2 NDS data. The data provided information on 1,951 trips that resulted in a crash and a random sample of 1,000,000 trips with no crash (from the full sample of 5.5 million trips). The data included trip data (such as start and end time, day of week, facility types and speeds, and max acceleration and deceleration), driver demographics (such as age, gender, education, income, and average annual mileage), crash event details (such as location details, collision type, crash severity, driver impairments, and weather). The list of variables examined in our study is summarized in Table 1. Several variables, such as total travel time, departure time of the trip, and the day of the week, were excluded from consideration because of many missing data points for those variables. Among the 1,951 trips resulting in a crash, 814 of those crashes were categorized as “low risk tire strike” and were excluded from the analysis, leaving 1,137 crashes to be analyzed. After further filtering the data, removing trips that had missing driver or trip information, we ended up with 928 trips resulting in a crash and 714,579 trips with no crash.

**Case-Control Design**

In case-control studies, *case* outcomes of interest (trips with a crash) are matched with a select number of *control* outcomes (trips without a crash). In our study we adopted the matched case-control approach. We selected the independent variables driver age, driver gender, and trip distance within a 20% margin for our matching exercise. With these criteria, we did not find enough controls for a small sample of crash trips. Therefore, we restricted our analyses to 914 crash trips (cases). For testing different case-to-control ratios, we created samples with the following case-to-control ratios 1:4, 1:9, 1:14, 1:19 and 1:29.

**Empirical Analysis**

**Parameter Variation across Various Samples**

The first part of our model development exercise was focused on parameter variability across the various samples. The binary logistic model was estimated for the largest sample, testing several variable specifications based on the variables described in the data preparation section. After a final specification was obtained for the 1:29 sample, the specification was estimated across all other samples. The final specification of the model was based on removing the statistically insignificant variables in a systematic manner based on the 90% confidence level. A summary of the model estimates across all control samples is presented in Table 2. A cursory examination of the parameters indicates reasonable agreement across all samples. The reader would note that the constant parameter across all models varies substantially. The variation across the constant parameter reflects the case-to-control sample share in the sample. Therefore, as the case-to-control ratio reduces, a reduction in the magnitude of the constant parameter is observed. While this is quite encouraging, the visual comparison does not indicate whether the difference across parameters for all the samples is within statistically acceptable levels.

To compare the parameters across the models, we employed the 1:29 control sample as the benchmark and evaluated whether the parameters for other models were statistically different relative to this sample. Toward making the comparison, a revised Wald test statistic relative to the 1:29 sample was generated as follows:

\[
\text{Parameter test statistic} = \frac{\text{(sample parameter} - \text{population benchmark)}}{\sqrt{\text{SE}_{\text{sample}}^2 + \text{SE}_{\text{population}}^2}}
\]

where SE denotes the standard error for the corresponding sample.
### Table 1. Summary of SHRP2 NDS Variables

#### Categorical variables

| Variable name               | Variable description                               | Share of category |
|-----------------------------|---------------------------------------------------|-------------------|
| Age 16–19                   | Driver age is between 16 and 19                   | 0.023             |
| Age 20–24                   | Driver age is between 20 and 24                   | 0.064             |
| Age 25–29                   | Driver age is between 25 and 29                   | 0.081             |
| Age 30–74                   | Driver age is between 30 and 74                   | 0.758             |
| Age >74                     | Driver age is greater than 74                     | 0.074             |
| Avg. annual miles <10,000   | Driver average annual mileage of less than 10,000 mi/yr | 0.229             |
| Avg. annual miles 10,000 to 25,000 | Driver average annual mileage between 10,000 and 25,000 mi/yr | 0.637             |
| Avg. annual miles >25,000   | Driver average annual mileage of greater than 25,000 mi/yr | 0.134             |
| Full-time worker            | Driver is full-time worker                         | 0.480             |
| Part-time worker            | Driver is part-time worker                         | 0.190             |
| Not working outside the home| Driver does not work outside the home              | 0.330             |
| Male                        | Driver is male                                     | 0.490             |
| Female                      | Driver is female                                   | 0.510             |
| Previous crash              | Driver has been in a crash in the last 3 years     | 0.260             |
| No previous crash           | Driver has not been in a crash in the last 3 years | 0.740             |

#### Continuous variables

| Variable name               | Variable description                               | Min. | Max. | Mean | SD   |
|-----------------------------|---------------------------------------------------|------|------|------|------|
| Years driving               | Number of years driver has been driving           | 0    | 74   | 33.132 | 17.732|
| Distance                    | Straight-line distance in miles between the start point and the end point of the trip | 0    | 577.135 | 7.531 | 14.869|
| Percent rural               | Percentage of the trip on rural roads             | 0    | 1    | 0.105 | 0.196|
| Percent urban               | Percentage of the trip on urban roads             | 0    | 1    | 0.550 | 0.285|
| Percent <30 mph             | Percentage of the trip where the speed was <30 mph | 0    | 1    | 0.388 | 0.313|
| Percent >70 mph             | Percentage of the trip where the speed was >70 mph | 0    | 1    | 0.018 | 0.089|
| Mean MPH                    | Mean speed of the vehicle in mph over the full trip | 0    | 88.487 | 28.630 | 12.276|
| Max MPH                     | Maximum speed of the vehicle in mph               | 0    | 93.206 | 46.879 | 17.558|
| Max acceleration            | Maximum longitudinal acceleration during the trip  | -1.367 | 3.210 | 0.287 | 0.096|
| Max deceleration            | Maximum longitudinal deceleration during the trip | -3.466 | 0.620 | -0.325 | 0.111|
| Max lateral accel           | Maximum lateral acceleration during the trip      | -0.238 | 3.483 | 0.381 | 0.131|
| Max turn rate               | Maximum turn rate during the trip                 | 344.057 | 399.990 | 26.673 | 10.216|

Note: SHRP2 = Second Strategic Highway Research Program; NDS = naturalistic driving study; avg. = average; mi/yr = miles per year; min. = minimum; max. = maximum; SD = standard deviation; accel = acceleration.

### Table 2. Crash Risk Estimates

| Parameters                             | 1:4 ratio | 1:9 ratio | 1:14 ratio | 1:19 ratio | 1:29 ratio |
|----------------------------------------|-----------|-----------|------------|------------|------------|
| Constant                               | -1.589 (0.174) | -2.390 (0.164) | -2.816 (0.160) | -3.144 (0.159) | -3.533 (0.152) |
| Trip variables                         |           |           |            |            |            |
| % Trip <30 mph                         | 0.383 (0.191) | 0.352* (0.180) | 0.3414* (0.176) | 0.363 (0.176) | 0.429 (0.167) |
| % Trip >70 mph                         | -0.792 (0.375) | -0.621* (0.348) | -0.606* (0.337) | -0.698 (0.336) | -0.004** (0.004) |
| Ln(Distance + 1)                       | 0.170 (0.057) | 0.144 (0.053) | 0.149 (0.052) | 0.153 (0.052) | 0.103 (0.049) |
| % Trip on urban roads                  | -0.54 (0.14) | -0.51 (0.13) | -0.54 (0.13) | -0.53 (0.13) | -0.48 (0.12) |
| Driver demographics                    |           |           |            |            |            |
| Drives <10,000 mi/yr                   | 0.384 (0.081) | 0.384 (0.076) | 0.398 (0.075) | 0.398 (0.074) | 0.386 (0.073) |
| Drives >25,000 mi/yr                   | 0.362 (0.121) | 0.388 (0.114) | 0.364 (0.111) | 0.372 (0.110) | 0.326 (0.109) |
| Full-time worker                       | -0.257 (0.082) | -0.178 (0.078) | -0.204 (0.076) | -0.196 (0.076) | -0.199 (0.075) |

Note: mi/yr = miles per year.
*Variable insignificant at 95% significance level; **Variable insignificant at 90% significance level.
If the parameter test statistic computed was higher than the 90% \( t \)-statistic, the result would indicate significant difference across the parameters. Employing the above test statistic computation, revised \( t \)-statistics for all the parameters across all sample were computed. Figure 1 provides a box plot summary of the variations across samples for all parameters. The figure clearly highlights that the range of the test statistic across all the parameters is quite narrow and exceeds the 90% significance only for one parameter. The parameter for “percentage of the trip at speeds greater than 70 mph” presents a range higher than the 90% confidence value of 1.65. This is not surprising given the variable was only marginally significant in the 1:29 control sample. We still retained the variable as it was intuitive. Given the stability across all samples, we selected the 1:9 control sample for further analysis and discussion.

**Methodological Framework**

Employing the 1:9 sample, a multi-level random parameters binary logit model was estimated. A brief mathematical description of the multi-level random parameters model follows:

Let \( q \) (\( q = 1, 2, 3, \ldots, m; M = 10 \)) represent the index for different samples for each stratum \( i \) (each case-control panel of 10 records). With this notation, the formulation takes the following familiar form:

\[
\nu_{iq}^* = \{ (\alpha + \gamma_{iq})z_{iq} + v_{iq} + \Omega_{iq} \}^+, \nu_{iq} = 1, \text{ if } \nu_{iq}^* > 0; \\
\nu_{iq} = 0, \text{ otherwise} \tag{1}
\]

where

\( \nu_{iq}^* \) represents the propensity for crash occurrence for sample \( q \) in stratum \( i \);

\( \nu_{iq}^* \) is 1 if the sample specific to a given stratum indicates a crash and 0 otherwise;

\( z_{iq} \) is a vector attribute associated with sample \( q \) in stratum \( i \) and \( \alpha \) is the vector of corresponding mean effects;

\( \gamma_{iq} \) is a vector of unobserved factors affecting probability of crash occurrence;

\( v_{iq} \) is an idiosyncratic error term assumed to be identically and independently standard logistic distributed; and

\( \Omega_{iq} \) is a vector of unobserved effects specific to stratum \( i \).

As highlighted earlier, within each stratum \( i \), we matched one crash with nine non-crash samples based on some similar characteristics including driver age, driver gender, and trip distance within a 20% margin. Therefore, there will be some common unobserved factors across the samples, and we capture such correlation using \( \Omega_{iq} \). Further, as we used 20% margin for trip distance to match crash:

\[
Q_{iq} = \beta + \eta \times \text{trip distance margin} \tag{2}
\]

where \( \beta \) (constant) and \( \eta \) are vectors of unknown parameters to be estimated. In estimating the model, it is necessary to specify the structure for the unobserved vectors \( \gamma \) and \( \Omega \) represented by \( \Omega \). In this paper, it is assumed that these elements are drawn from independent normal distribution: \( \Omega \sim N(0, (\pi^2, \Phi^2)) \). Thus, the equation system for modeling the probability of a crash takes the following form (conditional on \( \Omega \)):

\[
P_{iq} = p(\nu_{iq}^*)(\Omega) = \frac{\exp\{ (\alpha + \gamma_{iq})z_{iq} + v_{iq} + \Omega_{iq} \}}{1 + \exp\{ (\alpha + \gamma_{iq})z_{iq} + v_{iq} + \Omega_{iq} \}} \tag{3}
\]

The corresponding probability for non-crash is computed as

\[
Q_{iq} = 1 - P_{iq} \tag{4}
\]

Further, conditional on \( \Omega \), the joint probability \( L_i \) for each stratum \( i \) can be expressed as

\[
L_i = \left[ \prod_{q=1}^{M} \left( P_{iq}\right)^{\nu_{iq}} \times \left( Q_{iq}\right)^{1-\nu_{iq}} \right] f(\Omega)d\Omega \tag{5}
\]
As the integral defined in Equation 5 cannot be analytically estimated, we employ the maximum simulated estimation approach. The simulation technique approximates the likelihood function in Equation 5 by computing the $L_i$ for each stratum $i$ at different realizations drawn from a normal distribution, and averaging it over the different realizations (see Eluru and Bhat [32] for detail). For instance, if $DL_i$ is the realization of the likelihood function in the $c$th draw ($c = 1, 2, \ldots, C$), then the simulated log-likelihood function is as follows:

$$LL = \sum \ln \left( \frac{1}{C} \sum_{c=1}^{C} (DL_i) \right)$$

(6)

The parameters to be estimated in the model are: $\alpha$, $\gamma$, $\varphi$, $\beta$, $\eta$, $\pi$ and $\Phi$. To estimate the proposed model, we apply quasi-Monte Carlo simulation techniques based on the scrambled Halton sequence with $C$ set to 150 (see Eluru et al. [33] and Bhat [34] for examples of quasi-Monte Carlo approaches in literature). We tested the model with higher $C$ values and found the model estimation was stable. We estimate this model using GAUSS matrix programming language.

**Model Results**

The model estimates are presented in Table 3. A discussion of the model results follows.

**Trip-Level Characteristics.** The trip distance parameter was calculated as the natural log of the straight-line distance of the trip plus one. As the distance increases the crash risk associated also increases, highlighting that increased exposure to driving results in an increased risk of a crash. The percentage of trip in a speed category was tested in the model and offered interesting results. We employed the percentage of trip between 30 and 70 mph as the base category. The parameter results indicate that as the percentage of the trip under 30 mph increases, the risk associated with a trip of it resulting in a crash increases. On the other hand, when the percentage of trip over 70 mph increases, the crash risk for the trip reduces. The reader would note that the percentages by speed categories are likely to interact and therefore determining the net magnitude of the variable impact is not straightforward. In the model we considered rural and other roads as the base category and found that as the proportion of a trip on urban roads increases, the risk of a crash decreases. The result could be highlighting potential driver alertness in urban conditions as traffic conflicts are expected.

**Driver Characteristics.** We also examined driver annual mileage as a predictor of crash risk. The variable was categorized into three groups and the 10,000 to 25,000 range was considered as the base. The model estimates indicate that drivers in the lower range (<10,000) and the higher range (>25,000) are at a higher risk relative to the drivers in the normal range (10,000–25,000). It is also interesting to note that the magnitude of the impacts for lower and higher mileage ranges are reasonably close.

We examined whether the employment status had an impact on crash risk. The model parameter for full-time worker indicates these drivers are less at risk than are others.

**Panel and Random Effects.** The model estimation process considered multiple forms of unobserved variables. These include: (a) common unobserved effects for each case-control panel of 10 records, (b) common unobserved factors affecting the error margin in the trip distance variable, and (c) random effects for all independent variables. Among these parameters tested, only one random effect parameter offered a statistically significant result. The result related to full-time worker offered a significant variation indicating that while full-time workers are likely to experience a lower crash risk on average there is substantial variation in the actual reduction. In fact, the result indicates that among full-time drivers, about 82.1% of the time, the crash risk associated will be lower while for the remaining 17.9% of the time crash risk can increase.

**Model Application**

In order for this model to be applied, corrections would need to be made to the constant to match the actual crash to no-crash ratio in the general trip population. In the study we tested crash to no-crash ratios of 1:4, 1:9, 1:14, 1:19, and 1:29, but for the full data set the crash to no-

---

**Table 3. Multi-Level Random Parameters Binary Logit Model Results**

| Parameters | Estimate (SE) | T-Statistic |
|------------|---------------|-------------|
| Constant   | $-2.589 (0.179)$ | $-14.493$ |
| Trip variables | | |
| % Trip <30 mph | $0.515 (0.196)$ | $2.631$ |
| % Trip >70 mph | $-0.525 (0.425)^{**}$ | $-1.236$ |
| Ln(Distance + 1) | $0.194 (0.059)$ | $3.295$ |
| % Trip on urban roads | $-0.51 (0.15)$ | $-3.428$ |
| Driver demographics | | |
| Drives <10,000 mi/yr | $0.457 (0.088)$ | $5.197$ |
| Drives >25,000 mi/yr | $0.466 (0.141)$ | $3.310$ |
| Full-time worker | $-3.340 (2.193)^*$ | $-1.523$ |
| Full-time worker random effect | $3.634 (1.777)$ | $2.045$ |

*Note: SE = standard error; mi/yr = miles per year.

*Variable insignificant at 95% significance level; **Variable insignificant at 85% significance level.
crash ratio was 1:4,850. To calculate this, we adjusted the constant for random effect model so that the probability of a crash would match the 1:4,850 ratio of 0.0002. The resulting calibrated model parameter for the constant was $-8.5527$. This model was then tested on a sample data set of 4,500 randomly selected non-crash trips that had not been used in previous modeling and 500 randomly selected crash trips that were previously used for modeling. Reusing crash trips was necessary because of the limited number of crash trips available. A comparison of the results for the original and calibrated models is shown in Table 4. The results in Table 4 clearly indicate that the calibrated model captures the true ratio of crash to no-crash trips.

**Conclusion**

Traditional crash data has been instrumental in understanding the influence of various factors drawn from driver demographics, vehicle characteristics, roadway characteristics, crash characteristics, and environmental factors on crash frequency and severity. However, we still have challenges to truly understand the underlying cause of the crash as several important information elements, including characteristics of the trip (trip proportion on different facilities: speed limit, roadway functional class), and behavior (like eye movement) and action of the driver (actual speed of the vehicle) at the time of crash, are often missing from the data set. To that extent, the current research effort adopted the SHRP2 NDS, a detailed database recording real-time information for both crash and non-crash trips, to understand and predict the risk of crash occurrence at the finest resolution (trip level). As opposed to focusing on driver demographics, the NDS data allows us to truly understand the underlying timeline of the crash and account for driver behavior in the event of the crash. However, a limitation associated with NDS data is the rarity of crash samples in it relative to non-crash samples ($<0.01\%$). Estimating a binary outcome model for such rarity will be extremely challenging. Therefore, the current study employs a rigorous case-control study design for understanding trip-level crash risk.

For the case-control design, trips with a crash are matched with non-crash trips based on three common matching variables including driver age, driver gender, and trip distance within a 20% margin. Further, we vary the number of controls in the case-control design starting from four to 29 (to be specific, 1:4, 1:9, 1:14, 1:19 and 1:29) and conduct a revised Wald test statistic test to check for the parameter consistency across the samples. Specifically, we employ the 1:29 control sample as the population benchmark and evaluate whether the parameters for other models are statistically different or not. The result clearly highlights the stability in parameter estimates across the samples and, therefore, we restrict ourselves to the 1:9 case-control ratio for further analysis. In particular, employing the 1:9 sample, a multi-level random parameters binary logit model is estimated while considering a comprehensive list of factors including trip characteristics (like day of week, facility types, max acceleration and deceleration), driver demographics (age, gender, income), and crash-level factors (location, collision type, driver impairments, and weather). The model findings clearly illustrate the significant impact of several variables on the crash risk propensity including trip distance, trip proportion of different speed limit roads and facilities, and driver’s driving characteristics and employment status. Further, the proposed model also accommodates for the presence of several unobserved factors on trip-level crash risk with respect to correlation and random effects. However, we only find one random effect parameter offering a statistically significant result, for the full-time worker variable. The result indicates that among drivers employed full time, about 82.1% of the time, the crash risk associated with a trip will be lower while for the remaining 17.9% of the time crash risk associated with a trip can increase. The analysis is further augmented by conducting a prediction exercise on a hold-out sample of data records that is not used for model estimation. However, before generating the prediction, we calibrate the constant of the model to generate a population conforming crash risk model. Findings from the prediction exercise further reinforce the applicability of the model.

The study is not without limitation. The case-control design adopted in the study focused on matching the crashes with non-crashes based on three common attributes. However, there is scope to create multiple case-control designs considering different sets of common factors, such as trip spend on different facilities (rural/urban), trip spend on different speed limits, and other exogenous variables. It will be really interesting to see if the result varies across these different experimental
designs. Exploring these characterizations is an avenue for future research. Finally, recent advances in rare event literature to study skewed outcome contexts is also an avenue of research to address potential bias in binary logit model estimation for skewed samples (see King and Zeng [35], Calabrese and Osmetti [36], and Agarwal et al. [37]).

This study has contributed to safety research in two important ways. First, we presented a framework to employ NDS data to understand and predict crash risk at a disaggregate trip level accommodating for the influence of trip characteristics as well as traditional crash factors. Second, we employed a rigorous case-control study design for understanding trip-level crash risk. In the future, this research can serve as the foundation for safety researchers to employ SHRP2 and future NDS data for understanding and predicting crash risk.

Acknowledgments

The authors would like to gratefully acknowledge Virginia Tech Transportation Institute (VTTI) for providing access to the SHRP2 NDS database and for providing the trip- and crash-level data adopted in the current study.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Naveen Eluru, Tanmoy Bhowmik, Shamsunnahar Yasmin; data collection: Lauren Hoover, Tanmoy Bhowmik, Naveen Eluru; analysis and interpretation of results: Lauren Hoover, Tanmoy Bhowmik, Naveen Eluru; draft manuscript preparation: Lauren Hoover, Tanmoy Bhowmik, Naveen Eluru, Shamsunnahar Yasmin. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Tanmoy Bhowmik https://orcid.org/0000-0002-0258-1692
Shamsunnahar Yasmin https://orcid.org/0000-0001-7856-5376
Naveen Eluru https://orcid.org/0000-0003-1221-4113

Data Accessibility Statement

The data acquisition was part of an FHWA project titled “Development and Application of a Disaggregate Artificial Realistic Data Generator for Computationally Testing Safety Analysis Methods. The data that support the findings of this study are available on request from the corresponding author (TB)”

References

1. Vision Zero Network. Vision Zero Communities. Vision Zero Network. https://visionzeronetwork.org/resources/vision-zero-communities/. Accessed June 16, 2021.
2. Lord, D., and F. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A: Policy and Practice, Vol. 44, No. 5, 2010, pp. 291–305.
3. Yasmin, S., and N. Eluru. Latent Segmentation Based Count Models: Analysis of Bicycle Safety in Montreal and Toronto. Accident Analysis and Prevention, Vol. 95, 2016, pp. 157–171.
4. Bhowmik, T., M. Rahman, S. Yasmin, and N. Eluru. Exploring Analytical, Simulation-Based, and Hybrid Model Structures for Multivariate Crash Frequency Modeling. Analytic Methods in Accident Research, Vol. 31, 2021, p. 100167.
5. Yasmin, S., and N. Eluru. Evaluating Alternate Discrete Outcome Frameworks for Modeling Crash Injury Severity. Accident Analysis and Prevention, Vol. 59, No. 1, 2013, pp. 506–521.
6. Marcoux, R., S. Yasmin, N. Eluru, and M. Rahman. Evaluating Temporal Variability of Exogenous Variable Impacts over 25 Years: An Application of Scaled Generalized Ordered Logit Model for Driver Injury Severity. Analytic Methods in Accident Research, Vol. 20, 2018, pp. 15–29.
7. Kabli, A., T. Bhowmik, and N. Eluru. A Multivariate Approach for Modeling Driver Injury Severity by Body Region. Analytic Methods in Accident Research, Vol. 28, 2020, p. 100129.
8. Dingus, T. A., S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, et al. The 100-Car Naturalistic Driving Study: Phase II—Results of the 100-Car Field Experiment. Publication DOT-HS-810-593. NHTSA, US Department of Transportation, 2006.
9. Antin, J. F., S. Lee, M. A. Perez, T. A. Dingus, J. M. Hankey, and A. Brach. Second Strategic Highway Research Program Naturalistic Driving Study Methods. Safety Science, Vol. 119, No. 10, 2019, pp. 2–10.
10. Hankey, J. M., 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, Blacksburg, 2016.
11. Guo, F. Statistical Methods for Naturalistic Driving Studies. Annual Review of Statistics and Its Application, Vol. 6, 2019, pp. 309–328.
12. Abdel-Aty, M., and A. Pande. Crash Data Analysis: Collective vs. Individual Crash Level Approach. Journal of Safety Research, Vol. 38, 2007, pp. 581–587.
13. Xu, C., P. Liu, and W. Wang. Evaluation of the Predictability of Real-Time Crash Risk Models. Accident Analysis and Prevention, Vol. 94, 2016, pp. 207–215.
14. Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. The Impact of Driver Inattention on
15. Guo, F., and Y. Fang. Individual Driver Risk Assessment Using Naturalistic Driving Data. *Accident Analysis and Prevention*, Vol. 61, 2013, pp. 3–9.

16. Dingus, T. A., F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey. Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 113, No. 10, 2016, pp. 2636–2641.

17. Owens, J. M., T. A. Dingus, F. Guo, Y. Fang, M. Perez, and J. McClafferty. Crash Risk of Cell Phone Use While Driving: A Case-Crossover Analysis of Naturalistic Driving Data. *AAA Foundation for Traffic Safety*, Washington, DC, 2018.

18. Huisingh, C., C. Owsley, E. B. Levitan, M. R. Irvin, P. MacLennan, and G. McGwin. Distracted Driving and Risk of Crash or Near-Crash Involvement among Older Drivers Using Naturalistic Driving Data with a Case-Crossover Study Design. *Journals of Gerontology: Medical Sciences*, Vol. 74, No. 4, 2019, pp. 550–555.

19. Bärgman, J., 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, 2015, pp. 152–169.

20. Hamzeie, R., P. T. Savolainen, and T. J. Gates. Driver Speed Selection and Crash Risk: Insights From the Naturalistic Driving Study. *Journal of Safety Research*, Vol. 63, 2017, pp. 187–194.

21. Kamran, M., R. Arvin, and A. J. Khattak. The Role of Aggressive Driving and Speeding in Road Safety: Insights from SHRP2 Naturalistic Driving Study Data. Presented at 98th Annual Meeting of the Transportation Research Board Conference, Washington, DC, 2019.

22. Sacrist, T., E. C. Douglas, C. Hannan, R. Rogers, A. Belwadi, and H. Loeb. Near Crash Characteristics among Risky Drivers Using the SHRP2 Naturalistic Driving Study. *Journal of Safety Research*, Vol. 73, 2020, pp. 263–269.

23. Hao, H., Y. Li, A. Medina, R. B. Gibbons, and L. Wang. Understanding Crashes Involving Roadway Objects with SHRP 2 Naturalistic Driving Study Data. *Journal of Safety Research*, Vol. 73, 2020, pp. 199–209.

24. Simons-Morton, B. G., P. Gershon, F. O’Brien, G. Gensler, S. G. Klauer, J. P. Ehsani, C. Zhu, R. E. Gore-Langton, and T. A. Dingus. Crash Rates over Time among Younger and Older Drivers in the SHRP 2 Naturalistic Driving Study. *Journal of Safety Research*, Vol. 73, 2020, pp. 245–251.

25. Liu, S., M. A. Perez, and N. Lau. The Impact of Sleep Disorders on Driving Safety—Findings from the Second Strategic Highway Research Program Naturalistic Driving Study. *Sleep*, Vol. 41, No. 4, 2018.

26. Ali, E. M., M. M. Ahmed, and S. S. Wulff. Detection of Critical Safety Events on Freeways in Clear and Rainy Weather Using SHRP2 Naturalistic Driving Data: Parametric and Non-Parametric Techniques. *Safety Science*, Vol. 119, 2019, pp. 141–149.

27. Hickman, J. S., and R. J. Hanowski. An Assessment of Commercial Motor Vehicle Driver Distraction Using Naturalistic Driving Data. *Traffic Injury Prevention*, Vol. 13, No. 6, 2012, pp. 612–619.

28. Carney, C., D. McGehee, K. Harland, M. Weiss, and M. Raby. Using Naturalistic Driving Data to Assess the Prevalence of Environmental Factors and Driver Behaviors in Teen Driver Crashes. *AAA Foundation for Traffic Safety*, Washington, DC, 2015.

29. Mittleman, M. A., M. Macle, and J. M. Robins. Control Sampling Strategies for Case-Crossover Studies: An Assessment of Relative Efficiency. *American Journal of Epidemiology*, Vol. 142, 1995, pp. 91–98.

30. Zhang, J., and F. Y. Kai. What’s the Relative Risk?: A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA*, Vol. 280, 1998, pp. 1690–1691.

31. Mann, C. Observational Research Methods. Research Design II: Cohort, Cross Sectional, and Case-Control Studies. *Emergency Medicine Journal*, Vol. 20, 2003, pp. 54–60.

32. Eluru, N., and C. R. Bhat. A Joint Econometric Analysis of Seat Belt Use and Crash-Related Injury Severity. *Accident Analysis and Prevention*, Vol. 39, 2007, pp. 1037–1049.

33. Eluru, N., C. R. Bhat, and D. A. Hensher. A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Level in Traffic Crashes. *Accident Analysis and Prevention*, Vol. 40, 2008, pp. 1033–1054.

34. Bhat, C. R. Quasi-Random Maximumsimulated Likelihood Estimation of the Mixed Multinomial Logit Model. *Transportation Research Part B: Methodological*, Vol. 35, 2001, pp. 677–693.

35. King, G., and L. Zeng. Logistic Regression in Rare Events Data. *Political Analysis*, Vol. 9, No. 2, 2001, pp. 137–163.

36. Calabrese, R., and S. A. Osmetti. Modelling Small and Medium Enterprise Loan Defaults as Rare Events: The Generalized Extreme Value Regression Model. *Journal of Applied Statistics*, Vol. 40, No. 6, 2013, pp. 1172–1188.
