Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Computing Odds of Crashes and Near-crashes using Naturalistic Driving Study Data

Nipjyoti Bharadwaj[a], Praveen Edara[b] *, Carlos Sun[b]

[a]Graduate research assistant, University of Missouri - Columbia, Columbia, MO 65211, USA
[b]Professor, University of Missouri - Columbia, Columbia, MO 65211, USA

Abstract

Naturalistic driving study (NDS) data offers a different lens to examine the causal factors of crashes and near-crashes (CNC). In the U.S., a comprehensive NDS data collection effort was conducted as part of the Second Strategic Highway Research Program (SHRP 2). SHRP 2 NDS data includes information related to driver behavior and various non-driving related tasks performed while driving. Where the NDS data markedly differs from traditional crash databases is its capture of microscopic data pertaining to driver behavior (e.g. distraction, secondary tasks). In this article, we present the results of an exploration of NDS data. The exploration included developing a Logistic regression model to estimate crash risk for different factors using matched case-control design and odds ratios. The developed model's goodness of fit was evaluated using receiver operating characteristic curves for training and validation data sets. The study found that performing a non-driving related secondary task for more than 6 seconds increases the CNC risk by 5.48 times. Among different driver behavior factors, inattention was found to be the most critical factor contributing to CNC risk with an odds ratio of 16.16. Traffic conditions corresponding to Level of Service (LOS) D exhibited the highest level of CNC risk. The computation of odds ratios enables making informed decisions while designing countermeasures to enhance safety. NDS data provides an opportunity to perform these computations.

Keywords: Naturalistic driving; Safety; Modeling; Crashes and Near-Crashes

1. Introduction

Safety continues to be a high priority for transportation agencies around the world. According to National Highway Traffic Safety Administration (NHTSA), in 2016, 37,461 lives were lost on U.S. roads, an increase of 5.6 percent from 2015. Understanding the causes of crashes is essential for developing proactive countermeasures to improve safety.

* Corresponding author. Tel.: +1 573-882-1900; fax: (573) 882-4784.
E-mail address: EdaraP@missouri.edu

Traditional safety research relied on using statistical modeling approaches to capture the effect of various road, traffic, and environmental characteristics. Poisson and Negative Binomial models are well accepted approaches to model count data such as crashes [Guo et al. (2010), Lord and Mannering (2010), Lord et al. (2005), El-Basyouny and Sayed (2006)]. However, the effect of individual driver's behavior is typically not included in such approaches due to the difficulty in measurement. The Transportation Research Board's second Strategic Highway Research Program (SHRP 2) conducted a large naturalistic driving study (NDS) to investigate the role of driver and other factors in crash and near-crash events [Dingus et al. (2015)]. Over 3,000 drivers participated in the comprehensive driving experiment from six sites in Florida, Indiana, New York, North Carolina, Pennsylvania, and Washington. Nearly 50 million vehicle miles of data was recorded from trips made by these drivers. This unprecedented data allows for the investigation of the role of driver behavior in traffic safety. The interaction of driver with the vehicle, roadway, and the environment is captured in detail. Such highly detailed data enables a more accurate determination of the causes of crashes and near-crashes than the typical post-crash investigation using law enforcement data. The NDS data addresses a need that is not fulfilled by traditional data sources used in safety research.

This study was conducted to better understand the contributing factors of safety critical events using NDS data. This improved understanding of risk of a safety critical event occurring should enable design of countermeasures such as proactive warning systems to alert drivers of impending hazardous conditions. The dataset comprised of safety critical events related to drivers of all age groups and various weather and geometric conditions. The study has three key objectives: (1) identify factors associated with individual driver risk of being involved in a safety critical event, (2) develop a logistic regression model to predict this risk, and (3) quantify risk for different factors using matched case-control design and odds ratio (OR).

2. Literature Review

The SHRP 2 NDS data has only become publicly available recently. Thus, there are few research studies in the literature using this data. However, there have been studies using other types of naturalistic driving data from much smaller samples (e.g. 100-car study). The SHRP 2 experiment is the most comprehensive of the three data sources. It includes data from different road conditions, facilities, weather conditions, traffic conditions, and driving behavior.

Both statistical and machine learning methods have been explored in previous NDS research. Klauer et al. (2006) studied the relationship between driving behavior and CNC involvement. ORs were estimated using CNC and baseline driving data for various sources of inattention. The results indicated that driving while drowsy results in four- to six-times higher CNC risk than alert drivers. Drivers engaging in visually and/or manually complex tasks have a three-time higher CNC risk than drivers who are attentive. The number of actual crashes observed in NDS studies is relatively small, so near crashes are usually used as a crash surrogate. Guo et al. (2010) used precision and bias of risk estimation to validate near crash as a crash surrogate. The results suggested that near crash can be combined with crashes for statistical analysis. Klauer et al. (2010) calculated relative crash risk associated with various types of secondary tasks using case-crossover baseline. Conditional logistic regression was used to calculate ORs for drowsiness, secondary task engagement, and total time eyes-off-road. The study also assessed the relationship between task duration and eyes-off-road total time. Results indicated that drivers in the 100-Car Study engaged in secondary tasks 23.5 percent of the time that they were driving, approximately 40 percent higher than indicated in previous research. Guo and Fang (2013), used negative binomial method to identify factors related to crash and near-crash risk of individual drivers. After identifying the factors, drivers were classified into three risk groups using K-mean cluster method. The logistic regression method was applied to predict the high- and moderate-risk drivers. The study concluded that crash and near-crash risk for individual drivers is associated with the critical incident rate, and demographic and personality characteristics.

Xu and Fujimura (2014), employed Random Forest for driver's activity recognition. For training the model, a sequence of depth images were used as input, and output was an activity class among a predetermined set of driver activities. Geng et al. (2016) developed a neural network-based model for driver speed profile modeling at curvy paths. Five models with different prediction steps were developed to fit driver speed profiles under different driving situations. Chang and Edara (2017), applied classification methods to examine safety critical events in work zones. Four machine learning algorithms, Random forest, Deep Neural Network, Multilayer Feedforward Neural Network, and t-Distributed Stochastic Neighbor Embedding (t-SNE), were applied to work zone events within NDS data. The

Random forest algorithm performed the best in classifying NDS data into crashes, near-crashes, and baseline using pre-event variables.

The reviewed literature revealed that few studies used SHRP 2 NDS data for understanding crash causation. Majority of the existing using naturalistic driving data relied on significantly smaller datasets such as the 100-car study. The SHRP 2 NDS data is the most comprehensive NDS dataset available in the US, thus providing greater variability of conditions in which crashes occur.

3. SHRP 2 NDS data

SHRP2 NDS data consists of 36,103 crash, near-crash, and baseline events representing various drivers and conditions. For each event, data is available for 76 different variables. Each variable consists of several categories, which provide in depth details of driver behavior and other network characteristics. For example, "Driver behavior" variable includes the following categories: 'Distracted', 'Drowsy, sleepy, asleep, fatigued', 'Exceeded speed limit', 'Exceeded safe speed but not speed limit', 'Driving slowly: below speed limit', 'Passing on right', 'Illegal passing', 'Cutting in, too close in front of other vehicle', 'Making turn from wrong lane', 'Aggressive driving', and 'Following too closely'. Table 1 shows the sample sizes for different types of NDS events.

Table 1. Sample size of SHRP2 NDS events

Event	All Events
Crash	1,474 (4.1%)
Near-Crash	2,767 (7.7%)
Baseline	31,862 (88.3%)
Total	36,103 (100.0%)

The Virginia Tech Transportation Institute (VTTI) processed and provides SHRP 2 NDS data for researchers [Hankey et al. (2016)]. VTTI defines safety critical and baseline events as follows:

- Crash: Any contact that the subject vehicle has with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated is considered a crash.
- Near-Crash: Any circumstance that requires a rapid evasive maneuver by the subject vehicle, or any other vehicle, pedestrian, cyclist, or animal, to avoid a crash is considered a near-crash.
- Baseline: The goal of the baselines is to provide an estimate of what constitutes "normal driving" and "typical driver behavior" across the sample.

4. Logistic Regression for Prediction of Crashes and Near Crashes

In the NDS data, events were classified as crashes, near-crashes and baseline with a set of explanatory variables related to driver, roadway, and traffic conditions. The safety critical events are discrete responses and logistic regression is one of the most common methods used to investigate the relationship between discrete responses and a set of predictors [Kleinbaum et al. (2013)]. For binary responses, the outcome is expressed in terms of the probability of modeled response. For building a logistic regression model, a dichotomous variable was created by combining crash or near crash (CNC) as one response (i.e. a safety critical event) and baseline as the other. The setup of the model is illustrated as follows.

$$Y_i = \begin{cases} 1\ if\ crash\ or\ near\ crash \\ \quad 0\ if\ baseline \end{cases}$$

Let X_i is a matrix of predictors for an individual event i and p_i = Pr $(Y_i = 1)$ is the probability of occurrence of CNC, the logit function [Guo and Fang (2013)] can be defined as

$$\text{Logit } (p_i) = \log\left(\frac{p_i}{1-p_i}\right) = X_i\beta \qquad (1)$$

Where, β is the vector of regression parameters.

4.1. Development of the Prediction Model

In this study, potential risk factors related to safety critical events were identified from a previous study conducted by Chang and Edara (2017). These factors are duration of secondary task, driving behavior, maneuver judgement, traffic density and intersection influence. A secondary task in NDS data is defined as a non-driving related task performed by the driver while driving such as eating, talking, texting, etc. The dataset recorded up to three secondary tasks performed by the drivers with start and end time of each task. For this study, duration of secondary task was classified into two categories: 0-6 secs and greater than 6 secs. The 6 second cut off point was determined from quantile values. It was observed from the available NDS data that upper quantile value of secondary task duration is nearer to 6 seconds (secs) for safety critical events. Thus, a cut off value of 6 secs was established for comparison purposes.

The driving behavior variable is defined as behavior prior to a precipitating event or those resulting from the context of the driving environment, contributing to the crash or near crash. The dataset defines 55 different categories of driving behaviors. In compliance with the studies conducted by Reason et al. (1990), Åberg and Rimmö (1998), Rimmö and Åberg (1999), the driving behavior categories in the present study were merged into four groups, namely, violation, mistakes, inattention and experience.

Maneuver judgement is defined based on vehicle kinematics. Vehicle kinematics includes the subject vehicle's position and speed and direction of movement in relation to other vehicles, or environmental characteristics and surroundings. In present study, maneuver judgement is classified into two categories: safe and unsafe maneuvers.

The traffic density variable is classified into six categories from level of service A to F [HCM (2010)], based on the operating conditions. The six defined levels of service, A-F, describe operations, from the best to the worst, for a type of facility. Intersection influence is variable created based on subjective determination of whether the subject vehicle's safe movement, travel path, and travel speed, are under the influence of an intersection at the time of the event. The variable is defined in to six categories as no influence, interchange, parking, stop sign, signal and uncontrolled. Statistical analysis software (SAS, 9.4) was used to develop the logistic regression model for safety critical events. From the NDS dataset, 80% data were used to develop the model while remaining 20% were used for validation. The model estimates are summarized in Table 2.

Table 2. Summary of logistic model estimates

Effects	Estimate	P-value
Intercept	-4.1056	<.0001
Duration of secondary task: < 6 sec vs. 0-6 sec	1.5012	<.0001
Behavior : Violation vs. None	0.7471	<.0001
Behavior : Mistake vs. None	1.7807	<.0001
Behavior : Inattention vs. None	3.0398	<.0001
Behavior : Inexperience vs. None	3.4272	<.0001
Maneuver : Unsafe vs. Safe	1.8261	<.0001
Traffic density : LOS B vs. LOS A	0.5192	<.0001
Traffic density : LOS C vs. LOS A	1.3683	<.0001
Traffic density: LOS D vs. LOS A	1.5277	<.0001
Traffic density : LOS E vs. LOS A	1.0668	<.0001
Traffic density : LOS F vs. LOS A	0.3503	0.4341
Intersection influence : Interchange vs. No influence	1.7069	<.0001
Intersection influence : Parking vs. No influence	2.5028	<.0001
Intersection influence : Stop sign vs. No influence	1.0232	<.0001
Intersection influence : Signal vs. No influence	1.6645	<.0001
Intersection influence : Uncontrolled vs. No influence	2.3274	<.0001

Model fitness was evaluated using two approaches. The first approach was using goodness of fit statistics. The Hosmer-Lemeshow test [Hosmer et al. (2013)] can compute the goodness of fit for logistic regression models. A small Chi-squared value (with larger p-value i.e. >.05) indicates a good logistic regression model fit. For the estimated model, Chi-square value was 6.35 (p-value =0.17), indicating good fit. The second approach is to evaluate the predictive power of the model with the help of Receiver Operating Characteristic (ROC) Curves. An ROC curve is a plot of the proportion of sensitivity (CNC predicted to be CNC) versus the proportion of specificity (baseline predicted to be baseline). The output of logistic regression contains predicted probabilities of CNC for each observation. For any given threshold (i.e., predicted probability of CNC), there is a tradeoff between sensitivity and specificity. A set of possible thresholds between 0 to 1 were generated, and respective sensitivity and (1-specificity) were plotted for ROC curve in FIGURE 1 (a). The predictive power of models was estimated by the area under curve (AUC), where a higher area indicates better prediction ability. The AUC for prediction model was 0.877, indicating good prediction ability of CNC events (highest possible value of AUC=1).

4.2. Validation of the Prediction Model

The performance of the prediction model was evaluated by applying the model to the validation data set. The probability of CNC for each event is calculated as follows,

$$Probability\ (CNC) = \frac{\exp(-4.105+1.501*duration+0.747*violation+\cdots.........+2.327*uncontrolled)}{1+\exp(-4.105+1.501*duration+0.747*violation+\cdots.........+2.327*uncontrolled)} \qquad (2)$$

An ROC curve was developed for the predicted probabilities of CNC for the validation dataset. The AUC was 0.8862 for the validation dataset, indicating good predictive ability of the model and is plotted in FIGURE 1(b). Another approach to visualize the performance of the prediction model on the validation data set is to plot all CNC and baseline events with respect to predicted probabilities, as is depicted in FIGURE 2. Higher probability values indicate a higher likelihood of CNC event occurring.

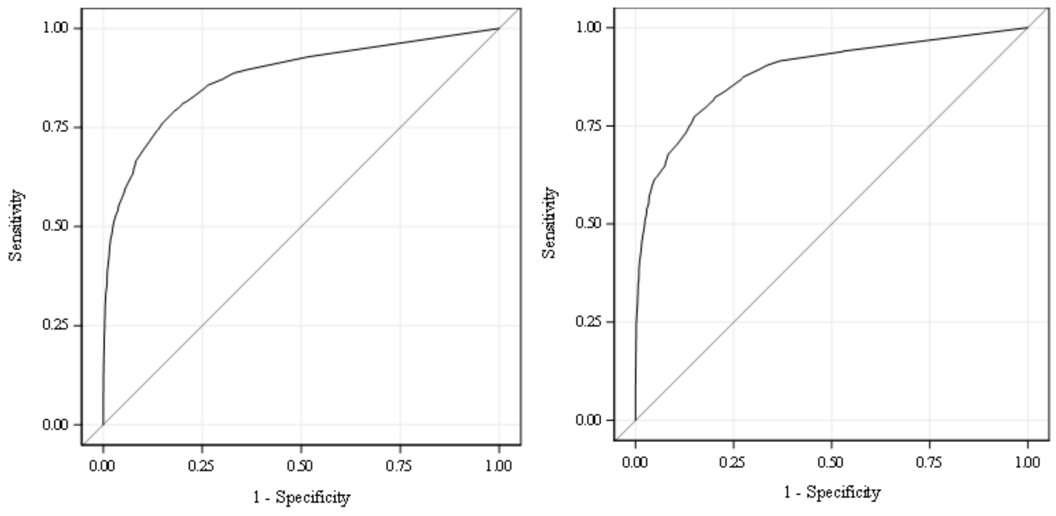

Fig. 1 (a) Fig. 1 (b)

Fig. 1(a) Prediction model ROC (b) Validation dataset ROC

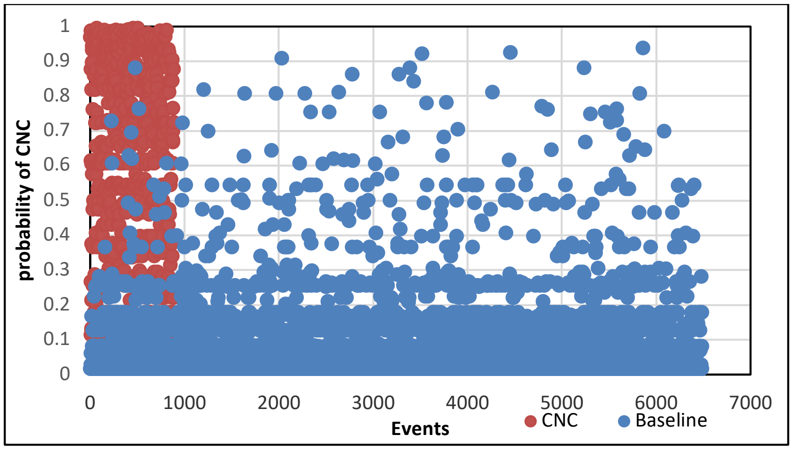

Fig. 2. Probability of CNC for CNC and baseline events

5. Crash and Near Crash Risk Assessment using Matched Case Control Design

Regression can be used to achieve two different goals. The first is to predict a dependent variable using a set of predictors. The second is to quantify the relationship between predictors and a dependent variable. In logistic regression, the exponential of regression parameter, exp (β), is the odds ratio (OR) for a variable. OR compares two or more groups of a predictor variable with regard to the dependent variable. To estimate accurately the elements of β of a variable, it is important to control for confounding variables. A confounding effect exists if inclusion or omission of an extraneous variable changes completely or partially the apparent association between an outcome and risk factor [Kleinbaum et al. (2013)]. In observational studies, potential confounding variables can be controlled directly by matching cases and controls during the design stage. In this study, events with CNC are considered as cases while baselines are considered as controls. In matched case control studies, each case is matched with one or more controls who have the same values or same categories of each potential confounding variable. The ratio of controls to cases depends on the availability of data, and, as the ratio increases, the power of the design increases but at a decreasing rate [Gross and Jovanis (2007), Woodward (2005)]. In this study, one control is selected for each case (a one to one ratio), which reaches about 90% power [Gross and Jovanis (2007), Woodward (2005)]. For a matched dataset, conditional logistic regression is appropriate to use [Kleinbaum et al. (2013), Gross and Jovanis (2007)]. The model is

$$Logit\ (p_i) = \beta_0 + \beta_1 E + \sum_{i=1}^{n-1} \gamma_i V_i \tag{3}$$

Where, p_i is the probability of occurrence of CNC; E is the predictor variable; V_i denote a set of variables distinguishing matched for n number of pairs. In this study, conditional logistic regression were applied to estimate the relative CNC risk for three predictor variables of interest – driver behavior, traffic density, and duration of secondary task.

5.1. Assessment of Confounding Variables

As previously stated, confounding effects exist if presence of an extraneous variable significantly changes the relationship between outcome and risk factor. Thus, confounding effect, on an outcome-risk factor relationship can be assessed by comparing the crude estimates (considering the effect of only risk factor at a time) with adjusted estimates (considering the effect of a risk factor along with its confounders). If estimates change by 10% or more, due to the addition of an extraneous variable, the confounding effect can be considered to be significant [Kleinbaum et al. (2013)]. The methodology was applied to identify the significant confounding variables of driver behavior, traffic density, and duration of secondary task. The identification procedure is illustrated in Table 2 for duration of secondary task.

Table 2. Identification of significant confounders for duration of secondary task

Models	β estimate for duration	Change in β
Duration of secondary task (crude model)	1.82	
Duration of secondary task + Behavior	1.55	14%
Duration of secondary task + Traffic density	1.82	0%
Duration of secondary task + Intersection influence	1.83	-1%
Duration of secondary task + Maneuver judgement	1.82	0%

From Table 2, it can be observed that behavior is a significant confounding variable for duration of secondary task variable. Due to the addition of the behavior variable, the crude estimates of duration changed significantly (14%). Therefore, behavior can be controlled using a criterion to match cases and controls.

The process was repeated for behavior and traffic density variables and their confounders. It was found that maneuver and intersection influence were significant confounders for driving behavior and duration, maneuver, intersection influence and behavior were significant confounders for traffic density.

5.2. Matching Cases and Controls

Matching of cases and controls was accomplished for duration, behavior, and traffic density variables. Three separate datasets were generated, with one for each variable of interest (duration, behavior and traffic density). Significant confounders of the variable were considered as matching criteria for each dataset. The matching process was accomplished with MatchIt package of R software. MatchIt package is developed for non-parametric preprocessing of data with various matching methods [Ho et al. (2011)]. The goal of the matching process was to create a dataset where the distribution of confounding variables is similar in both case and control groups, which is known as balance [Ho et al. (2011)]. The balance was evaluated by comparing the differences in means between the groups. Upon matching, 2673 pairs of case-control were identified for the secondary task duration variable, 2945 pairs for behavior and 1680 pairs for traffic density were identified. An example of matching data set for the duration variable, where behavior is considered as a criterion for matching, is shown in Table 3. Case and controls were selected randomly without replacement and matched one-to-one for the most significant variable i.e. behavior. Each selected case control pair have same category of behavior while other variables were unmatched.

Table 3. Sample matching for secondary task time duration

Pair	Outcome	Behavior	Traffic density	Maneuver	Influence	Duration
1	CNC	Inattention	LOS C	Safe	Signal	<6 secs
	baseline	Inattention	LOS B	Safe	None	0-6 secs
2	CNC	Inattention	LOS D	Safe	None	<6 secs
	baseline	Inattention	LOS A	Safe	None	<6 secs
3	CNC	None	LOS B	Safe	None	0-6 secs
	baseline	None	LOS A	Safe	None	0-6 secs
4	CNC	Violation	LOS A	Unsafe	Parking	0-6 secs
	baseline	Violation	LOS A	Unsafe	None	0-6 secs

5.3. Analysis and Results

Logistic regression was used to estimate both crude and adjusted odds ratios. Three univariate logistic regression models were developed for duration, behavior and traffic density to estimate the relationship of respective variables with safety critical events. The crude odds ratios estimated from the univariate models were unadjusted for confounders and reported in Table 4 with 95% confidence intervals.

Table 4: Crude odds ratios from regression

Risk Factors	Categories	Odds Ratio (95% CI in parenthesis)
Duration of secondary task	<6 secs vs. 0-6 secs	6.15 (5.63-6.71)
Behavior	Violation vs. None	7.75 (6.93-8.67)
	Mistake vs. None	13.65 (11.96-15.58)
	Inattention vs. None	29.71 (26.39-33.43)
	Inexperience vs. None	50.53 (26.80-95.29)
Traffic density	LOS B vs. LOS A	1.72 (1.59-1.87)
	LOS C vs. LOS A	3.65 (3.21-4.16)
	LOS D vs. LOS A	4.82 (3.89-5.96)
	LOS E vs. LOS A	2.57 (1.83-3.62)
	LOS F vs. LOS A	3.12 (1.63-5.96)

From Table 4, it can be observed that there was significant unadjusted association between safety critical events and risk factors. The risk of crash (or near-crash) is significantly higher for behavior risk factors. Table 4 shows that, among listed levels of behavior, risk increased gradually from violation to inexperience level. For traffic density, the risk of CNC increased by 1.72 times for LOS B as compared to LOS A. The risk is further increased in LOS C and LOS D, beyond which a decreasing trend in risk was observed. LOS D is found to have the highest level of risk of CNC. The results also indicate that duration is a significant factor contributing to risk of CNC.

As discussed in the previous section, for more precise estimation of the elements of β of a variable, it is important to control for confounding variables. The confounding status can be elucidated by matching the cases and controls for significant confounders. After matching the significant confounders, three conditional logistic regression models were developed to investigate the relationships of CNC with duration, behavior and traffic density. In all the three models, significant confounders were matched for case controls while remaining covariates were unmatched. ORs obtained from the conditional logistic regressions for a variable were adjusted for all potential confounders. Adjusting for confounders, the odds of CNC for duration, behavior and traffic density are provided in Table 5 along with 95% confidence intervals.

TABLE 5: Adjusted odds ratios for safety critical events

Risk Factors	Categories	Odds Ratio (95% CI in parenthesis)
Duration of secondary task:	<6 secs vs. 0-6 secs	5.48 (4.35-6.91)
Behavior	Violation vs. None	1.97 (1.39-2.79)
	Mistake vs. None	5.09 (3.97-6.51)
	Inattention vs. None	16.16 (11.65-22.43)
	Inexperience vs. None	9.27 (3.35-25.69)
Traffic density	LOS B vs. LOS A	1.53 (1.29-1.81)
	LOS C vs. LOS A	4.37 (3.18-6.01)
	LOS D vs. LOS A	4.65 (2.45-6.72)
	LOS E vs. LOS A	1.92 (0.98-3.77)
	LOS F vs. LOS A	0.38 (0.04-3.34)

The adjusted odds ratios overall trends are consistent with those observed for crude odds ratios. Although similar trend can be observed for different categories of variables, the values changed significantly after adjusting for confounders. From the odds ratio estimates of TABLE 5, the duration of secondary task greater than 6 secs increases the CNC risk by 5.48 times. Driving behavior was found to be the most critical risk factor. Any kind of deviation from normal driving can cause or contribute to higher CNC risk. The risk of CNC increased by 1.97 times for any 'violation' and 5.09 times for any 'mistake'. Among all types of driver behavior exhibited in a safety critical event, 'inattention' has the highest level of CNC risk (16.16 times). For traffic density, there is an incremental trend of CNC risk from LOS B to LOS D then there is decreasing trend in risk. LOS D was found to have highest level of risk for CNC (4.65 times). The highest CNC risk in LOS D could be due to traffic flow transitioning from free flowing to congested conditions and possibly less opportunities for lane changing. The reversal of trend after LOS D could be due to the lower speeds encountered in LOS E and F conditions.

6. Conclusion

This study used NDS data to explore the relationship between safety critical events and various potential contributing factors. A simple binary Logistic regression model was found to accurately capture the relationship. The study found that the duration of secondary task, driving behavior, traffic density, maneuver judgement and intersection influence were significant risk factors.

Driving behavior was found to be the most critical risk factor. Among all types of driver behavior exhibited in a safety critical event, inattention has the highest level of CNC risk. This finding points to the need for developing countermeasures such as warning messages to alert the driver. There was also evidence that performing a non-driving related secondary task for more than 6 seconds increases the CNC risk (by 5.48 times). Traffic conditions corresponding to Level of Service (LOS) D exhibited the highest level of CNC risk. The computation of odds ratios enables making informed decisions while designing countermeasures to enhance safety.

The study results provide motivation to apply the developed methodology to compute the odds for different subsets of NDS data such as: 1) older drivers, 2) younger drivers, 3) intersection crashes, 4) weather, and 5) rural intersections. Having separate models for each emphasis area will enable the development of optimal countermeasures targeted to that specific area.

7. Acknowledgments

The research presented in this paper was generated as part of an ongoing FHWA SHRP 2 BAA grant. The authors would like to acknowledge the assistance provided by the FHWA Project Manager, Mr. Charles Fay. Dr. Yohan Chang provided assistance with processing the NDS data used in this study. Assistance provided by VTTI team in supplying the NDS data is also respectfully acknowledged.

8. References

1. Aberg, L. and Rimmo, P.A., 1998. Dimensions of aberrant driver behaviour. *Ergonomics*, *41*(1), pp.39-56.
2. Chang, Y. and Edara, P., 2017, October. Predicting hazardous events in work zones using naturalistic driving data. In *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on* (pp. 1-6). IEEE.
3. Dingus, T.A., Hankey, J.M., Antin, J.F., Lee, S.E., Eichelberger, L., Stulce, K.E., McGraw, D., Perez, M. and Stowe, L., 2015. *Naturalistic driving study: Technical coordination and quality control* (No. SHRP 2 Report S2-S06-RW-1).
4. El-Basyouny, K. and Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. *Transportation Research Record*, *1950*(1), pp.9-16.
5. Geng, X., Liang, H., Xu, H., Yu, B. and Zhu, M., 2016, June. Human-driver speed profile modeling for autonomous vehicle's velocity strategy on curvy paths. In *Intelligent Vehicles Symposium (IV), 2016 IEEE* (pp. 755-760). IEEE.
6. Gross, F. and Jovanis, P.P., 2007. Estimation of the safety effectiveness of lane and shoulder width: Case-control approach. *Journal of transportation engineering*, *133*(6), pp.362-369.
7. Guo, F. and Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, *61*, pp.3-9.
8. Guo, F., Klauer, S.G., Hankey, J.M. and Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. *Transportation Research Record*, *2147*(1), pp.66-74.

9. Hankey, J.M., Perez, M.A. and McClafferty, J.A., 2016. *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets*. Virginia Tech Transportation Institute.
10. Ho, D.E., Imai, K., King, G. and Stuart, E.A., 2011. MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), pp.1-28.
11. Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression* (Vol. 398). John Wiley & Sons.
12. Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D. and Ramsey, D.J., 2006. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data.
13. Klauer, S.G., Guo, F., Sudweeks, J. and Dingus, T.A., 2010. *An analysis of driver inattention using a case-crossover approach on 100-car data* (No. HS-811 334).
14. Kleinbaum, D., Kupper, L., Nizam, A. and Rosenberg, E., 2013. *Applied regression analysis and other multivariable methods*. Nelson Education.
15. Lord, D. and Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), pp.291-305.
16. Lord, D., Washington, S.P. and Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, *37*(1), pp.35-46.
17. Manual, H.C., 2010. HCM2010. *Transportation Research Board, National Research Council, Washington, DC*.
18. National Highway Traffic Safety Administration, 2016. 2015 motor vehicle crashes: overview. *Traffic safety facts research note*, *2016*, pp.1-9.
19. Reason, J., Manstead, A., Stradling, S., Baxter, J. and Campbell, K., 1990. Errors and violations on the roads: a real distinction?. *Ergonomics*, *33*(10-11), pp.1315-1332.
20. Rimmö, P.A. and Åberg, L., 1999. On the distinction between violations and errors: sensation seeking associations. *Transportation Research Part F: Traffic Psychology and Behaviour*, *2*(3), pp.151-166.
21. Woodward, M., 2013. *Epidemiology: study design and data analysis*. Chapman and Hall/CRC.
22. Xu, L. and Fujimura, K., 2014, September. Real-time driver activity recognition with random forests. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 1-8). ACM.