2021

Modeling of Crash Risk for Realistic Artificial Data Generation: Application to Naturalistic Driving Study Data

Lauren Hoover
University of Central Florida

MODELING OF CRASH RISK FOR REALISTIC ARTIFICIAL DATA GENERATION: APPLICATION TO NATURALISTIC DRIVING STUDY DATA

by

LAUREN HOOVER
B.S. University of Central Florida, 2016
B.S.C.E. University of Central Florida, 2018

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2021

Major Professor: Naveen Eluru

ABSTRACT

Most safety performance analysis employs cross-sectional and time-series datasets, posing an important challenge to safety performance and crash modification analysis. The traditional safety model analysis paradigm relying on observed data only allows relative comparisons between analysis methods and is unable to establish how well the methods mimic the true underlying crash generation process. Assumptions are made about the data, but whether the assumptions truly characterize the safety data generation in the real world remains unknown. To address this issue, this thesis proposes the generation of realistic artificial data (RAD). In developing a prototype RAD generator for crash data, we mimic the process of crash occurrence, simulating daily traffic patterns and evaluating each trip for crash risk. For each crash, details such as crash location, crash type, and crash severity are also generated. As part of the artificial data generation, this thesis also proposes a framework for employing naturalistic driving study (NDS) data to understand and predict crash risk at a disaggregate trip level. This framework proposes a case-control study design for understanding trip level crash risk. The study also conducts a comparison of different case to control ratios and finds the model parameters estimated with these control ratios are reasonably similar. A multi-level random parameters binary logit model was estimated where multiple forms of unobserved variables were tested. This model was calibrated by modifying the constant parameter to generate a population conforming risk model, and then tested on a hold-out sample of data records. This thesis contributes to safety research through the development of a prototype RAD generator for traffic crash data, which will lead to new information about the underlying causes of crashes and ways to make roadways safer.

ACKNOWLEDGEMENTS

I would like to express a deep gratitude to my supervisor, Dr. Naveen Eluru, for his valuable guidance and support in the process of developing this thesis.

I would also like to gratefully acknowledge Dr. Tanmoy Bhowmik, whose support and guidance in developing this thesis have been invaluable.

I would also like to gratefully acknowledge the contributions of Dr. Naveen Eluru, Dr. Tanmoy Bhowmik, and Dr. Shamsunnahar Yasmin in the development of Chapter 3.

I would like to acknowledge the Federal Highway Administration for their financial support. I would also 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 this thesis.

Finally, I would like to express the deepest gratitude to my husband, Matthew Hoover, who has provided constant support and encouragement through the process of completing this thesis.

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES viii

CHAPTER 1: INTRODUCTION 1

Thesis Structure 5

CHAPTER 2: DEVELOPMENT OF A DISAGGREGATE REALISTIC ARTIFICIAL DATA (RAD) GENERATOR FOR TRAFFIC CRASHES 6

RAD Conceptual Framework 6

Prototype RAD Generator Development 6

RAD Generator Testing 7

CHAPTER 3: UNDERSTANDING CRASH RISK USING A MULTI-LEVEL RANDOM PARAMETER BINARY LOGIT MODEL: APPLICATION TO NATURALISTIC DRIVING STUDY DATA 11

Earlier Research 11

Data Preparation 14

Case Control Design 14

Empirical Analysis 16

Parameter Variation Across Various Samples 16

Methodological Framework 18

Model Results 21

Model Application 22

Conclusion 23

CHAPTER 4: CONCLUSION 26

APPENDIX A: CRASH GENERATION PYTHON CODE 28

APPENDIX B: IRB WAIVER 30

REFERENCES 32

LIST OF FIGURES

Figure 1: Crash Generation Conceptual Framework 6

Figure 2: Test Statistics (t-statistics) for Parameter Estimates Across Samples for each Variable 18

LIST OF TABLES

Table 1: Crash Risk Dummy Model 8

Table 2: Crash Location Dummy Model 8

Table 3: Crash Type Dummy Model 9

Table 4: Crash Type Results 9

Table 5: Crash Severity Dummy Model 10

Table 6: Crash Severity Results 10

Table 7: Summary of SHRP2 NDS Variables 15

Table 8: Crash Risk Estimates 16

Table 9: Multi-Level Random Parameters Binary Logit Model Results 21

Table 10: Comparison of Model Predictions for Crash and No Crash Testing Datasets 23

CHAPTER 1: INTRODUCTION

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 (Vision Zero Network, 2021). 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 (Vision Zero Network, 2021). 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 (Lord & Mannering, 2010; Yasmin & Eluru, 2016; Bhowmik, Rahman, Yasmin, & Eluru, 2021). The second direction of research examines factors affecting crash consequence (usually injury severity) conditional on the occurrence of a crash (Yasmin & Eluru, 2013; Marcoux, Yasmin, Eluru, & Rahman, 2018; Kabli, Bhowmik, & Eluru, 2020). 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, environmental factors on crash frequency and severity. However, the data does not allow us to examine the underlying cause of crash. Additionally, when crash frequency and severity are modeled, they are modeled using one dataset, allowing a

comparison between analysis methods, but not an understanding of the underlying crash generation process. 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 D.C. area monitoring 100 cars for about a year (Dingus, et al., 2006). More recently, another naturalistic driving study 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 (Antin, et al., 2019). 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. The NDS data allows for understanding the underlying timeline of the crash and account 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. However, while NDS data is useful in understanding the underlying cause of a crash, it still can't be used to understand the underlying process of crash generation.

To understand the underlying crash generation process, it would be useful if crash models could be tested on a large number of datasets. While real data cannot do this, artificial data could be a solution. Dr. Ezra Hauer proposed "one way to address this issue is to generate an artificial dataset i.e. to synthesize the data by making assumptions about the underlying crash generation process" (Bonneson & Ivan, 2013). This dataset, also known as Realistic Artificial Data (RAD), would allow researchers to test their models against multiple generated datasets. RAD generation has been used in multiple different fields. In *medical science*, synthetic data has been generated to simulate cancer survival data to evaluate parametric and non-parametric models (Gamel & Vogel, 1997). Synthetic data has also been used to generate time series data using only a small amount of ground truth data (Dahmen & Cook, 2019). In *data science*, artificial data has been generated to evaluate the performance of data mining procedures (Scott & Wilkins, 1999), evaluating frequent episode mining approaches employed for recovering sequential patterns (Zimmermann, 2012), and monotone ordinal data sets have been generated to be used in multi-attribute ordinal problems (Potharst, Ben-David, & van Wezel, 2009). In *education*, simulated data has been used to assess methods for evaluating school performance (Bifulco & Bretschneider, 2001). In *ecology*, data generation has been used to generate realistic plant species distributions using direct and indirect gradients to evaluate statistical methods (Austin, Belbin, Meyers, Doherty, & Luoto, 2006). In *information technology*, realistic artificial testing datasets have been generated based on real data for use in research (Syahaneim, et al., 2016) and for evaluating information analytics applications (Whiting, Haack, & Varley, 2008). In *traffic safety*, simulated data has been used in simulating

roadway intersections (Salim, Loke, Rakotonirainy, & Krishnaswamy, 2007), daily travel patterns (Ye & Lord, 2011), traffic crash data (Geedipally, Lord, & Dhavala, 2012; Cummings, McKnight, & Weiss, 2003), traffic crash sites (Lord & Kuo, 2012), traffic crash severity (Eluru, 2013), and crash modification factors (Wu, Lord, & Zou, 2015). In *travel behavior* research, generated data has been used to simulate a host of discrete choice models (Bhat, 2003; Bhat, Castro, & Khan, 2013; Paez & Scott, 2007; Bhat, Sener, & Eluru, 2010; Bhat & Sidharthan, 2011; Pinjari & Bhat, 2010; Ferdous, Eluru, Bhat, & Meloni, 2010).

From our review of earlier literature, the RAD frameworks considered are consistently single level frameworks, i.e. the underlying decision process consists of only one layer of decisions. To elaborate, in modeling crash occurrence, earlier research has related the crash occurrence to roadway geometry and traffic volume under pre-specified assumptions of what variables will influence crash occurrence (say AADT and lane width). The proposed research effort will be the first effort that will attempt the development of RAD datasets using a multi-layered decision process. Thus, it is expected to be challenging. Drawing on the earlier literature on RAD, the goal of this thesis is the development and implementation of a prototype RAD generator that mimics the true process of crash occurrence to generate a list of traffic crashes (and crash characteristics) to be used for safety model analysis. The development of a realistic data for the aforementioned framework requires substantial data processing across multiple safety datasets and is beyond the scope of a MS thesis. Hence, the current thesis has two objectives. First, we develop a software prototype development for all modules with place holder models to be estimated later. Second, using NDS data, we develop an innovative framework for crash risk at a

trip level. The prototype RAD framework proposed and tested can enhance the current state of the art in RAD generation across various domains.

Thesis Structure

The remainder of this thesis is organized as follows. Chapter 2 discusses the development and testing of the disaggregate prototype RAD generator for simulating traffic crashes and subsequent crash characteristics. Chapter 3 discusses the development of a multi-level random parameter binary logit model using NDS data to predict crash risk. Chapter 4 presents the conclusions and recommendations based on the empirical results of the study.

CHAPTER 2: DEVELOPMENT OF A DISAGGREGATE REALISTIC ARTIFICIAL DATA (RAD) GENERATOR FOR TRAFFIC CRASHES

RAD Conceptual Framework

As the first part of this study, a disaggregate prototype RAD generator was developed to simulate traffic crashes at a trip level. This generator was designed to resemble the true process of crash occurrence, as part of a trip from an origin to a destination. It considers a series of trips that simulate daily traffic patterns, evaluating each trip's risk profile based on trip level factors, demographic characteristics, roadway facility attributes, and vehicle attributes. Once a crash is determined to occur (within a stochastic framework), crash characteristics are generated including crash location, crash type and crash severity. This list of trips in which a crash occurs, along with their generated crash characteristics, are provided as output to the user. The full conceptual framework for crash generation is shown in Figure 1.

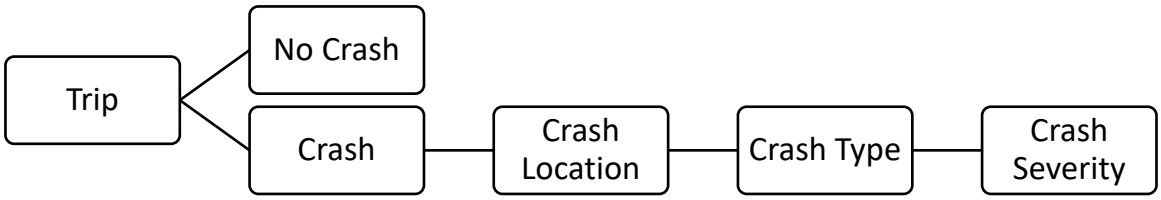

Figure 1: Crash Generation Conceptual Framework

Prototype RAD Generator Development

As described earlier, the development of a realistic data for safety analysis requires substantial data processing across multiple safety datasets. Hence, we focus on a prototype RAD with place holder models to test the software developed for RAD generation. The prototype RAD begins with a trip level file with details on trip data (such as travel distance, and overall trip level segment characteristics) and driver demographics (such as age and gender). The final desired

output is a list of crashes with crash details simulated by the RAD generator. This RAD generator hypothesized comprised of four modules – crash risk, crash location, crash type, and crash severity. The first module, crash risk, uses a binary logit model to determine if a crash occurs during a specified trip. For those trips where a crash occurs, the second module determines the location of the crash. Using the trip path as input and trip segments as the alternatives, a multinomial logit model is used to determine the segment of the trip where the crash occurs. The third module then determines the type of crash that occurs. Using a list of crash types as alternatives (such as rear-end, sideswipe, head-on, single vehicle, or non-motorized), a multinomial logit model is used to determine the type of crash that occurs. The fourth and final model determines the severity of the crash using an ordered logit model. For crash severity we use the KABCO crash injury severity model defined by The Federal Highway Administration (FHWA) (FHWA, 2011) which has five categories of crash injury severity: fatal (K), incapacitating injury (A), non-incapacitating injury (B), possible injury (C), and no injury (O). As trip records are processed, a subset of these trips is selected to be involved in a crash and the subsequent crash characteristics are generated for these trips. It is important to recognize that the sequence of the crash characteristic generation is important as the variable generated can be employed as independent variable in downstream variable generation.

RAD Generator Testing

The prototype software is developed with appropriate econometric model systems with assumed model parameters. For the crash risk module, the assumed model is shown in Table 1. In this model, driving during morning or evening peak hours increases the risk of a crash, young drivers are at an increased crash risk, senior drivers are at a decreased crash risk, and longer trips

incrementally increase crash risk. This model was applied to 2,256,502 trips and resulted in about 1,100 crashes on average. This is in agreement with the daily number of crashes that occur in Florida according to the Florida Highway Safety and Motor Vehicles (Florida Highway Safety and Motor Vehicles, 2021), which states that 403,626 crashes occurred in Florida in 2018 (about 1,106 crashes per day) and 401,868 crashes occurred in Florida in 2019 (about 1,101 crashes per day).

Table 1: Crash Risk Dummy Model

Variables	**Coefficient**
Constant	-16.000
Trip Length	0.001
Morning Peak (6am-9am)	2.300
Evening Peak (4pm-7pm)	0.970
Young Driver (<=20 years old)	0.650
Senior Driver (>=65 years old)	-0.890

Table 2 shows the model used to test the crash location module. This model considered each trip resulting in a crash as input and used a multinomial logit model to determine the road segment in the trip path where the crash occurred. The alternative set includes all segments along the trip. The average number of segments in a path was 21.8, with a minimum of 1 segment and a maximum of 685 segments. In this model, roads with a higher speed had an increased crash risk and roads with wider lanes and wider shoulders had a decreased crash risk.

Table 2: Crash Location Dummy Model

Variables	**Coefficient**
Speed	0.1
Lane Width	-0.3
Shoulder Width	-0.1

The third module uses a multinomial logit model to determine the crash type for each trip resulting in a crash. The assumed model used is shown in Table 3. This model used trip duration, lane width at crash location, shoulder width at crash location, and driver age to determine if the

crash was rear-end, sideswipe, head-on, single vehicle, or non-motorized. The results in Table 4 show that the expected probability closely matches with the resulting proportions for all crash types. Rear-end crashes are the most prevalent at about 40% of crashes, followed by sideswipes at about 20%, then head-on and single vehicle crashes at about 15% each, and non-motorized vehicle crashes were the least prevalent at about 10% of crashes.

Table 3: Crash Type Dummy Model

Variables	**Rear-end**	**Sideswipe**	**Head-on**	**Single Vehicle**	**Non-motorized**
Constant	0	-0.02	-20.9	-10.3	-7.3
Duration	0.003	0.002	0.004	0.003	0
Lane Width	0	0.6	0.22	0.43	2.17
Shoulder Width	0	0.5	0.01	0.35	0.9
Age	0	0	0	0	0.09

Table 4: Crash Type Results

Crash Type	**Probability**	**Proportion**	**30 Day Total**	**30 Day Average**
Rear-end	0.3970	0.4054	13,406	447
Sideswipe	0.2061	0.2109	6,975	232
Head-on	0.1429	0.1407	4,653	155
Single Vehicle	0.1584	0.1577	5,216	174
Non-motorized	0.0957	0.0852	2,819	94
Total	1.0000	1.0000	33,069	1102

The fourth and final module uses an ordered logit model to determine crash severity based on the KABCO injury scale. The assumed model for crash severity is shown in Table 5. In this model, crash severity is influenced by speed and crash type. Higher speeds increase severity, rear-end crashes decrease severity, and head-on and non-motorized crashes increase severity. The results in Table 6 show the proportion of crashes at each severity level, with about 54% of crashes with no injury, about 18% of crashes with possible injury, about 10% of crashes with non-incapacitating injury, about 8% of crashes with incapacitating injury, and about 10% of crashes that were fatal.

Table 5: Crash Severity Dummy Model

Propensity Variables	**Coefficient**
Speed	0.002
Rear-End Crash	-0.1
Head-On Crash	0.8
Non-Motorized Crash	0.7
Threshold between O and C	0.4
Threshold between C and B	1.2
Threshold between B and A	1.8
Threshold between A and K	2.5

Table 6: Crash Severity Results

Crash Severity	**Probability**	**Proportion**	**30 Day Total**	**30 Day Average**
O	0.5364	0.5368	17,752	592
C	0.1792	0.1818	6,013	200
B	0.1031	0.1034	3,418	114
A	0.0811	0.0786	2,600	87
K	0.1002	0.0994	3,286	110
Total	1.0000	1.0000	33,069	1102

These placeholder models were useful for testing, but to effectively use the prototype RAD generator, realistic models are needed. The prototype software works well for the placeholder models with adequate variability across different realizations. In Chapter 3 we describe how we developed a crash risk binary logit model using naturalistic driving study data which can be applied to the first module of the RAD generator.

CHAPTER 3: UNDERSTANDING CRASH RISK USING A MULTI-LEVEL RANDOM PARAMETER BINARY LOGIT MODEL: APPLICATION TO NATURALISTIC DRIVING STUDY DATA[1]

Earlier Research

This chapter 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. 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 datasets include 100-Car NDS (Klauer, Dingus, Neale, Sudweeks, & Ramsey, 2006; Guo & Fang, 2013) or the SHRP2 NDS (Dingus, et al., 2016; Owens, et al., 2018; Huisingh, et al., 2019). The dimensions affecting crash /near crash risk examined in these NDS studies include various driver behaviors such as driver inattention (Klauer, Dingus, Neale, Sudweeks, & Ramsey, 2006; Dingus, et al., 2016), glance behavior (Bärgman, Lisovskaja, Victor, Flannagan, & Dozza, 2015), aggressive/risky driving and speeding (Guo & Fang, 2013; Hamzeie, Savolainen, & Gates, 2017; Kamrani, Arvin, & Khattak, 2019; Seacrist, et al., 2020) and secondary task involvement (Huisingh, et al., 2019). Apart from the two major NDS studies, a small number of studies examined role of driver actions in crash/near crash events for commercial drivers (Hickman

[1] The contents of this chapter have been previously published in a paper accepted for presentation at the 2022 Transportation Research Board Annual Meeting. This paper by myself, Dr. Tanmoy Bhowmik, Dr. Shamsunnahar Yasmin, and Dr. Naveen Eluru is titled "Understanding Crash Risk using a Multi-Level Random Parameter Binary Logit Model: Application to Naturalistic Driving Study Data". This paper is also under consideration for publication in the Transportation Research Record. I contributed to the study conception and design, data collection, model estimation and validation, analysis and interpretation of results, and manuscript preparation.

& Hanowski, 2012), and influence of behavioral and environmental factors present prior to a crash for teenage drivers (Carney, McGehee, Harland, Weiss, & Raby, 2015).

Analysis of NDS data is conducted using two main types of case-control study designs: (a) case-cohort design and (b) case-crossover design (Guo F. , 2019). 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., (1995) suggested a 1:4 ratio for case-crossover studies. Most of the existing literature in safety employ a ratio ranging from 1:1 to 1:10. However, it is important that an examination of 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 & Kai, 1998) for a detailed discussion). The case-control model outputs can only be used to calculate the odds ratio (Mann, 2003). 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 on 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.

In this context, this chapter 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 towards understanding potential crash occurrence and/or severity. Given the rarity of crashes, even an exhaustive exercise as SHRP2 produced only 1,951 crash events from 5,512,900 trips (Hankey, Perez, & McClafferty, 2016). Hence, 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 naturalistic driving studies (See (Guo F. , 2019) for a detailed review) and real-time crash risk models developed in safety literature (Abdel-Aty & Pande, 2007; Xu, Liu, & Wang, 2016). The current research will draw on earlier case-control literature in transportation safety to customize the case control study design for our analysis.

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, 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 7. 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 adopt 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. Hence, we restricted our analyses to 914 crash trips (cases). For testing different case to control ratios, we create samples with the following case to control ratios 1:4, 1:9, 1:14, 1:19 and 1:29.

Table 7: Summary of SHRP2 NDS Variables

Variable Name	Variable Description	Min.	Max.	Mean	Std. Dev.
Driver Demographics					
Age 16-19	Driver age is between 16 and 19	0	1	0.023	0.151
Age 20-24	Driver age is between 20 and 24	0	1	0.064	0.245
Age 25-29	Driver age is between 25 and 29	0	1	0.081	0.273
Age > 74	Driver age is greater than 74	0	1	0.074	0.263
Avg. annual miles < 10,000	Driver average annual mileage of less than 10,000 mi/yr	0	1	0.229	0.420
Avg. annual miles > 25,000	Driver average annual mileage of greater than 25,000 mi/yr	0	1	0.134	0.341
Years driving	Number of years driving	0	74	33.132	17.732
Full-time worker	If full time worker, 1, else, 0	0	1	0.480	0.500
Part-time worker	If part time worker, 1, else, 0	0	1	0.190	0.392
Gender	1 if male, 0 if female	0	1	0.490	0.500
Previous Crash	1 if driver has been in a crash in the last 3 years, 0 otherwise	0	1	0.260	0.439
Trip Variables					
Distance	Straight line distance 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	100	10.497	19.566
Percent Urban	Percentage of the trip on urban roads	0	100	54.985	28.534
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 value during the trip	-1.367	3.210	0.287	0.096
Max deceleration	Maximum longitudinal deceleration value during the trip	-3.466	0.620	-0.325	0.111
Max lateral accel.	Maximum lateral acceleration value during the trip	-0.238	3.483	0.381	0.131
Max turn rate	Maximum turn rate turing the trip	344.057	399.990	26.673	10.216

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. A summary of the model estimates across all control samples is presented in Table 8. 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 if the difference across parameters for all the samples is within statistically acceptable levels.

Table 8: 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	-0.005 (0.001)	-0.005 (0.001)	-0.005 (0.001)	-0.005 (0.001)	-0.005 (0.001)
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)

* Variable insignificant at 95% significance level; ** Variable insignificant at 90% significance level

To compare the parameters across the models, we employ the 1:29 control sample as the benchmark and evaluate if the parameters for other models are statistically different relative to this sample. Towards making the comparison, a revised Wald test statistic relative to the 1:29 sample is generated as follows:

$$\text{Parameter test statistic} = abs\left[\frac{(sample\ parameter - population\ benchmark)}{\sqrt{SE_{sample}{}^2 + SE_{population}{}^2}}\right]$$

If the parameter test statistic computed is 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 are computed. Figure 2 provides a box plot summary of the variations across samples for all parameters. The figure clearly highlights 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 was 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.

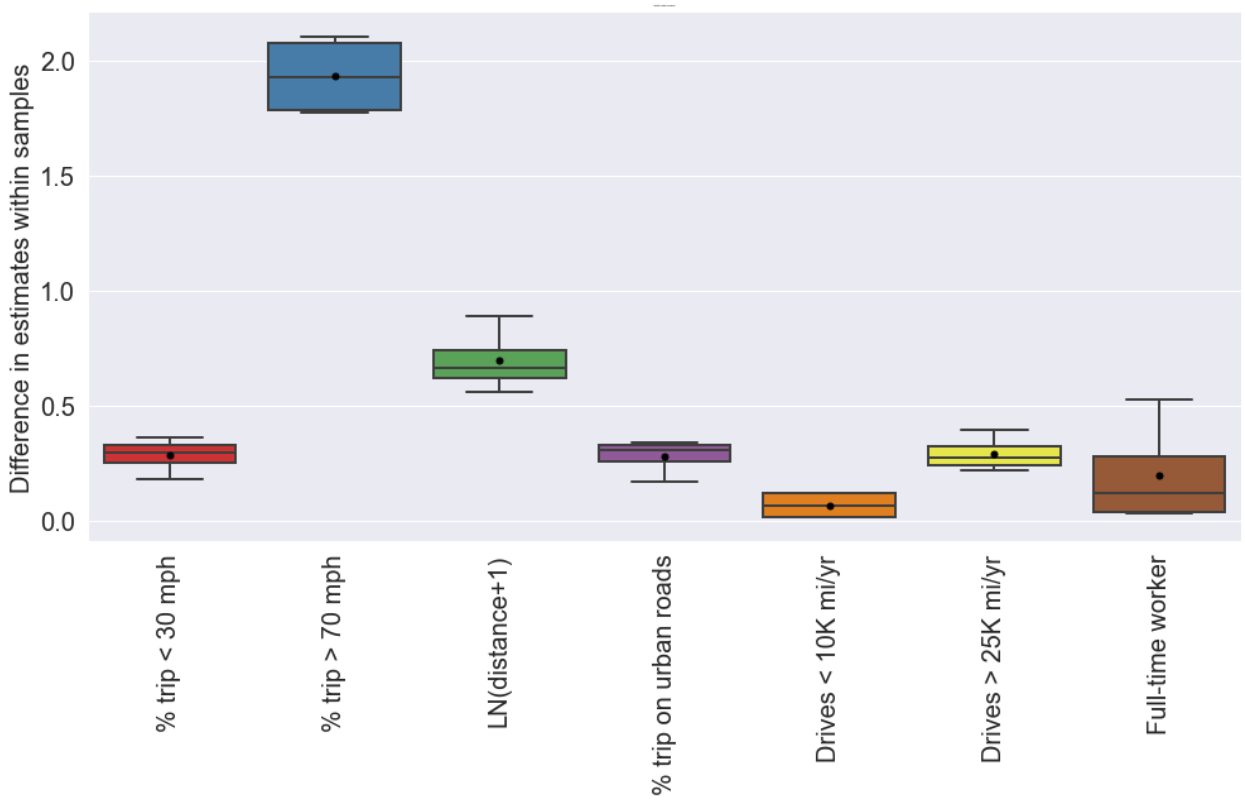

Figure 2: Test Statistics (t-statistics) for Parameter Estimates Across Samples for each Variable

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 $q(q = 1,2,3, \ldots \ldots \ldots . . m; M = 10)$ represents 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:

$$v_{iq}^{*} = \{(\alpha + \gamma_{iq})z_{iq} + \varepsilon_{iq} + \varrho_{iq}\}, v_{iq} = 1, if\ v_{iq}^{*} > 0;\ v_{iq} = 0, otherwise \quad (1)$$

where, v_{iq}^{*} represents the propensity for crash occurrence for sample q in stratum i; $v_{iq}^{*}\ is$ 1 if sample specific to a given stratum indicates crash and 0 other wise. z_{iq} is a vector attributes associated with sample q in stratum i and α is the vector of corresponding mean effects. γ_{iq} is a vector of unobserved factors affecting probability of crash occurrence. ε_{iq} is an idiosyncratic error

term assumed to be identically and independently standard logistic distributed. ϱ_{iq} is a vector of unobserved effects specific to stratum i. As highlighted earlier, within each stratum i, we matched 1 crash with 9 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 ϱ_{iq}. Further, as we used 20% margin for trip distance to match crash: non-crash, it is quite possible that the correlation across the samples might vary based on this margin. To be specific, sample with lower trip distance margin (let's say 0-5%) might exhibit stronger correlation in comparison to the sample with higher margins (like 20%). Hence, as opposed to fixing the correlation, we allow it to vary across samples by parameterizing the ϱ_{iq} term as a function of trip distance margin as follows:

$$\varrho_{iq} = \beta + \eta * trip\ distance\ margin \tag{2}$$

where, β (constant) and η are vectors of unknown parameters to be estimated. In estimating the model, it is necessary to specify the structure for the unobserved vectors $\gamma\ and\ \varrho$ represented by Ω. 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 crash takes the following form (conditional on Ω):

$$P_{iq} = p((v_{iq}^*)|(\Omega) = \frac{exp\{(\alpha + \gamma_{iq})z_{iq} + \varepsilon_{iq} + \varrho_{iq}\}}{1 + exp\{(\alpha + \gamma_{iq})z_{iq} + \varepsilon_{iq} + \varrho_{iq}\}} \tag{3}$$

The corresponding probability for non-crash is computed as

$$Q_{iq} = 1 - P_{iq} \tag{4}$$

Further, conditional on Ω, the joint probability L_i for each stratum i can be expressed as:

$$L_i = \int \left[\prod_{q=1}^{M} \{(P_{iq})^{v_{iq}} * (Q_{iq})^{(1-v_{iq})}\} \right] f(\Omega) d\Omega \tag{5}$$

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 & Bhat, 2007) for detail). For instance, if DL_i is the realization of the likelihood function in the cth draw (c = 1, 2, …, C), then the simulated log-likelihood function is as follows:

$$LL = \sum Ln\left(\frac{1}{C}\sum_{c=1}^{C}(DL_i)\right) \quad (6)$$

The parameters to be estimated in the model are: $\alpha, \gamma, \varrho, \beta, \eta, \pi \; and \; \Phi$. 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, Bhat, & Hensher, A mixed generalized ordered response midel for examining pedestrian and bicyclist injury severity level in traffic crashes, 2008; Bhat, Quasi-random maximumsimulated likelihood estimation of the mixed multinomial logit model, 2001) 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 9. A discussion of the model results follows.

Table 9: Multi-Level Random Parameters Binary Logit Model Results

Parameters	**Estimate (std. err.)**	**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	-0.005 (0.002)	-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

* Variable insignificant at 95% significance level; ** Variable insignificant at 85% significance level

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 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 hence 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 3 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 if the employment status had an impact on crash risk. The model parameter for full-time worker indicates these drivers are less at risk compared to 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 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 dataset the crash to no crash ratio was 1:4,850. In order 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 dataset of 4,500 randomly selected non-crash trips that had not been used in previous modeling and 500 randomly selected crash trips. A comparison of the results for the original and calibrated models is shown in Table 10. The results in Table 10 clearly indicate that the calibrated model captures the true ratio of crash to no crash trips.

Table 10: Comparison of Model Predictions for Crash and No Crash Testing Datasets

	Original Random Effect Model	Calibrated Random Effect Model
Probability of crash using 500 crash trip testing set	0.0534	0.0002
Probability of no crash using 4,500 no crash trip testing set	0.9466	0.9998

Conclusion

Traditional crash data has been instrumental in understanding the influence of various factors drawn from driver demographics, vehicle characteristics, roadway characteristics, crash characteristics, 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 including characteristics of the trip (trip proportion on different facilities: speed limit, roadway functional class), behavior (like eye movement) and action of the driver (actual speed of the vehicle) at the time of crash are often missing from the dataset. To that extent, the current research effort adopted the Second Strategic Highway Research Program (SHRP2) naturalistic driving study data (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 its' rarity in crash sample relative to non-crash samples (<0.01 %). Estimating a binary outcome model for such rarity will be extremely challenging. Hence, 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 4 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 if the parameters for other models are statistically different or not. The result clearly highlights the stability in parameter estimates across the samples and hence, we restrict 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 was 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, 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 offered 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, prior to generating the prediction, we calibrate the constant of the model to generate a population conforming crash risk model. Findings from the prediction exercise further reinforces the applicability of the model.

The study is not without limitations. 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 set of common factors such as, trip spend on different facilities (rural/urban), trip spend on different speed limit 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 & Zeng, 2001; Calabrese & Osmetti, 2013; Agarwal, Narasimhan, Kalyanakrishnan, & Agarwal, 2014)).

CHAPTER 4: CONCLUSION

The traditional analysis paradigm relying on observed data only allows relative comparisons between analysis methods and is unable to establish how well the methods mimic the true underlying crash generation process - often unobserved or known only partially with various degrees of uncertainty. At the same time, existing data sources and availability of data for model calibration and validation pose an important challenge to safety performance and crash modification analysis. Most safety performance analysis employs cross-sectional and time-series datasets. Assumptions are made about the data, but whether the assumptions truly characterize the safety data generation in real world remains unknown. To address this issue, this thesis proposes the generation of an artificial dataset based on a stochastic but well-defined data generation process. As part of the artificial data generation, this thesis also proposes a framework for employing NDS data to understand and predict crash risk at a disaggregate trip level.

In this thesis we first propose a conceptual framework for realistic crash data generation that mimics the true process of crash occurrence. A series of trips simulate daily traffic patterns, and each trip is evaluated for crash risk. Once a crash is established to occur, crash details such as crash location, crash type, and crash severity are generated. Given the complexity and data processing challenges with generating models, the software was coded assuming place holder models for crash risk, crash location, crash type, and crash severity. As a second part of my thesis, we propose a framework for predicting crash risk (first module) using NDS data. This framework proposes a case-control study design for understanding trip level crash risk, matching crash and non-crash trips based on driver age, driver gender, and trip distance within a 20% margin. In this study we vary the number of controls, conducting a revised Wald test statistic test on control samples of 1:4, 1:9, 1:14, 1:19, and 1:29, employing the 1:29 control sample as the benchmark.

Since there is stability in parameter estimates across the samples, the 1:9 sample is used in estimating a multi-level random parameters binary logit model. In estimating the model, several variables were found to have a significant impact on crash risk propensity, including trip distance, trip proportion of different speed limit roads and facilities, driver's driving characteristics and employment status. In developing this model, multiple forms of unobserved variables were also tested, including common unobserved effects for each case-control panel, common unobserved factors affecting the error margin in the trip distance variable, and random effects for all independent variables. However, the only random effect parameter that offered statistically significant results was for the full-time worker variable, indicating 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. This model was calibrated by modifying the constant parameter to generate a population conforming risk model, and then tested on a hold-out sample of data records.

This thesis contributes to safety research through the development of a prototype RAD generator for traffic crash data, which will lead to new information about the underlying causes of crashes and ways to make our roadways safer. In future research, realistic models for other modules will need to be developed and then embedded within the prototype simulator.

APPENDIX A:
CRASH GENERATION PYTHON CODE

```
# Generate list of crashes for the number of days requested
# Input:  Number of days to test, list of trips, dictionary of crash risk model coefficients
#           (int, list[Dictionary], Dictionary)
# Output: List of crashes for each day tested
#            (list[list[Dictionary]])
def generate_crashes(numRuns, tripsList, crashCoeff):

    # Create list of list of crashes to output
    crashList = [[] for i in range(numRuns)]

    # For each day tested
    for i in range(numRuns):

        # Set random seed for crashes
        random.seed(100000+i)

        # For each trip in list of trips
        for record in tripsList:

            # Generate random number between 0 and 1
            rand = random.random()

            # Determine utility value of trip
            record["util"] = util.util_calc(crashCoeff, record)

            # Calculate probability of no crash based on trip data
            prob = 1 – (1 / (1 + math.exp(-(record["util"]))))

            # If rand is greater than crash probability, crash; else, no crash
            if rand > prob:
                crash = 1
            else:
                crash = 0

            # If crash, add record to crash list
            if crash:
                crashRecord = record.copy()
                crashList[i].append(crashRecord)

    # Return list of crashes
    return crashList
```

APPENDIX B:
IRB WAIVER

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

NOT HUMAN RESEARCH DETERMINATION

October 18, 2021

Dear Lauren Hoover:

On 10/18/2021, the IRB reviewed the following protocol:

Type of Review:	Initial Study
Title of Study:	Modeling of Crash Risk for Realistic Artificial Data Generation: Application to Naturalistic Driving Study Data
Investigator:	Lauren Hoover
IRB ID:	STUDY00003505
Funding:	Name: US Dept of Transp Fed Highway Adm (FHA)
Grant ID:	
Documents Reviewed:	• HRP-251- FORM - Faculty Advisor Scientific-Scholarly Review fillable form.pdf, Category: Faculty Research Approval; • Data Abstraction Form, Category: Other; • HRP-250-FORM- Request for NHSR.docx, Category: IRB Protocol

The IRB determined that the proposed activity is not research involving human subjects as defined by DHHS and FDA regulations.

IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities are research involving human in which the organization is engaged, please submit a new request to the IRB for a determination. You can create a modification by clicking **Create Modification / CR** within the study.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Katie Kilgore
Designated Reviewer

REFERENCES

Abdel-Aty, M., & Pande, A. (2007). Crash data analysis: Collective vs. individual crash level approach. *Journal of Safety research, 38*, 581-587.

Agarwal, A., Narasimhan, H., Kalyanakrishnan, S., & Agarwal, S. (2014). GEV-Canonical Regression for Accurate Binary Class Probability Estimation when One Class is Rare. *Proceedings of the 31 st International Conference on Machine Learning. 32*, pp. 1989-1997. Beijing, China: JMLR.

Antin, J. F., Lee, S., Perez, M. A., Dingus, T. A., Hankey, J. M., & Brach, A. (2019). Second strategic highway research program naturalistic driving study methods. *Safety Science*, 2-10.

Austin, M. P., Belbin, L., Meyers, J. A., Doherty, M. D., & Luoto, M. (2006). Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecological Modelling, 199*, 197-216.

Bärgman, J., Lisovskaja, V., Victor, T., Flannagan, C., & Dozza, M. (2015). 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, 35*, 152-169.

Bhat, C. R. (2001). Quasi-random maximumsimulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological, 35*, 677-693.

Bhat, C. R. (2003). Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences. *Transportation Research Part B, 37*(9), 837-855.

Bhat, C. R., & Sidharthan, R. (2011). A Simulation Evaluation of the Maximum Approximate Composite Marginal Likelihood (MACML) Estimator for Mixed Multinomial Probit Models. *Transportation Research Part B, 45*(7), 940-953.

Bhat, C. R., Castro, M., & Khan, M. (2013). A New Estimation Approach for the Multiple Discrete-Continuous Probit (MDCP) Choice Model. *Transportation Research Part B, 55*, 1-22.

Bhat, C. R., Sener, I. N., & Eluru, N. (2010). A Flexible Spatially Dependent Discrete Choice Model: Formulation and Application to Teenagers' Weekday Recreational Activity Participation. *Transportation Research Part B, 44*(8-9), 903-921.

Bhowmik, T., Rahman, M., Yasmin, S., & Eluru, N. (2021, September). Exploring Analytical, Simulation-Based, And Hybrid Model Structures For Multivariate Crash Frequency Modeling. *Analytic Methods in Accident Research, 31*.

Bifulco, R., & Bretschneider, S. (2001). Estimating school efficiency: A comparison of methods using simulated data. *Economics of Education Review, 20*, 417-429.

Bonneson, J., & Ivan, J. (2013). Theory, explanation, and prediction in road safety promising directions. *Transportation Research Circular.* Washington, D.C.: Transportation Research Board of the National Academies.

Calabrese, R., & Osmetti, S. A. (2013). Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics*, *40*(6), 1172-1188.

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

Cummings, P., McKnight, B., & Weiss, N. S. (2003). Matched-pair cohort methods in traffic crash research. *Accident Analysis & Prevention, 35*(1), 131-141.

Dahmen, J., & Cook, D. (2019). SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors, 19*(5), 1181.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the United States of America, 113*(10), 2636-2641.

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., . . . Knipling, R. R. (2006). *The 100-Car Naturalistic Driving Study: Phase II – Results of the 100-Car Field Experiment.* U.S. Department of Transportation, National Highway Traffic Safety Administration. Springfield, VA: National Technical Information Service.

Eluru, N. (2013). Evaluating Alternate Discrete Choice Frameworks for Modeling Ordinal Discrete Variables. *Accident Analysis & Prevention, 55*(1), 1-11.

Eluru, N., & Bhat, C. R. (2007). A Joint econometric analysis of seat belt use and crash-related injury severity. *Accident Analysis and Prevention, 39*, 1037-1049.

Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response midel for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis and Prevention, 40*, 1033-1054.

Ferdous, N., Eluru, N., Bhat, C. R., & Meloni, I. (2010). A Multivariate Ordered Response Model System for Adults' Weekday Activity Episode Generation by Activity Purpose and Social Context. *Transportation Research Part B, 44*(8-9), 922-943.

FHWA. (2011, July 15). *Highway Safety Improvement Program Manual*. Retrieved from Federal Highway Administration: https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec4.cfm

Florida Highway Safety and Motor Vehicles. (2021, April 25). *Florida Crash Dashboard*. Retrieved from Florida Highway Safety and Motor Vehicles: https://www.flhsmv.gov/traffic-crash-reports/crash-dashboard/

Gamel, J. W., & Vogel, R. L. (1997). Comparison of Parametric and Non-Parametric Survival Methods Using Simulated Clinical Data. *Statistics in Medicine, 16,* 1629-1643.

Geedipally, S. R., Lord, D., & Dhavala, S. S. (2012). The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis & Prevention, 45*, 258-265.

Guo, F. (2019). Statistical methods for naturalistic driving studies. *Annual review of statistics and its application, 6*, 309-328.

Guo, F., & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. *Accident Analysis and Prevention, 61*, 3-9.

Hamzeie, R., Savolainen, P. T., & Gates, T. J. (2017). Driver speed selection and crash risk: Insights from the naturalistic driving study. *Journal of safety research, 63,* 187-194.

Hankey, J. M., Perez, M. A., & McClafferty, J. A. (2016). *Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets.* Blacksburg, VA: Virginia Tech Transportation Institute.

Hickman, J. S., & Hanowski, R. J. (2012). An Assessment of Commercial Motor Vehicle Driver Distraction Using Naturalistic Driving Data. *Traffic Injury Prevention, 13*(6), 612-619.

Hoover, L., Bhowmik, T., Yasmin, S., & Eluru, N. (2022). Understanding Crash Risk using a Multi-Level Random Parameter Binary Logit Model: Application to Naturalistic Driving Study Data. *Transportation Research Board Annual Meeting.* Washington D.C.

Huisingh, C., Owsley, C., Levitan, E. B., Irvin, M. R., MacLennan, P., & McGwin, G. (2019). 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, 74*(4), 550-555.

Kabli, A., Bhowmik, T., & Eluru, N. (2020). A Multivariate Approach For Modeling Driver Injury Severity By Body Region. *Analytic Methods in Accident Research, 28*.

Kamrani, M., Arvin, R., & Khattak, A. J. (2019). The Role of Aggressive Driving and Speeding in Road Safety: Insights from SHRP2 Naturalistic Driving Study Data. *Transportation Research Board 98th Annual Meeting.* Washington DC.

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis, 9*(2), 137-163.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data.* U.S. Department of Transportation, National Highway Traffic Safety Administration. Springfield, Virginia: National Technical Information Service.

Lord, D., & Kuo, P. F. (2012). Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures. *Accident Analysis & Prevention, 47*, 52-63.

Lord, D., & 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), 291-305.

Mann, C. (2003). Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal, 20*, 54-60.

Marcoux, R., Yasmin, S., Eluru, N., & Rahman, M. (2018). 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, 20*, 15-29.

Mittleman, M. A., Maclure, M., & Robins, J. M. (1995). Control sampling strategies for case-crossover studies: an assessment of relative efficiency. *American Journal of Epidemiology, 142*, 91-98.

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

Paez, A., & Scott, D. M. (2007). Social influence on travel behavior: a simulation example of the decision to telecommute. *Environment and Planning A, 39*(3), 647-665.

Pinjari, A. R., & Bhat, C. R. (2010). A Multiple Discrete-Continuous Nested Extreme Value (MDCNEV) Model: Formulation and Application to Non-Worker Activity Time-Use and Timing Behavior on Weekdays. *Transportation Research Part B, 44*(4), 562-583.

Potharst, R., Ben-David, A., & van Wezel, M. (2009). Two algorithms for generating structured and unstructured monotone ordinal data sets. *Engineering Applications of Artificial Intelligence, 22*, 491-496.

Salim, F. D., Loke, S. W., Rakotonirainy, A., & Krishnaswamy, S. (2007). Simulated intersection environment and learning of collision and traffic data in the u&i aware framework. *International Conference on Ubiquitous Intelligence and Computing*, (pp. 153-162). Springer, Berlin, Heidelberg.

Scott, P. D., & Wilkins, E. (1999). Evaluating data mining procedures: techniques for generating artificial data sets. *Information and Software Technology, 41*, 579-587.

Seacrist, T., Douglas, E. C., Hannan, C., Rogers, R., Belwadi, A., & Loeb, H. (2020). Near crash characteristics among risky drivers using the SHRP2 naturalistic driving study. *Journal of safety research, 73*, 263-269.

Syahaneim, Hazwani, R. A., Wahida, N., Shafikah, S. I., Zuraini, & Ellyza, P. N. (2016). Automatic Artificial Data Generator: Framework and Implementation. *International Conference on Information and Communication Technology (ICICTM)* (pp. 56-60). Kuala Lumpur, Malaysia: IEEE.

Vision Zero Network. (2021, June 16). *Vision Zero Communities*. Retrieved from Vision Zero Network: https://visionzeronetwork.org/resources/vision-zero-communities/

Whiting, M. A., Haack, J., & Varley, C. (2008). Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. *BELIV '08: Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization* (pp. 1-9). Florence, Italy: Association for Computing Machinery.

Wu, L., Lord, D., & Zou, Y. (2015). Validation of crash modification factors derived from cross-sectional studies with regression models. *Transportation research record, 2514*(1), 88-96.

Xu, C., Liu, P., & Wang, W. (2016). Evaluation of the predictability of real-time crash risk models. *Accident Analysis and Prevention, 94*, 207-215.

Yasmin, S., & Eluru, N. (2013). Evaluating Alternate Discrete Outcome Frameworks for Modeling Crash Injury Severity. *Accident Analysis and Prevention, 59*(1), 506-521.

Yasmin, S., & Eluru, N. (2016). Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. *Accident Analysis and Prevention, 95*, 157-171.

Ye, F., & Lord, D. (2011). Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit. *Transportation Research Record, 2241*(1), 51-58.

Zhang, J., & Kai, F. Y. (1998). What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *Jama, 280*, 1690-1691.

Zimmermann, A. (2012). Generating Diverse Realistic Data Sets for Episode Mining. *12th IEEE International Conference on Data Mining Workshops, ICDMW* (pp. 611-618). Brussels, Belgium: IEEE.