

Role of Human Factors, Driving Instability, and Roadway Environment in Safety Critical Events: Safe System Approach

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Numan Ahmad

December 2021

Copyright ©2021 by Numan Ahmad

All rights reserved.

This dissertation is dedicated to my parents, my beloved family, and my professors who have always been a source of encouragement in this long journey to accomplish my doctoral studies

ACKNOWLEDGMENTS

First and most of all, I am very thankful to Almighty Allah (for His blessings) Who enabled me to complete my Ph.D. studies. Without His guidance, it would have not been possible for me to complete my dissertation and my dream of accomplishing a doctoral degree would have been merely a dream. This is Him who bestowed me with the opportunity, strength, ability, and patience to accomplish my doctoral degree.

I wish to extend my sincere gratitude to my professor Dr. Asad J. Khattak who has been a source of learning, motivation, and guidance during the four years of my Ph.D. studies. This was his powerful mentorship and support which have enabled me to contribute to the research program at the University of Tennessee Knoxville. Being an international student, I must acknowledge and sincerely appreciate Prof. Khattak for his fatherly affection towards all his students including me due to which I never felt being away from my family. It has been an utmost honor for me to be your student and I shall always take pride in my association with you.

Also, it has been an immense honor and pleasure to appreciate my dissertation committee members, Prof. Hamparsum Bozdogan, Dr. Candace Brakewood, and Dr. Shau Li for their tremendous guidance, help, and support in my dissertation. Also, I am especially indebted to Prof. Hamparsum Bozdogan to serve as a committee member for my M.S. Statistics degree.

I would also like to acknowledge all my Professors some of whom include Ahmad Jan, Dr. Anwaar Ahmed, Dr. Candace Brakewood, Dr. Lee D. Han, Dr. Airton Kohls, Dr. Muhammad Irfan, Dr. Akhtar Ali Shah, and Dr. Arshad Hussain. To my co-authors, Dr. Asad J. Khattak, Dr. Behram Wali, Dr. Eric Dumbaugh, and Dr. Arvin Ramin, thank you for your valuable knowledge and insights. I would also like to thank all my lab seniors and colleagues including Dr. Behram Wali, Dr. Alexandra Boggs, Dr. Armin Ramin, Dr. Xiaobing Li, Dr. Mohsen Kamrani, Dr. Meng Zhang, Abdul Rashid Mussah, Dr. Nima Hoseinzadeh, Iman Mahdinia, Amin MohammadNazar, Steve Lee, Antora Mohsena Haque, and A. Latif Patwary for their support and memorable experiences. Special thanks to the Department of Civil and Environmental Engineering faculty and staff. My research work was greatly supported by the Collaborative Sciences Center for Road Safety, the US Department of Transportation, and the Tennessee Department of Transportation.

In addition, I am sincerely grateful to my Professor, Dr. Anwaar Ahmed through whom I stepped into the field of research during my M.S. at the National University of Sciences and Technology, Pakistan. He has been a source of tremendous support and encouragement to date. It would be inappropriate to acknowledge the support of all my dear friends some of whom include Dr. Izaz Ali, Sanaullah Saif, Tufail Muhammad,

Muhammad Shoaib, Muhammad Arif, Aimal Khan, Dr. Tahir, and my roommates in the U.S. Also, it would be an injustice to acknowledge the support of my dearest friend Dr. Behram Wali who has been a credible guide both in social and professional endeavors. It has always been a tremendous learning experience for me to work with you – thank you, Dr. Wali. As your junior and colleague, I would not hesitate to mention that your fine mind and research skills have always impressed me.

My acknowledgment would be incomplete without thanking my biggest pillar of unconditional support and guidance, my parents and family (Kamran Jan, Imran Khan, Farhan Ahmad, Mr. & Mrs. Kazim Jan, and Fatima). I wish to strongly acknowledge my parents, aunt, and my siblings for their involvement, concern, and unconditional support. I wish my uncle (late Muhammad Tahir Khan) and grandfather (late Muhammad Mumtaz Khan) could be alive to see their dream coming true. Finally, I wish to acknowledge and thank my fiancée (Ms. H. Tahir Khan) for her prayers and best wishes.

ABSTRACT

Human factors including driver behavior and performance are highly relevant to the safe systems approach as these factors tend to be predominant causes of road crash occurrence. Transportation safety can be enhanced by applying the safe systems approach to harness new forms of large-scale naturalistic driving study data. The naturalistic driving study data, collected through the 2nd Strategic Highway Research Program, include real-world microscopic details about pre-crash driving behaviors, performance, instantaneous vehicular movements (speed, acceleration, and deceleration), and distracted driving (along with its duration) which were collected via recording equipment in the instrumented vehicles. One of the unique aspects of the naturalistic driving data is that it includes data about baselines (non-event driving) and near-crashes where the earlier ones could be used as a surrogate to the exposure to computing the risk of a crash or near-crash. To examine the role of human factors and improve crash investigations, this dissertation first develops a systematic taxonomy of driving errors and violations using a subsample of naturalistic driving data. Next, after developing a taxonomy of naturalistic driving errors and violations, and deriving measures of driving volatility, this dissertation investigates the role of driving errors, violations, duration of distraction, and driving volatility in safety-critical events (crashes and near-crashes) in diverse roadway environments. To unveil the aforementioned key relationships, more robust path analytic approaches including the ones via joint estimation are used to account for potential correlation between the unobserved factors in the two models. From the prediction standpoint, this dissertation applies machine learning and artificial intelligence methods to understand the significance of human factors (driving errors, violations, duration of distraction, and instability in driving) compared to roadway environmental factors in predicting safety-critical events. Finally, the dissertation looks at the effects of errors, violations, and impairment by both pedestrians and drivers on pedestrian injuries using comprehensive data of North Carolina Pedestrian-vehicle crashes coded by professional data experts. One of the unique methodological aspects of this dissertation is to account for the unobserved heterogeneity in pedestrian injuries and in the effects of observed variables (e.g., light conditions and impairment) on pedestrian injuries due to some unobserved factors where the later ones could be used to classify pedestrians into latent (unobserved) classes using finite mixture models. Overall, this dissertation contributes by providing firm knowledge and foundation about the role of human factors in road crashes which could help in selecting proactive countermeasures thus enhancing the overall safety. Based on the findings, this dissertation discusses the practical implications and discusses potential countermeasures including traditional roadway (built-environment) changes, policy interventions, vehicle technologies (e.g., adaptive cruise control, automatic braking system, dilemma zone mitigation systems, and collision warning systems) which have the potential to proactively reduce driving errors, violations, and driving instability thus

mitigating the chance of safety-critical events. From the automated vehicles standpoint, this research discusses the potential procedure which can be used to predict driving errors, violations, instability in driving, and especially safety-critical events in real-time based on human factors in diverse roadway environments.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION	1
Summary	2
Problem Justification	10
Literature Review, Gaps, and Contributions.....	12
Developing Taxonomy of Naturalistic Driving Errors and Violations	12
Exploring Effects of Roadway Environment on Driving Errors, Violations, and Crashes.....	14
Quantifying the Effects of Driving Errors, Violations, and Duration of Distraction on SCES Through Instability in Driving.....	17
Predicting Safety-Critical Events Using Driver Behaviors and Performance: Application of Artificial Intelligence.....	21
Understanding Effects of Human Factors on Pedestrian Injury Severity Given a Crash.....	23
Methodological Framework.....	26
Unique Datasets and Deriving Human Factors.....	28
SHRP2-NDS Data: Deriving Human Factors	28
NCPBCAT Data: Deriving Human Factors.....	32
CHAPTER 2. A TAXONOMY OF DRIVING ERRORS AND VIOLATIONS: EVIDENCE FROM THE NATURALISTIC DRIVING STUDY	37
Abstract.....	38
Introduction.....	39
Methodology	40
Data Source.....	40
Safety Matrix	43
A Systematic Taxonomy for Driving Errors and Violations.....	45
Diverse Roadway Localities: NDS Data.....	50
Results and Discussion	52
Safety Matrix	52
Prevalence of Driving Errors in Safety-Critical Events: Evidence from NDS Data.....	55
Distribution of Safety-Critical Events across Various Roadways and Environments	57
Variation of Driving Errors across Various Roadways and Environments	58
Countermeasures to Reduce Driving Errors and Violations	63
Limitations	70
Conclusions.....	70

CHAPTER 3. BUILT ENVIRONMENT, DRIVING ERRORS AND VIOLATIONS, AND CRASHES IN NATURALISTIC DRIVING ENVIRONMENT	73
Abstract	74
Introduction	74
Research Objectives and Contribution	77
Methodology	78
Data Source	78
Systematic Classification of Driver Errors and Violations – NDS Data	80
Diverse Roadway Environment and Localities	81
Path Analysis Framework	81
Results	86
Descriptive Statistics	86
Modeling Results	87
Discussion	98
Limitations and Future Directions	101
Conclusions	102
CHAPTER 4. EXPLORING PATHWAYS FROM DRIVING ERRORS, VIOLATIONS, AND DURATION OF DISTRACTION TO CRASHES: THE ROLE OF INSTABILITY IN DRIVING	105
Abstract	106
Introduction	107
Literature Review	108
Driving Errors and Violations, Driving Instability, and SCEs	108
Duration of Distraction, Driving Instability, and SCEs	109
Study Scope and Contribution	111
Methodology	112
Conceptual Framework	112
Modeling Framework	112
Data	118
Data Source and Systematic Classification of Driving Errors	118
Descriptive Statistics	121
Results	122
Separate Estimation: Speed Volatility Model and Event Outcome Model	122
Joint Estimation (Systems of Equations) for Speed Volatility and Event Outcome	127
Path Analysis Results: Joint Estimation	127

Discussion	131
Practical Applications	138
Limitations	138
Conclusions.....	141
CHAPTER 5. PREDICTING SAFETY-CRITICAL EVENTS USING DRIVER BASED ON DRIVER BEHAVIORS AND PERFORMANCE: APPLICATION OF MACHINE LEARNING METHODS	143
Abstract.....	144
Introduction and Literature Review	144
Methodology	147
Study Framework.....	147
Ordered Probit Regression.....	149
Naïve Bayes Classifier	150
K Nearest Neighbor Classifier	150
Gradient Boosting Trees Classification	151
Performance Measures.....	151
Data	152
Data Source.....	152
Human Factors	154
Results.....	155
Descriptives.....	155
Empirical Results	155
Discussion	165
Limitations	166
Conclusions.....	168
CHAPTER 6. THE ROLE OF HUMAN FACTORS IN PEDESTRIAN INJURIES GIVEN A CRASH: APPLICATION OF FINITE MIXTURE MODEL	170
Abstract.....	171
Introduction.....	171
Methodology	173
Data	173
Conceptual and Methodological Framework.....	174
Methods.....	176
Classification of Human Factors and Descriptive Statistics	179
Classification of Human Behaviors	179

Descriptive Statistics.....	179
Results.....	182
Ordered Probit Model	182
Finite Mixture Ordered Probit Model	184
Discussion	184
Practical Implications.....	190
Limitations	192
Conclusions.....	192
CHAPTER 7. OVERALL RESEARCH CONTRIBUTION AND PRACTICAL IMPLICATIONS	194
CHAPTER 8. CONCLUSIONS	201
REFERENCES	206
VITA.....	221

LIST OF TABLES

Table 2. 1 Definition of Various Roadway Localities in the NDS Data (Source: NDS Data Dictionary) .	51
Table 2. 2 Safety Matrix: Contribution of Human, Vehicle, and Roadway Factors	54
Table 2. 3 Prevalence of Driving Errors in Safety-Critical Events.....	56
Table 2. 4 Distribution of Crashes, Near-Crashes, and Baselines across Different Roadways and Environments	59
Table 2. 5 ITS Countermeasures Versus Different Driving Errors and Violations	68
Table 3. 1 Summary of Previous Taxonomies, and Types of Driving Errors and Violations	76
Table 3. 2 Descriptive Statistics (Row and column percentages of Baselines and SCEs).....	88
Table 3. 3 Driver Error Model Results (Multinomial Logit Model).....	90
Table 3. 4 Driver Error Model Results (Multinomial Logit Model): Marginal Effects for Only Significant Variables	91
Table 3. 5 Ordered Probit Models with Different levels of Proportional Odds Assumptions	93
Table 3. 6 Crash Outcome Model Results (Ordered Probit Models: Accounting for Proportional Odds Assumptions)	94
Table 3. 7 Direct, Indirect, and Total Effects of Explanatory Variables (Path Analysis).....	95
Table 4. 1 Relevant Literature: The Role of Driving Volatility in Roadway Safety	110
Table 4. 2 Descriptive Statistics of Key Variables	124
Table 4. 3 Results of Speed Volatility Model and Event Outcome Model (Separate Estimation)	125
Table 4. 4 Results of Speed Volatility Model and Event Outcome Model (Jointly Estimated)	128
Table 4. 5 Path Analysis Results (Based on Joint Estimation)	130
Table 4. 6 Path Analysis Results (Based on Joint Estimation): Effects of Duration of Distraction on SCEs	133
Table 5. 1 Sample Confusion Matrix for Baselines, Near-Crashes, and Crashes	153
Table 5. 2 Descriptive Statistics of Key Variables (N = 6,466).....	156
Table 5. 3 Estimation Results of the Ordered Probit Model	158
Table 5. 4 Comparing NB Classifiers: Default Prior, Informative Prior, and Optimal Kernel Width.....	159
Table 5. 5 Confusion Matrices for Ordered Probit Model and NB, KNN, and GBT Classifiers.....	164
Table 5. 6 Comparing Overall and Class-level Out-of-Sample Prediction Accuracy	164
Table 6. 1 Descriptive Statistics of Key Variables	181
Table 6. 2 Results of Simple Ordered Probit (Baseline) Model	183
Table 6. 3 Results of Finite Mixture Ordered Probit Model	185
Table 6. 4 Probabilities of Two Classes: Postestimation Results FMOPM	188
Table 6. 5 Mean Posterior Probabilities of Pedestrian Injuries: Base versus FMOPM Models	188

Table 7. 1 Key Relationships of Human Factors and Roadway Environments with SCEs: Insights from SHRP2-NDS data.....	200
Table 7. 2 Potential Countermeasures: Improving Elements of Safe Systems	200

LIST OF FIGURES

Figure 1. 1 Safe System Approach: Principles and Elements (Source: FHWA)	3
Figure 1. 2 Overall Conceptual Framework of the Proposal	8
Figure 1. 3 Proposal Outline	9
Figure 1. 4 Model for Human Information Processing	13
Figure 1. 5 Proposal Methodological Framework	27
Figure 1. 6 Illustration of SHRP2-NDS Video: Capturing Driver Distraction and Gaze (Source: (Campbell 2012))	33
Figure 1. 7 Pre-Crash Risky Behaviors by Pedestrians: Coded in Crash Type Variable (NCPBCAT).....	34
Figure 1. 8 Pre-Crash Risky Behaviors by Drivers: Coded in Crash Type Variable (NCPBCAT).....	34
Figure 1. 9 Proposed framework for future study	36
Figure 2. 1 Evolution of Driving Errors and Violations in Space-Time Dimension	46
Figure 2. 2 Systematic Taxonomy for Driver Errors and Violations	47
Figure 2. 3 Contribution of Human, Vehicle, and Roadway Factors in Crashes	54
Figure 2. 4 Driving Errors and Violations on Interstates and Bypass/Divided Highways with Traffic Signals (in Crashes)	59
Figure 2. 5 Driving Errors and Violations in Business/Industrial and Moderate Residential Areas (in Crashes)	61
Figure 2. 6 Driving Errors and Violations in School Zones (in Crashes)	61
Figure 2. 7 Systematic Classification for Roadway Localities	62
Figure 3. 1 Methodological Framework	82
Figure 3. 2 Direct and Indirect (through Driving Errors and Violations) Effects of Roadways and Environments Crash Occurrence	97
Figure 4. 1 Conceptual Framework: Path Analysis through Joint Estimation	113
Figure 4. 2 Modified Version of Systematic TDEV “Source: (Khattak <i>et al.</i> 2021a)”	120
Figure 4. 3 Distribution of Drivers based on COV of Speed in Baselines, Near-Crashes, and Crashes...	123
Figure 4. 4 Distribution of Baselines and SCEs on Secondary Task Duration (seconds).....	123
Figure 4. 5 Pathways to SCEs through COV of Speed (based on Joint Estimation Analysis)	132
Figure 4. 6 Predicted Instability in Driving versus Duration of Distraction	136
Figure 4. 7 Probabilities of Safety-Critical Events versus Distraction Duration	136
Figure 4. 8 Probabilities of Safety-Critical Events versus Instability in Driving	137
Figure 4. 9 Illustration of Key Findings through Infographics	139
Figure 4. 10 Potential Strategies and technologies for reducing driving distraction and instability.....	140
Figure 5. 1 Study Framework	148

Figure 5. 2 Search for Minimum Objective Function (Optimization for NB)	159
Figure 5. 3 Minimum Objective Function vs. Kernel Width (Optimization for NB Classifier)	159
Figure 5. 4 Posterior Predicted Probabilities versus Driving Instability and Distraction Duration	160
Figure 5. 6 Minimum Objective Function vs. Number of Nearest Neighbors (Optimization of KNN) ...	162
Figure 5. 7 Overall Classification Accuracy versus Tuning Parameters for the GBT Classifier	163
Figure 5. 8 Feature Importance Plot for the GBT Classifier	167
Figure 6. 1 Conceptual Framework.....	175
Figure 6. 2 Classifying Pre-Crash Behaviors of Drivers and Pedestrians in Crashes	180
Figure 6. 3 Probability Distribution of Pedestrians (or Crashes): Two Latent Classes	188
Figure 6. 4 Posterior Probabilities Distribution of Various Pedestrian Injury Severity Levels	189
Figure 6. 5 Potential Countermeasures for Pedestrian-vehicle Crashes.....	191

CHAPTER 1. INTRODUCTION

Summary

The safe system approach requires transportation experts to consider transportation systems more holistically. Based on the approach, the safety of road users could be substantially improved by providing various protective layers including safe road users, safe vehicles, safe speeds, safe roads, and appropriate post-crash care systems (Figure 1.1). The motivation behind the multi-layered protection systems for road users relates to the six principles which include: 1). Humans make mistakes, 2). Death/serious injuries are unacceptable, 3). Humans are vulnerable, 4). Responsibility is shared, 5). Safety is proactive, and 6). Redundancy is crucial (Figure 1.1). Based on the principles, road user safety could be substantially improved through the provision of tolerating, human-centric, proactive, and redundant roadway systems which could tolerate human mistakes, accommodate human vulnerabilities, understand and reduce unobserved crash risk, and strengthen the overall system respectively. In terms of elements, the key goal of this dissertation is to provide firm knowledge and foundation which could help in providing/improving some of the protective layers including safe road users, safe vehicles, safe speeds, and safe road (environments). Human factors (driving errors and violations) are highly relevant to the safe systems approach as they tend to be the predominant cause of crash occurrence contributing to more than 90 percent of overall road crashes. Understanding driver behavior, performance, and decisions can help in identifying more effective strategies to enhance safety. Transportation safety can be enhanced by applying a safe systems approach to harness new forms of large-scale data.

Real-world driving behavior and performance in time-to-collision can be examined using data from numerous geo-coded (location-based) driving parameters such as pre-crash driving behaviors, distractions (and their durations), speed, acceleration, and eye glance behavior of the subject (in instrumented vehicles) drivers being continuously captured via advanced equipment. Naturalistic Driving Study (NDS) data, collected as a part of the 2nd Strategic Highway Research Program (SHRP2), includes all of the aforementioned variables which make it more unique and reliable especially for the analysis of driving behavior and performance. Sophisticated methods including frequentist techniques, artificial intelligence, and machine learning techniques can be utilized to harness new microscopic level data including SHRP2-NDS and connected and automated vehicles (CAVs) data to generate new knowledge which could help in identifying more effective countermeasures (compared to traditional police crash data) to improve safety. For instance, SHRP2-NDS data include information about real-world pre-crash driving behaviors which could provide a more realistic understanding of risky human factors contributing to safety-critical events (SCEs). This could not be possible using traditional police crash data which could be subjective as the reporting officer may not be present at the time of crash occurrence. Furthermore, no information about baselines is available in police crash data which can be used to quantify crash or near-crash risk.



Figure 1. 1 Safe System Approach: Principles and Elements (Source: FHWA)

Source: https://safety.fhwa.dot.gov/zerodeaths/zero_deaths_vision.cfm

Importantly, the new large-scale data includes detailed data about vehicle kinematics including speed, acceleration/deceleration, and jerk which could be used to assess driving instability using new measures. With the availability of new databases (particularly SHRP2-NDS data), research efforts, focusing on different perspectives, were made to conduct safety analysis at different levels; however, certain important gaps remain. For instance, previous studies have developed systematic taxonomies of human factors to understand the role of driving errors and violations, at a deeper level, in crashes using police crash data. None of the studies has developed the taxonomy of driving errors and violations using real-world SHRP2-NDS data which we consider imperative as the nature and sources (e.g., cell phone use, climate control) of one or more driving errors and violations might have significantly changed compared to past decades. After developing systematic taxonomy of driving errors and violations (TDEV) and quantifying the contribution of driving errors and violations in SCEs using SHRP2-NDS data, some key research questions need to be answered. For instance, one of the key questions is how diverse roadway environments relate to SCEs both directly and through one or more driving errors and violations. This question is based on our hypothesis that certain roadway environments (school zones, urban areas, and business/industrial locations) would encourage certain types of driving errors (recognition errors and decision errors) which in turn would have a significant influence on SCEs. Our next key hypothesis is that drivers after making an error or violation and realizing it may take corrective actions which may, in turn, induce some crash risk. We also presume that roadway environments may have a significant influence on SCEs through driving instability which is defined as deviation from the norm (normal values) of specific driving parameters including speed, acceleration, deceleration, and jerk. Besides the important hypotheses, this dissertation focuses on understanding how distraction duration (seconds) relates to SCEs both directly and through driving instability while accounting for non-linearity in the effects of distraction duration on both SCEs and driving instability. The motivation behind this was based on our hypothesis where we consider that effects of distraction duration (seconds) on driving instability and SCEs would significantly vary at different values of distraction. After exploring the role of roadway environments, driving errors, violations, and driving instability while harnessing SHRP2-NDS data, this dissertation seeks an answer to the question “whether and how errors, violations, and impairment by both pedestrians and drivers relate to pedestrian injury severity (given pedestrian-vehicle crash)”. Besides, investigating the inferential relation of human factors including driver errors, violations, duration of distraction, and instability in driving on SCEs, this dissertation applies sophisticated machine learning methods to predict SCEs based on all the aforementioned human factors using real-world SHRP2-NDS data. To test the hypotheses and answer the relevant questions, this dissertation makes an effort to analyze both SHRP2-NDS data and enriched data obtained from North Carolina Pedestrian and Bicycle Crash Data Tool (NC PBCAT). To conclude, my Ph.D. dissertation will explore both SHRP2-NDS data and conventional NC PBCAT data to explore the

role of human behaviors, decisions, and performance in SCEs from multiples aspects. The motivation behind this dissertation is to generate new knowledge while applying new concepts and methods which could help in selecting more effective countermeasures to road user safety through multiple protective layers including safe road users, safe vehicles, safe roadway environments, and safe speed. My dissertation contributes to the existing state-of-the-art through five chapters (some of which are published/presented in peer-reviewed journals/conferences) as discussed below.

Chapter 2 focuses on understanding the role of naturalistic driving errors and violations in crashes and identifying strategies which could reduce them where possible. The new idea is to measure and quantify pre-crash driving behaviors, bring them more substantively into the behavioral framework, and then understand how these errors and violations vary across diverse roadway environmental contexts. The results reveal that human factors still prevail contributing to around 93% of the observed crashes. The most common human errors were recognition and decision errors, which occurred in 39% and 34% of crashes, respectively. These two error types occurred more frequently (each contributing to ~39% of crashes) in business or industrial land use environments (but not in dense urban localities).

Chapter 3 conducts rigorous path analysis supported by TDEV to explore multiple pathways for diverse roadway environments to crashes both directly and through one or more driving errors and violations. Multidimensional connections between diverse roadway environments, driving errors and violations, and the ultimate safety outcomes are explored through two-stage path analysis including Tobit and ordered probit model. The results reveal that all key types of driving errors and violations significantly increase crash risk. Results also reveal that roadway environments including school zones, urban areas, business/industrial locations induce crash risk not only directly but also through each recognition error and decision error. The findings are expected to help in identifying more effective countermeasures which could bring more safety benefits to the system.

Chapter 4 aims to answer the question of whether a driver, after making errors and violations, becomes more instable (volatile), and has a thus higher chance of involving in SCEs. Besides, this chapter also investigates how the duration of distraction relates to SCEs while accounting for the non-linearity in its effects on SCEs both directly and through instability in driving. Using a safe system approach, this chapter applies path analysis to explore the role of driving errors, violations, and duration of distraction in SCEs through jointly driving instability to account for the potential correlation between the unobserved factors in the two equations. The findings reveal that all driving errors and violations are associated with increased crash risk and near-crash risk both directly and indirectly through driving instability. Furthermore, the results indicate that both instability in driving and the chance of SCEs increase non-linearly with an increase

in the duration of distraction. Also, findings suggest sufficient statistical evidence of the correlation between unobserved factors in the two equations which if not taken into account through joint estimation could lead to inefficient estimates.

Chapter 5 applies machine learning methods to predict SCEs based on human factors including driving errors, violations, duration of distraction, and instability in driving (captured via a coefficient of variation of speed, deceleration, and negative jerk) in diverse roadway environments. To achieve the study objectives, the study applies sophisticated machine learning methods including naïve Bayes classifier, K-nearest neighbor, and gradient boosting tree. For the sake of comparing out-of-sample prediction accuracy and to understand the inferential relation of human factors with SCEs, the ordered Probit regression is also used to predict the event outcome (baseline, near-crash, and crash). After applying the conventional statistical model and ML methods to the training sample, both the overall out-of-sample prediction accuracy and accuracy for a specific class are computed and compared. The findings show that all the three ML methods showed higher out-of-sample prediction accuracy compared to the ordered Probit model; however, the gradient boosting tree showed the most accurate performance among with 91.23% overall accurate predictions in the test sample. Referring to the individual class prediction, the gradient boosting tree again showed the highest out-of-sample accuracy in predicting both crashes and near-crashes.

Chapter 6 quantifies the contribution of human factors (errors, violations, and impairment by both pedestrian and driver) on pedestrian injury severity (given a crash) using NC PBCAT data. First, all risky pre-crash behaviors of pedestrians and drivers in pedestrian-vehicle crashes are identified and brought under separate classification systems. Then finite mixture models are applied to account for unobserved variation in pedestrian injury severity across unobserved (latent classes) based on some unobserved factors (aggressiveness of driver/pedestrian, pedestrian conspicuity, and impact speed). The initial findings reveal that 60.78% of pedestrian crashes occurred due to human factors where violation and recognition errors by pedestrians were the predominant errors which contributed to 28% and 22% of overall pedestrian crashes respectively. Results reveal that recognition and performance errors by a driver and violations by pedestrians are associated with a higher chance of a pedestrian being killed and seriously injured. Results also suggest significant evidence of heterogeneity in pedestrian injury severity across two different latent (unobserved) classes which if not considered could lead to inaccurate parameter estimates and predictions.

In summary, the key contribution of this dissertation is to explore the effects of human factors on SCEs and pedestrian injury severity given a crash. Both large-scale SHRP2-NDS data and NC PBCAT crash data are harnessed. Findings obtained through this dissertation are relevant to intelligent transportation systems. The findings provide a firm foundation to select more effective strategies including multiple vehicle

technologies, roadway “built” environmental changes, and policy interventions that could potentially reduce one or more driving errors, violations, and driving instability, thus consequently road user safety. For the conceptual framework and outline of the proposal, please refer to Figure 1.2 and Figure 1.3 respectively. Under this dissertation, it is expected to publish at least five papers in high impact factor (IF) transportation journals listed below (along with my contribution):

- Khattak, A., Ahmad, N., Wali, B., & Dumbaugh, E. (2021). A taxonomy of driving errors and violations: Evidence from the naturalistic driving study.
 - Journal article: Published in *Accident Analysis & Prevention*
 - Peer-reviewed conference paper: Presented in 100th Transportation Research Board Annual Meeting 2021, Washington DC.
 - My Overall Contribution: 65% (Conceptual design, data preparation, methodology, analysis and interpretation, software analysis, original draft preparation).
- Ahmad, N., Wali, B., Khattak, A., & Dumbaugh, E. Built environment, driving errors and violations, and crashes in naturalistic driving environment.
 - Journal article: Published in *Accident Analysis & Prevention*
 - Peer-reviewed conference paper: Presented in 98th Transportation Research Board Annual Meeting 2019, Washington DC.
 - My Overall Contribution: 75% (Conceptual design, data preparation, methodology, analysis and interpretation, software analysis, original draft preparation).
- Ahmad, N., Arvin, R., & Khattak, A. (2021). Exploring Pathways from Driving Errors and Violations to Crashes: The Role of Speed Volatility.
 - Under Review in Journal of *Accident Analysis & Prevention*.
 - Peer-reviewed conference paper: Presented at 100th Transportation Research Board Annual Meeting 2021, Washington D.C.
 - My Overall Contribution: 80% (Conceptual design, data preparation, methodology, analysis and interpretation, software analysis, original draft preparation).
- Ahmad N., Arvin, R., & Khattak, A. Exploring non-linear effects of distraction duration on driving instability and safety-critical events.
 - In Review in Journal of *Accident Analysis & Prevention*.
 - Peer-reviewed conference paper: Accepted for Presentation at 101st Transportation Research Board Annual Meeting 2022, Washington D.C.
 - My Overall Contribution: 85% (Conceptual design, data preparation, methodology, analysis and interpretation, software analysis, original draft preparation).



Figure 1. 2 Overall Conceptual Framework of the Proposal

Note: The sources to the picture in Figure 1.2 are: <https://www.safekids.org/infographic/school-zone-infographic>; <https://medium.com/@rohitgurjar009/various-road-patterns-2-3-e55d48c947fa>; <https://venturebeat.com/2020/03/22/the-covid-19-crisis-could-show-us-how-to-cure-our-congested-cities/>; <https://www.shutterstock.com/image-photo/no-distracted-driving-sign-red-stop-206848018>; <https://www.bestbuy.com/site/shop/bluetooth-headset-for-phone>; <https://www.cnet.com/tech/mobile/apple-listens-to-some-siri-recordings-to-make-it-better/>; <https://wallpapersafari.com/w/XEGOf>; <https://www.bosch-mobility-solutions.com/en/solutions/assistance-systems/adaptive-cruise-control/>; <https://pmrpressrelease.com/forward-collision-warning-systems-market-industry-perspective/>; <https://www.shutterstock.com/search/distracted+driving+sign>

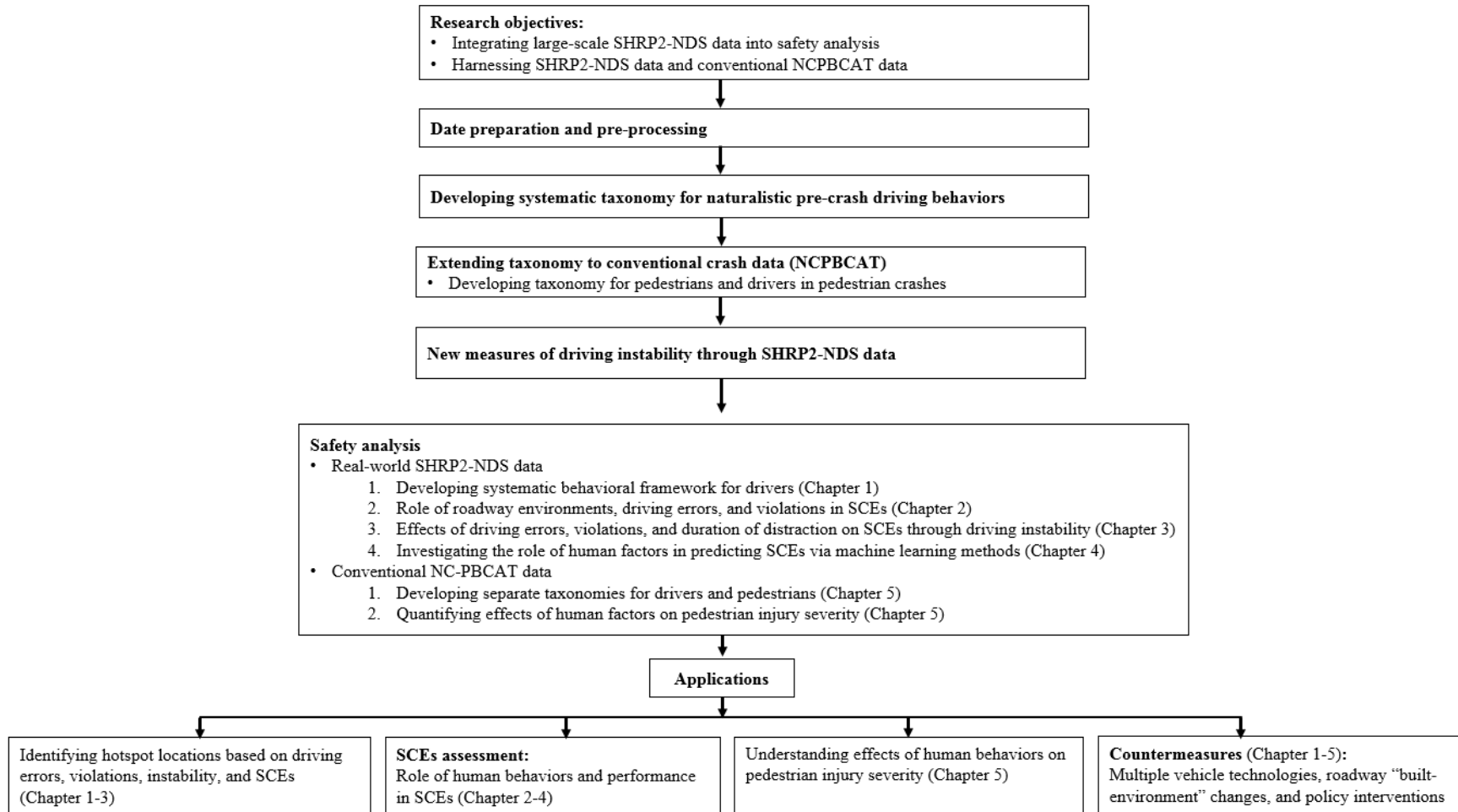


Figure 1. 3 Proposal Outline

Note: Chapter 1 is published in the Accident Analysis and Prevention journal and presented in TBR2021; Chapter 2 is published in Accident Analysis and Prevention journal and presented in TRB2019; Chapter 3 has resulted in 2 papers both of which are in review with Accident Analysis and Prevention journal. Note that 1st paper from Chapter 3 was presented in TRB2021 while the 2nd one is accepted for presentation in TRB2022; Chapter 4 is in preparation and will be submitted to an impact factor journal for publication review; while a portion of Chapter 5 is submitted to the TRB2022 for both presentation and publication review.

- Ahmad N. & Khattak, A. The role of human factors in pedestrian injuries given a crash: application of finite mixture model.
 - A portion of this chapter is under review in TRB2022 and Journal of Transportation Research Record for presentation and publication respectively.
 - My Overall Contribution: 80% (Conceptual design, data preparation, methodology, analysis and interpretation, software analysis, original draft preparation).

Problem Justification

According to recent estimates, more than 38,000 people die every year as a result of road crashes across the U.S with a fatality rate of 12.4 per 100,000 population. Moreover, 4.4 million people get seriously injured in road crashes on U.S. roadways every year. The social and economic impact of road crashes leads to a cost of \$871 billion per year to the U.S. population which includes more than \$380 million of direct medical costs. Human factors, including driving errors and violations, are highly relevant to the safe system approach as they contribute to more than 90% of road crashes (Sabey and Staughton 1975, Treat *et al.* 1979, Singh 2015). Humans use the cognition process for conducting activities where deficiency could lead to potential errors. Human cognition is a complex mechanism that involves a series of sequential phases including sensing, perceptual encoding (recognition), central processing (decision), responding (performance) (Figure 1.4). A similar cognition mechanism relates to the human driver where a driver may sense, perceive/recognize, process (decide), and respond to potential safety risks during driving (Figure 1.4). During the sensing phase, a driver may sense a potential safety risk that involved the transformation of physical energy into neural energy. Next comes the perceptual encoding (recognition) phase where a driver may extract meaningful information from sensory details received during the input phase (Figure 1.4). In the central processing (decision), a driver, through the central nervous system, may decide a course of action to avoid the potential safety risk (Figure 1.4). After sensing, perceiving, and processing the information, a human driver may execute his/her intended course of action either physically or mentally to overcome the potential safety risk (Figure 1.4). To develop an understanding of driving behaviors, at a deeper level, studies proposed a systematic behavioral framework of driving error and violations based on the human cognitive process (Treat *et al.* 1979, Singh 2015). Based on these taxonomies driving behaviors, which contributed to crashes, were mainly classified into recognition errors, decision errors, performance errors, and violations (Treat *et al.* 1979, Singh 2015). Studies reveal that recognition errors and decision errors are the two predominant driving errors that contributed to a majority of road crashes (Treat *et al.* 1979, Singh 2015). Dr. Michael Clamann (Human Factor Engineer at UNC, North Carolina) conducted a thorough literature review which shows similar findings of the predominance of recognition errors and decision errors in crashes. The aforementioned studies proposed a systematic behavioral framework that

provides a more holistic picture of the origin and nature of driving errors and violations based on the human cognition process. However, most of the previous taxonomies of driving behaviors were based on police crash reports which could be subjective and bias (Yamamoto *et al.* 2008, Washington *et al.* 2010, Ye and Lord 2011). Importantly, the human cognitive process is a sequential process that could be more relevant to real-world data rather than traditional police crash data. On other hand, emerging data sources like SHRP2-NDS include detailed information on real-world pre-crash driving behavior, performance, vehicle kinematics, and other important factors (e.g., distraction duration) which enables researchers to think out of the box and generate new knowledge through applying new methods and concepts. The SHRP2-NDS data provides us opportunities to analyze driving behaviors and performance at a microscopic level due to the availability of real-world (second-to-second) information on pre-crash driving behavior and performance. Besides crashes and near-crashes, the SHRP2-NDS data also includes data on baselines (used to quantify crash risk) which could not be available in data obtained through police crash reports. Given the advantages associated with SHRP2-NDS data, this dissertation mainly harnesses SHRP2-NDS data to understand the role of diverse roadway environments, driving errors and violations, and driving instability on SCEs. In addition to investigating and quantifying the effects of human factors in SCEs, this study applies more sophisticated machine learning methods including naïve Bayes classification, K-Nearest Neighbor classification, and gradient boosting tree to predict SCEs based on the human factors (driving errors, violations, duration of distraction, and instability in driving) in diverse roadway environments using real-world SHRP2-NDS data.

After thoroughly investigating SHRP2-NDS data to understand the role of roadway environments, driving errors, violations, and driving instability in SCEs, we are interested to understand how human factors (referring to both pedestrians and drivers) influence pedestrian injury severity (given a crash) which has been lightly addressed. To pursue the later objective, we selected data from NC PBCAT to understand the effects of human factors on pedestrian injury severity given a crash. The reasons for selecting conventional crash data (NC PBCAT) to quantify the effects of human factors on pedestrian injury severity are as 1). While SHRP2-NDS data have plentiful advantages (as discussed above), it includes very few pedestrian crashes, 2). The SHRP2-NDS crashes were mostly minor and low-risk/tire strike crashes (78%), whereas only 22.14% of the overall crashes were police-reportable making SHRP2-NDS data not fully comparable with police crash reports based when it comes to injury severity, 3). Also, SHRP2-NDS data only provide detailed information about pre-crash behaviors and performance for subject drivers, whereas such information for other drivers or pedestrians is not available. Given the aforementioned reasons, we selected NC PBCAT crash data (2007-2019) to analyze pedestrian injury severity given a crash. Note that NCPBCAT data is based on police crash reports; however, more detailed information about risky behaviors

of pedestrians, drivers, and roadway environments, etc. was additionally extracted from the narratives in the crash report forms by NC DOT.

Literature Review, Gaps, and Contributions

Developing Taxonomy of Naturalistic Driving Errors and Violations

Literature Review

Pre-crash factors leading to a crash event belong to one or more of the three categories: driver-related, vehicle-related, and environment-related factors (Sabey and Staughton 1975, Treat *et al.* 1979, Frawley and Eisele 2004, Huang *et al.* 2010, Klauer *et al.* 2014, Singh 2015). Police-accident report-based studies have found that driver-related factors are the prevailing contributing factor in crash events, associated with more than 90% of all crashes (Sabey and Staughton 1975, Treat *et al.* 1979, Singh 2015). While police-reported crash data sources provide valuable information, the interpretation of the data they contain is necessarily subjective; since most of the information in police crash reports is not based on direct observations of the events preceding a crash, analysts are forced to infer these behaviors based on data collected after the crash event has occurred (Hankey *et al.* 2016, Wali *et al.* 2018e). Other inaccuracies, such as non-reporting of low severity (property damage) crashes, exist in traditional crash data (Yamamoto *et al.* 2008, Washington *et al.* 2010, Ye and Lord 2011).

Real-world driving performance and behaviors in time-to-collision can be examined using data from numerous geo-coded (location-based) driving parameters such as speed, acceleration, time to collision, secondary tasks/durations, distractions, and eye glance behavior of the drivers being continuously captured via advanced equipment (Dingus *et al.* 2006, Dingus *et al.* 2016). All these advanced features make NDS databases more unique and reliable compared to conventional police crash reports. Earlier studies have classified errors and developed taxonomies for driver errors and violations through different themes (Treat *et al.* 1979, Reason *et al.* 1990, Petridou and Moustaki 2000, Wierwille *et al.* 2002, Stanton and Salmon 2009). Driver errors have been classified as slips, lapses, or mistakes (Reason *et al.* 1990). In another study, the classification focused on differentiating behavioral factors affecting driving capabilities from those encouraging risky driving (Petridou and Moustaki 2000). From previous taxonomies, key driving errors include recognition errors, decision errors, and performance errors (Treat *et al.* 1979, Wierwille *et al.* 2002). Recognition and decision errors were found to be the leading and critical errors resulting in crashes (Treat *et al.* 1979, Wierwille *et al.* 2002, Iden and Shappell 2006). Due to variations in driver characteristics and behavioral patterns, different roadway environments can induce different errors and violations (Dumbaugh and Li 2010).

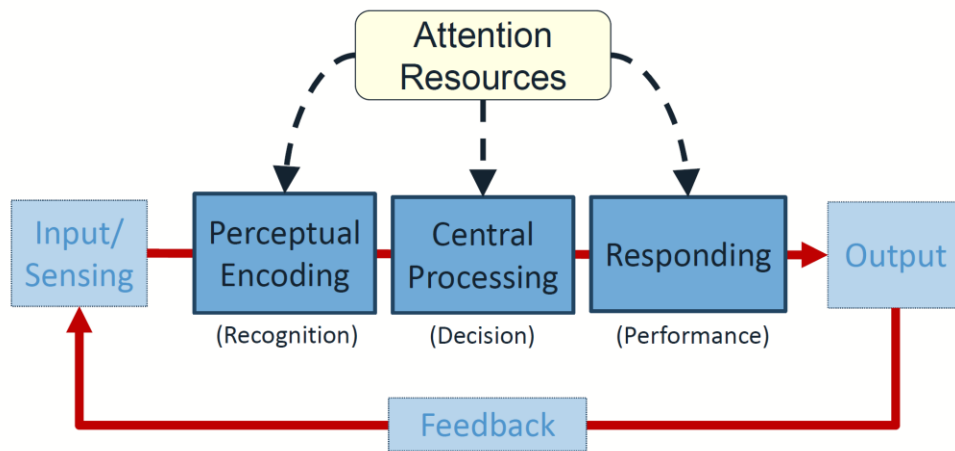


Figure 1. 4 Model for Human Information Processing

Source: Dr. Michael Clamann's Presentation (UNC, North Carolina)

Research Gaps

While previous taxonomies of driving errors and violations have led to valuable insights, none (to the best of our knowledge) have utilized real-world naturalistic driving datasets which provide detailed pre-crash driver behavior information for crashes, near-crashes, and baseline driving. Therefore, there is a need to develop a detailed and systematic taxonomy of real-world safety-critical driving errors and violations. Also, understanding how different naturalistic driving errors and violations vary across diverse roadway environments is something that has been lightly addressed.

Objectives and Contribution

Given the gaps in the literature, this paper aims to achieve some of the key objectives given below.

- To develop a systematic taxonomy of driving errors and violations (TDEV) using real-world SHRP2-NDS data.
- To understand how different driving errors and violations vary across diverse roadway environmental contexts.

While using new SHRP2-NDS data, this study reinforces and validates the need to address human response and behavior as the primary driver of safety, the new idea is to measure and quantify pre-crash behaviors and bring them more substantively into the behavioral framework. A different set of errors and factors are identified as contributing to crashes than before. For instance, SHRP2-NDS provides real-world data on pre-crash driving behaviors (e.g., exceeded safe speed but not speed limit or driving slowly with other traffic) and secondary tasks (e.g., cell phone use, talking/listening to other occupants or music, and using hands-free communication) regarded by data reductionists as crash causative factors which are not available in police crash reports. Furthermore, data availability on baselines (exposure without an event) and near-crashes (i.e., safety-critical events) are new aspects of our research. Using the unique set of information on pre-crash driving behavior, secondary tasks, baselines, and near-crashes, we first bring human crash-contributing factors into a systematic framework and then compute their associated crash risk. These efforts can in turn help accident investigations and provide insights into the development of automated vehicles.

Exploring Effects of Roadway Environment on Driving Errors, Violations, and Crashes

Literature Review

In general, human errors can be classified using a generic-error-modeling system (Reason 1990). According to the system, human factors can be classified into skill-based slips (and lapses), knowledge-based mistakes, and rule-based mistakes (Reason 1990). Slips (e.g., misperception, action intrusion, and disordering of

action) and lapses (e.g., forgetting intended action and loss of action sequence) include those types of human errors in which actions of an operator deviate from intentions as a result of storage and/or execution failures (Norman and Shallice 1980, Reason and Mycielska 1982, Reason 1990). Mistakes (e.g., poor or inappropriate decision making and overconfidence) indicate human errors where actions run according to intention; however, intended outcomes are not achieved due to inadequate plans (Reason 1990). Besides, violations were placed in a separate category and were further classified into intentional and unintentional violations (Reason 1990). Human errors become more critical in driving where the human driver is required to interact with complex and diverse roadway and traffic environments. To understand the origin and nature of various types of driving errors, studies proposed and/or discussed various taxonomies of driving errors and violations to conceptualize human factors, which contribute to crashes, into a systematic framework using diverse themes (Treat *et al.* 1979, Reason *et al.* 1990, Petridou and Moustaki 2000, Wierwille *et al.* 2002, Stanton and Salmon 2009, Salmon *et al.* 2010). While previous studies generated useful knowledge about driving errors and violations, most of them were based on either police crash reports or self-reported behavioral surveys which could be subjective (Ahmad *et al.* 2019, Wali and Khattak 2020). Only one of the studies (to the best of our knowledge) used data obtained from “external video cameras” installed at various locations to monitor accidents (Wierwille *et al.* 2002). While this could be a relatively better approach to determine the true cause of road crashes, still it might have not captured the exact driver behavior (especially in-vehicle behaviors) and performance. Besides, it did not include information on baselines that could be used as an exposure to explore crash risk (Wierwille *et al.* 2002). Real-world driving behavior and performance before safety-critical events can be explored using SHRP2-NDS data (Dingus *et al.* 2016, Khattak *et al.* 2021a). NDS data include objectively collected information that is unique and more reliable (Dingus *et al.* 2016, Khattak *et al.* 2021a). While realizing the need to systematically classify human factors using real-world data on pre-crash driver behavior and performance, one of the recent studies developed a systematic taxonomy of driving errors and violations (TDEV) using SHRP2-NDS data mainly based on the perception-reaction process (Khattak *et al.* 2021a). Using real-world SHRP2-NDS data, pre-crash human behaviors, which have contributed to crashes, were classified into recognition errors, decision errors, performance errors, physical condition-related errors, and exposure errors (Khattak *et al.* 2021a). Violations were kept in a separate category and were further classified into intentional and unintentional violations (Khattak *et al.* 2021a). The TDEV is a timely effort to bring all real-world pre-crash driving behaviors into a systematic framework which is more relevant as we consider that the nature and sources (e.g., cell phone use) of one or more driving errors might have significantly changed during the recent years (Khattak *et al.* 2021a). Based on SHRP2-NDS data, human factors were found to still contribute to 93% of crashes whereas recognition error and decision errors contributed to almost 39% and 34% of crashes respectively (Khattak *et al.* 2021a).

Numerous studies have explored key factors contributing to crash occurrence and/or injury outcomes given a crash (Frawley and Eisele 2004, Huang *et al.* 2010, Klauer *et al.* 2014). Several behavioral factors correlate with crashes (Khattak *et al.* 2002, Frawley and Eisele 2004, Dumbaugh and Li 2010, Huang *et al.* 2010, Klauer *et al.* 2014, Dingus *et al.* 2016). Among them, driver distraction was found to be one of the most critical driver behavior (Klauer *et al.* 2014). Studies proposed methods for detecting risky driving behaviors using video surveillance (Chen *et al.* 2017). Driver must continuously interact with various roadway entities (e.g., roadway environment, traffic control devices, and other traffic), which in response influence driver performance as compared to the hypothetical scenarios where a driver does not interact with any roadway entities. For instance, studies suggest that driving behavior is significantly influenced by roadway infrastructure (Schorr *et al.* 2017). Driver assistance systems (such as in-vehicle navigation systems) were deployed to assist drivers in overcoming driving errors. However in real scenarios, even with a driving assistance system, drivers may make certain errors and mistakes due to several reasons like the poor design of human-machine interface or simple driver carelessness (Ng and Sathasivan 2014). In response, erroneous driving behavior increases crash risk. Studies also reveal that local demographics and socio-economic conditions can contribute to higher crash risks (Huang *et al.* 2010). The chances of crashes are higher with a higher density of local population, higher traffic intensity, and urbanization (Huang *et al.* 2010). Similarly, freeways were associated with a lower crash risk than arterials (Huang *et al.* 2010). From a roadway standpoint, the existence of a construction zone on a road facility increases crash risk (Khattak *et al.* 2002).

Research Gaps

While the literature acknowledges the contributions of the roadway and environmental factors on crash occurrence (Dumbaugh and Li 2010), this issue is lightly addressed. The role of roadway environments in encouraging or discouraging certain naturalistic risky behaviors is not well-understood. The multidimensional connections between roadway environment, driving errors and violations, and the ultimate safety outcomes are not explicitly quantified. Compared to subjective data sources (Anastasopoulos *et al.* 2012, Venkataraman *et al.* 2013, Ahmad *et al.* 2019), the use of objective real-world safety data (such as SHRP2-NDS data used in this study) could lead to a more enriched understanding of the key behavioral elements influencing safety - ultimately leading to more effective crash prevention strategies.

Objectives and Contribution

Given the aforementioned gaps, this study adds to the existing state-of-the-art regarding crash prevention by focusing on the key gaps listed below.

- To examine, in a greater depth, the role of diverse roadway environments in encouraging or discouraging different driving errors and violations using real-world SHRP2-NDS data.
- To explore complex pathways which may lead from diverse roadway environments to crash both directly and through one or more driving errors and violations.

To achieve the research objectives, this study applies a rigorous path analytic framework to unveil the key pathways that lead from diverse roadway environments to SCEs both directly and through driving errors and violations using the new SHRP2-NDS data. From a conceptual standpoint, the study tests the hypothesis that diverse roadway environments can significantly encourage or discourage certain types of naturalistic driving errors and violations which in turn could significantly induce crash and/or near-crash risk – something which was not explored earlier. From the methodological standpoint, this study applies a rigorous path analytic approach comprising of two constituent models including multinomial logit and generalized ordered Probit (goprobit) models. To model the ordinal data in the path analysis, previous studies have widely used conventional ordered Probit or ordered logit (proportional odds) models (Liu *et al.* 2015, Zhang *et al.* 2018, Kamrani *et al.* 2019). Regardless of the path analysis, studies generally use traditional ordered Probit or logit models (Williams 2016, Wali *et al.* 2017, Ahmad *et al.* 2019) which could be inappropriate if one or more of the explanatory variables violates the proportional odds (or parallel line) assumption. If the proportional odds assumption is violated by at least one of the explanatory variables then the application of proportional odds models becomes inappropriate (Williams 2005, 2016). To account for the parallel line assumption violation, the current study applies the partial proportional ordered Probit (goprobit) model for the event outcome (baseline, near-crash, and crash) in the path analytic framework. Studies suggest that goprobit models are more parsimonious and accurate when the parallel line assumption is violated (Williams 2005, 2016). To sum it up, besides the conceptual rigor, from a methodological standpoint, this study contributes by using a partial proportional ordered Probit model for the event outcome in the path analysis which enables the study to capture the heterogeneity across different categories of the event outcome while providing more parsimonious and easily interpretable results.

Quantifying the Effects of Driving Errors, Violations, and Duration of Distraction on SCES Through Instability in Driving

Literature Review

To have a deeper understanding of driving behaviors and their contribution to crashes, studies proposed systematic classification systems of human factors based on different themes which provide a more inclusive and systematic understanding of their contribution to crashes (Treat *et al.* 1979, Reason *et al.* 1990, Petridou and Moustaki 2000, Wierwille *et al.* 2002, Stanton and Salmon 2009). For instance, human

factors in crashes were classified into recognition errors, decision errors, performance errors, and violations (Treat *et al.* 1979). While previous taxonomies provide useful insights about how human factors contribute to crashes, they were based on police crash reports which could be subjective and bias (Yamamoto *et al.* 2008b, Kamrani *et al.* 2019, Wali and Khattak 2020, Khattak *et al.* 2021a). Moreover, police crash reports use indirect observations as it is unrealistic for a police officer to be present at the crash scene before crash occurrence (Hankey *et al.* 2016). Furthermore, police crash reports can be prone to underreporting issues as property damage only and minor crashes might not be reported (Yamamoto *et al.* 2008, Washington *et al.* 2010, Ye and Lord 2011).

To bridge the aforementioned research gaps, one of the recent studies used the comprehensive and objective SHRP2-NDS data to classify driving behaviors that contributed to crashes (Khattak *et al.* 2021a). Naturalistic driving behaviors were systematically categorized into recognition errors, decision errors, performance errors, physical condition-related errors, experience errors, and violations which include both intentional and unintentional violations (Khattak *et al.* 2021a). Note that all naturalistic driving behaviors (e.g., distraction due to cell phone browsing, cell phone holding, cell phone texting, drinking with lid and straw, adjusting/monitoring radio, and looking at object external to the vehicle) due to which a subject driver failed to appropriately recognize the true situation (which consequently led to a crash) were classified as recognition errors (Khattak *et al.* 2021a). Given that the subject driver perceived and recognized the situation appropriately, but took some inappropriate decisions (e.g., exceeded safe speed but not speed limit and following too closely), which resulted in a crash, were regarded as decision errors (Khattak *et al.* 2021a). Next comes performance errors and violations which could occur in the reaction phase when a subject driver had already recognized the situation and had decided his intended course of actions appropriately to overcome the safety hazards (Khattak *et al.* 2021a). In the reaction phase, subject drivers might either make performance errors (e.g., improper turn, wide left turn, steering too far) or violate the traffic regulations including both intentional violations (e.g., signal violation - intentionally disregarded signal, failed to signal, and signal violation – tried to beat a signal change) and unintentional violations (e.g., signal violation, apparently did not see the signal) (Khattak *et al.* 2021a). For further details about TDEV and constituent driving errors and violations, please see (Khattak *et al.* 2021a). The TDEV can be considered as important contributors to the previous taxonomies (based on police crashes) as driving behaviors especially sources of distraction (e.g., cell phone use, cell phone, browsing) have significantly changed compared to the past (Khattak *et al.* 2021a). This study takes advantage of the aforementioned TDEV to bring naturalistic driving behaviors into a systematic framework using SHRP2-NDS data which are then used to compute pathways which leads from these driving errors and violations to SCEs both directly and through instability in driving.

Besides pre-crash driving behaviors, studies reveal that driving volatility (instability) significantly influences crash risk (Wali *et al.* 2019, Wali and Khattak 2020) and crash severity (Arvin *et al.* 2019). Driving volatility can be defined as deviations from the normal values of driving parameters like speed, accelerations or decelerations, and jerk (Arvin *et al.* 2019, Wali *et al.* 2019, Wali and Khattak 2020). Some of the pioneer studies which introduced driving volatility in surface transportation are listed in parenthesis (Liu 2015, Wang *et al.* 2015, Liu *et al.* 2017b). The SHRP2-NDS data includes information on spatiotemporal variations in vehicle parameters like speed, accelerations, decelerations, and jerks which can be used to derive various volatility indices (Arvin *et al.* 2019, Wali *et al.* 2019, Wali and Khattak 2020). Studies reveal interesting findings of how various volatility indices relate to crash risk and crash severity (Arvin *et al.* 2019, Wali *et al.* 2019, Wali and Khattak 2020). For instance, it is found that crash risk in school zones increases with an increase in vehicular jerk in both longitudinal and lateral directions (Wali and Khattak 2020). It is also suggested driving volatility increases the likelihood of violation and crash (Yeo *et al.* 2020). Referring to the association between volatility and crash severity, studies suggest that driving volatility positively correlates with crash severity (Arvin *et al.* 2019). In summary, the above discussion reveals that 1) studies have investigated how instability (volatility) in driving relates to SCEs (Arvin *et al.* 2019, Wali *et al.* 2019, Wali and Khattak 2020) and 2) studies have also investigated the effects of naturalistic driving errors and violations on SCEs (Ahmad 2021). However, this study considers that drivers after making errors and violations and realizing them may take corrective actions which could increase their instability in driving thus further inducing crash or near-crash risk – something which remains unexplored to the best of authors' knowledge.

Harnessing the SHRP2-NDS data source, studies suggest that subject driver, if involved in distracting activities, has a higher risk of involving in a crash (Khattak *et al.* 2019, Khattak *et al.* 2021a). The aforementioned studies add useful knowledge to the literature highlighting how real-world pre-crash distractions, by the subject drivers, relate to crash risk; however, they do not provide a deeper understanding of how the duration of distraction relates to crash or near-crash risks. Using data from instrumented vehicles, studies have looked into how the duration of distraction may affect crash risk (Simons-Morton *et al.* 2014, Gershon *et al.* 2019, Arvin and Khattak 2020). One of the key studies explored how distraction duration due to different sources including cell phone use, eating, and monitoring/control of radio or climate change, etc. induce the risk of SCEs (Arvin and Khattak 2020). It is found that distraction duration due to various sources increases the crash risk (Arvin and Khattak 2020). These studies provide a useful and deeper understanding of how the duration of distraction relates to the chance of crash (Gershon *et al.* 2019) or SCEs (Simons-Morton *et al.* 2014, Arvin and Khattak 2020); however certain research gaps exist. First, we consider that distraction duration may not only induce crash and near-crash risk directly but also through

instability in driving. The key rationale behind this is that drivers distracted for a longer duration may become more volatile; therefore, their chance of involving in a crash or near-crash may further increase through its influence on instability in driving. Furthermore, there could be non-linearity in the effects of duration of distraction on both instability in driving and probability of SCEs which needs to be accounted for – something which has been lightly addressed.

Research Gaps

In summary, from the aforementioned discussion, it can be concluded that 1) None of the studies has investigated the underlying pathways from naturalistic driving errors and violations to SCEs through instability in driving, and 2) no research effort was carried out to investigate how the duration of distraction relate to SCEs directly and through instability in driving.

Objectives and Contribution

Consider the literature review and research gaps, this study applies a rigorous path analytic approach to achieve some of the objectives listed below.

- To explore the relationship of driving errors and violations with SCEs through instability in driving in diverse roadways and environments.
- To investigate how the duration of distraction relates to SCEs both directly and through instability in driving while accounting for the non-linearity in the effects of duration of distraction in the path analysis.
- To access and account for potential correlation between the unobserved factors associated with both event outcome (baseline, near-crash, and crash) and COV of speed through joint estimation.

Besides quantifying complex pathways that lead from driving errors, violations, and duration of distraction to SCEs both directly and through instability in driving, this study applies a joint estimation procedure to account for the potential correlation between the unobserved factors which may significantly influence both the event outcome (baseline, near-crash, and crash) and instability in driving. Note that if a significant correlation between unobserved factors exists and is not taken into account, the results would be inefficient (Ahmad 2021). In simple words, while using standard path analysis (where constituent models are developed independently and not in a joint fashion) maybe still consistent, more efficient parameter estimates could be obtained through joint estimation when a potential correlation between the unobserved factors exists (Ahmad 2021). We note that most of the previous studies have used standard path analysis while ignoring the potential correlation between the common unobserved factors (Liu and Khattak 2017, 2018, Arvin *et al.* 2019). Furthermore, previous studies have used average marginal effects (a constant

change in the response variable or probability of a specific category of a response variable) in the path analysis to compute direct, indirect, and overall effects of a specific explanatory variable. While these studies may provide useful insights, using average marginal effects for continuous explanatory variables (e.g., duration of distraction in this study) becomes less appropriate. To account for the non-linearity in the association of duration of distraction with SCEs and instability in driving, this study uses a squared term for the duration of distraction in the two constituent models, and then uses the predictive margins (which shows the predicted value and predicted probabilities of instability in driving and SCEs respectively) to access its effects on SCEs. In summary, from a methodological standpoint, this study not only captures the non-linearity in the relationship of duration of distraction with SCEs and instability in driving in a path analytic framework but also accounts for potential correlation between the unobserved factors associated with the instability in driving (in Tobit model) and SCEs (in ordered Probit model) something which has been lightly addressed in the context of safety literature. Note that the selection of Tobit regression for modeling the COV of speed is based on the distribution which exhibits left spikes – such a scenario is termed as a corner solution problem. While the corner solution problem is not a data observability issue like truncation and censoring, studies suggest that the Tobit regression can be applied in all three scenarios including truncation, censoring, corner solution problems.

Predicting Safety-Critical Events Using Driver Behaviors and Performance: Application of Artificial Intelligence

Literature Review

Previous studies reveal that human factors contribute to more than 90% of road crashes (Sabey and Staughton 1975, Treat *et al.* 1979, Singh 2015). Note that most of the relevant studies in the past have analyzed police crash data which could be subjective when it comes to pre-crash driving behavior (Dingus *et al.* 2006a, Arvin *et al.* 2019, Ahmad *et al.* 2020, Ahmad 2021, Khattak *et al.* 2021a). On other hand, the SHRP2-NDS data includes real-world microscopic details about pre-crash driving behaviors, performance, vehicle kinematics (speed, acceleration, and deceleration), and other roadway and environment variables captured through fitted cameras and recording equipment in the subject (instrumented) vehicles (Dingus *et al.* 2006a, Arvin *et al.* 2019, Khattak *et al.* 2021a). Utilizing the SHRP2-NDS data, one of the recent studies shows that human factors still contribute to 93% of road crashes (Khattak *et al.* 2021a). The study has also proposed a systematic taxonomy of naturalistic driving errors and violations to understand the nature and role of various driver behaviors in crashes (Khattak *et al.* 2021a). Based on the TDEV, driving behaviors could be classified into 1) Recognition errors, 2) Decision errors, 3) Performance errors, 4) Physical conditions-related errors, 5) Experience or exposure errors, and 6) Violations (Khattak *et al.* 2021a). Recognition errors were found to be the predominant among the key naturalistic driving errors that

contributed to 38.63% of road crashes, followed by decision errors (34.32%) and violations (9.06%) (Khattak *et al.* 2021a). Building upon the TDEV, One of the recent studies have applied rigorous path analysis to see how roadway and environmental factors may relate to SCEs both directly and through driving errors and violations (Ahmad *et al.* 2021). This study provided some useful insights. For instance, it was found that specific roadway environments (e.g., urban areas) significantly increases the crash risk not only directly but also through recognition error, decision errors, and performance error as the chance of each of these driving errors was significantly higher in urban areas (Ahmad *et al.* 2021). The findings suggest that there is a 6.44% higher crash risk in urban areas; where the urban area can also induce an additional crash risk of 2.16% through recognition errors due to which the total increase in crash risk in urban areas could be 8.60% (Ahmad *et al.* 2021). These studies provide useful insights about the effects of the roadway and environmental factors and driving errors and violations on SCEs (Ahmad *et al.* 2021, Khattak *et al.* 2021a). However, besides roadway environment and driving errors and violations, we consider that driving volatility (instability) could significantly relate to SCEs which was not captured in the aforementioned study. To this effect, some studies have investigated the role of driving instability in SCEs (Wali *et al.* 2019, Wali and Khattak 2020); however, these studies have not included the naturalistic driving errors and violations in the analysis which could significantly induce SCEs as discussed earlier. The aforementioned studies reveal that volatile or unstable drivers have a higher chance of crash and near-crash (Wali *et al.* 2019, Wali and Khattak 2020). Referring to the effects of driver distraction (seconds) on SCEs, studies have found that the duration of distraction positively correlates with SCEs (Ahmad *et al.* 2021). In conclusion, previous studies provide useful insights into how one or more human factors relate to SCEs in a naturalistic driving environment, there is a need to understand the role of all human factors (driving errors, violations, duration of distraction, and instability in driving) in SCEs in a more comprehensive manner. Importantly, all of the relevant studies have conducted an empirical investigation to quantify the effects of human factors on SCEs; however, the key issue with conventional statistical is that these models are weaker in prediction. On other hands, machine learning methods show promising performance in both in-sample and out-of-sample prediction.

Research Gaps

In summary, the aforementioned discussion reveals that previous studies provide useful inferential insights of how one or more human factors including driving errors, violations, instability in driving, and duration of distraction relate to SCEs while applying statistical models. However, some research gaps still exist. First, while statistical models may help in understanding the nature and magnitude of association of human factors with SCEs, these models show weaker performance in prediction. On other hand, the application of interpretable machine learning methods not only helps in generating useful inferential knowledge of how

human factors relate to SCEs but can improve out-of-sample prediction accuracy. Second, previous studies have applied statistical models to understand the relation of one or two human factors and not all of the aforementioned human factors in the analysis to look at the role of human factors in SCEs in a comprehensive manner.

Objectives and Contribution

Bridging the aforementioned research gaps, the key focus of this study is to apply various ML/AI methods to predict SCEs based on all of the available human features including driving errors, violations, driving instability, and distraction duration. Note that this study also considers some of the key control factors including intersection influence and roadway environments in the analysis to investigate the risk of crash and near-crash in diverse roadway environments. First, the study estimates an ordered Probit model to understand how human factors relate to the event outcome (baseline, near-crash, and crash). Next, the study applies three key ML methods including Naïve Bayes, K-Nearest Neighbor (KNN), and gradient boosting tree (GBT) classifiers to predict the event outcome based on the human factors. After training the statistical model and machine learning methods to the training data, the study accesses the out-of-sample prediction accuracy for each model to select the most accurate model/method which could predict the event outcome using the real-world information about human factors in the SHRP2-NDS data. The application of ML methods to predict SCEs based on human factors while using real-world SHRP2-NDS data may provide an opportunity to understand the role of all of these factors in predicting the SCEs in new data. Note that after checking the out-of-sample prediction accuracy of the ML methods may also provide a foundation for further research related to developing and implementing such algorithms in instrumented (subject) vehicles in the future which could compute the potential risk of crash and near-crash based on real-world information about driver behavior and performance in diverse roadway environments.

Understanding Effects of Human Factors on Pedestrian Injury Severity Given a Crash

Literature Review

Human factors including driving errors and violations are found to be the key contributing factors in road crashes (Treat *et al.* 1979, Singh 2015, Khattak *et al.* 2021a). Studies suggest that in more than 90% of road crashes, it is human factors that contribute (Treat *et al.* 1979, Singh 2015, Khattak *et al.* 2021a). Rather than focusing on individual risky driving behaviors, studies developed taxonomies of driving errors and violations through different themes which provide a deeper understanding about different driving errors and violations and their contribution to crashes (Treat *et al.* 1979, Wierwille *et al.* 2002, Stanton and Salmon 2009, Thomas *et al.* 2013, Singh 2015, Khattak *et al.* 2021a). For instance, driving behaviors could be

classified into recognition errors, decision errors, performance errors, physical condition (or impairment) errors, and experience errors (Khattak *et al.* 2021a). Violations including intentional or unintentional violations could be placed in a separate category (Khattak *et al.* 2021a). While providing useful knowledge about the contribution and classification of pre-crash driving behaviors, these studies had a broader focus using overall road crashes. Driving behaviors could be significantly different if a driver had to interact with vulnerable road users (VRUs) including pedestrians, bicyclists, and motorcyclists. Interaction with VRUs and especially pedestrians may change driver expectancy, recognition, decision, and performance. Studies also suggest a significant positive correlation between pedestrian and driver behaviors (Şimşekoğlu 2015). Besides driving behaviors, there could be pedestrian behavioral factors which could significantly contribute to crashes. Studies have explored how individual driving behaviors and pedestrian behaviors contribute to crashes (Holubowycz 1995, Miles-Doan 1996, Öström and Eriksson 2001, Pour-Rouholamin and Zhou 2016, Kemnitzer *et al.* 2019).

Studies reveal that drunk driving is found to be a significant contributing factor that positively correlates with pedestrian injury given a crash (Pour-Rouholamin and Zhou 2016). Studies also suggest that driving distraction, speeding, and violation of traffic signals positively influence injury severity given a pedestrian-vehicle crash (Casado-Sanz *et al.* 2019). Studies also suggest that aggressive driving behaviors such as speeding significantly increase the chance of most severe pedestrian injury given a crash (Tefft 2013). For instance, the risk of severe pedestrian injury is 10% at an impact speed of 16 miles per hour which becomes 90% if the impact speed is 46 miles per hour (Tefft 2013). Referring to pedestrian behaviors, studies suggest that older or intoxicated pedestrians and those having reduced reaction or vision due to weather or lighting conditions when involved in crashes are more likely to have a higher injury (Lee and Abdel-Aty 2005). Other studies reveal similar findings which suggest that the chance of a fatality or serious injuries for pedestrians increases in a crash if pedestrians are intoxicated (Holubowycz 1995, Miles-Doan 1996, Öström and Eriksson 2001, Pour-Rouholamin and Zhou 2016, Kemnitzer *et al.* 2019). Studies also reveal that the chance of a pedestrian being injured increases if a pedestrian is distracted (Stavrinos *et al.* 2009). Pedestrian distraction specifically due to cell phone use has a higher odd of involving in an injury crash (Stavrinos *et al.* 2011, Byington and Schwebel 2013, Nasar and Troyer 2013). Studies also explored how pedestrian behaviors related to traffic violations especially traffic or pedestrian signals relate to pedestrian injury (Poudel-Tandukar *et al.* 2007). Pedestrian darting is also more likely to get injured in a crash (Kemnitzer *et al.* 2019).

Research Gaps

The studies provide useful knowledge about how pre-crash behaviors of both pedestrians and drivers relate to pedestrian injury given a pedestrian-vehicle crash; however, few research gaps exist to the best of our

knowledge. For instance, most of the studies have looked at how individual pre-crash behaviors of drivers or pedestrians relate to pedestrian injury. We consider that there is a need first to bring pre-crash behaviors of both drivers and pedestrians into a systematic framework which could be then used to see their correlation with pedestrian injury. Secondly, there is a need to see how different combinations of impairment by pedestrians and drivers involved in pedestrian crashes relate to pedestrian injury given a crash that has been lightly addressed.

Objectives and Contribution

To bridge the research gaps, this study aims to achieve the objectives given below.

- 1). To develop systematic frameworks for the classification of pre-crash behaviors of both drivers and pedestrians involved in pedestrian-vehicle crashes.
- 2). To explore how key errors, violations, and impairment by both drivers and pedestrians correlate with pedestrian injury.
- 3). To identify accurate estimates and predictions through finite mixture (latent class) models which could capture the heterogeneity in pedestrian injuries across various classes based on some unobserved variable or set of variables (impact speed, conspicuity of pedestrians, aggressiveness of both pedestrian and driver).

Besides the rigorous conceptual design, a unique methodological aspect of this study is that it applies a finite mixture model to capture the potential unobserved heterogeneity in pedestrian injuries (given a crash) while accounting for latent (unobserved) class variables or a set of class variables. For instance, the NCPBCAT data include information about lighting conditions and not about the conspicuity of the pedestrian involved in a crash. Based on this, this study hypothesizes that chance of a fatality or serious injuries for a pedestrian being killed in a time a crash with bright color clothes may be different (most probably lower) than those with dark color clothes. In case of unavailability of such data, the application of finite mixture ordered Probit regression to model pedestrian injury could help in understanding the difference in the effects of light conditions on pedestrian injury severity while accounting for the unobserved class. Note that if latent classes in the data based on some unobserved factors exist but are not taken into account, the results obtained could be less accurate and biased. While realizing the conceptual importance of the finite mixture models in pedestrian crashes, this study contributes by applying the finite mixture ordered Probit regression to see how behaviors of drivers and pedestrians, and roadway and environment factors relate to pedestrian injury using unique NCPBCAT data.

Methodological Framework

This dissertation aims to perform safety analysis, using both naturalistic driving study and traditional crash data., at various levels. Figure 1.5 illustrates the overall framework of the proposal. Referring to the data, we used real-world SHRP2-NDS data and conventional crash data (NC PBCAT). The data sets were pre-processed and prepared before conducting thorough exploratory analyses. NC PBCAT data was processed and cleaned with more care because the basic source of this data was police crash reports.

First, systematic taxonomy is developed for real-world pre-crash driving behaviors using SHRP2-NDS data which is then extended and modified for both drivers and pedestrians using NC PBCAT data. Referring to the driving instability measures in SHRP2-NDS data analysis, various indices (e.g., coefficient of variation and quartile coefficient of variation) of driving instability were computed using SHRP2-NDS data. Referring to the safety analysis, real-world SHRP2-NDS data were used to conduct a multi-layered in-depth investigation to explore the role of human factors in SCEs and to explore the pathways which run from diverse roadway environment, distraction duration, and other key factors to SCEs through driving errors, violations, and driving instability in naturalistic driving (Figure 1.5). Several types of path analysis including standard path analysis, path analysis through the joint estimation, and path analysis through predictive margins are used to explore the key pathways. While conducting path analysis various types of constituent models including ordered probit, partial proportional odds, Tobit, and multinomial logit models are used to explore the key association of explanatory variables with the response variables. To understand the role of human factors (driving errors, violations, duration of distraction, and instability in driving) in predicting SCEs and to select the most accurate method for predicting SCEs, more sophisticated machine learning methods including naïve Bayes, K-Nearest Neighbors, and gradient boosting tree classification algorithms are applied. After thoroughly investigating the role of human factors in SCEs using SHRP2-NDS data, the dissertation aimed to understand how human factors (pre-crash behaviors of both pedestrians and drivers) relate to pedestrian injury severity. Due to the limited sample size of pedestrian-vehicle crashes in SHRP2-NDS data, we have selected a comprehensive North Carolina PBCAT dataset for analysis. To explore the effect of human factors (errors, violations, and impairment by both drivers and pedestrians) on pedestrian injury severity, first, risky behaviors by both driver and pedestrian are classified. Next, a finite mixture ordered probit model is used to account for heterogeneity in the effects of observed variables (especially human factors) on pedestrian injury severity across different unobserved (latent) classes of pedestrians.

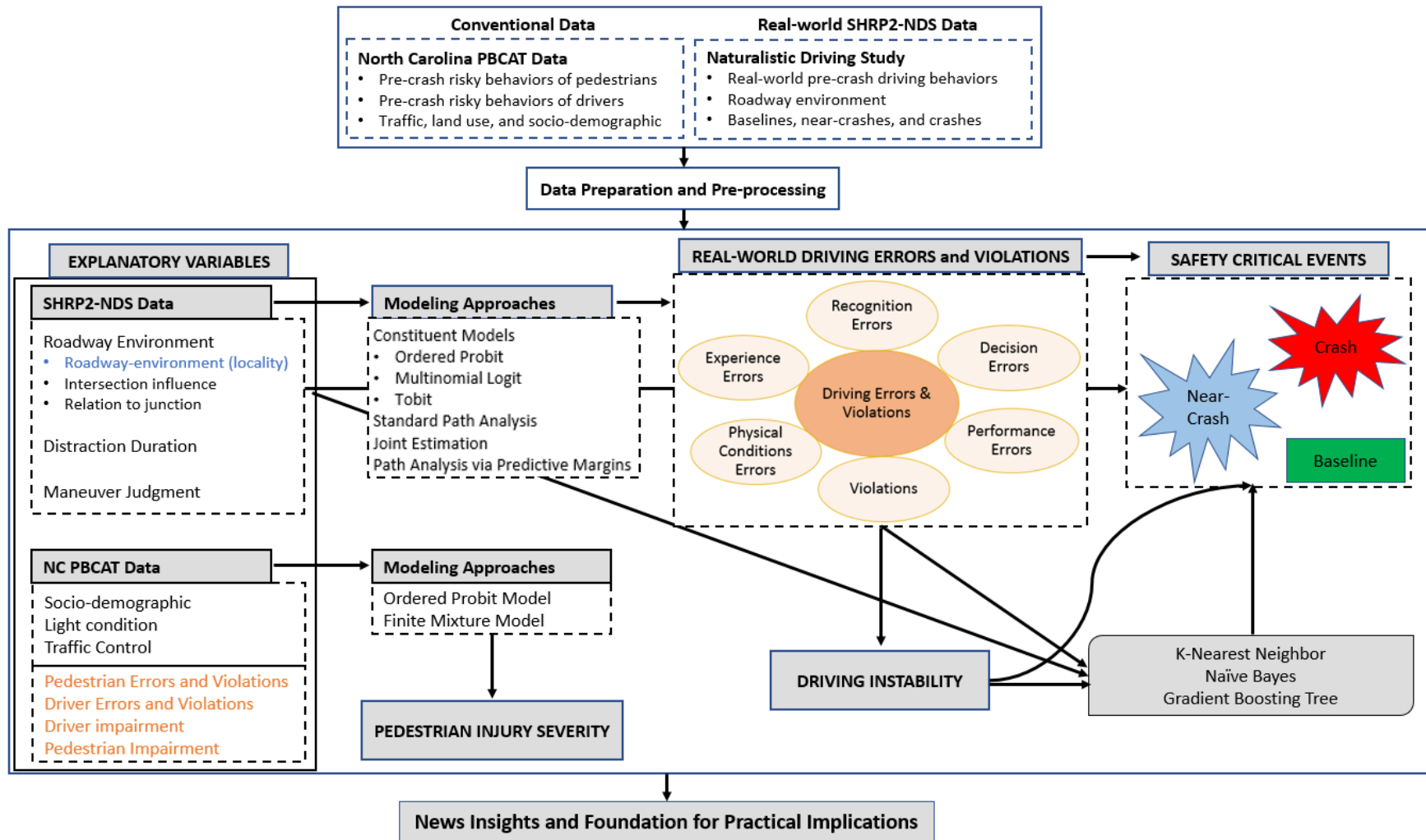


Figure 1. 5 Proposal Methodological Framework

Unique Datasets and Deriving Human Factors

This section discusses the significance of the data analyzed in this dissertation. The dissertation includes a total of five chapters. The first four chapters (Chapter 2-5) uses the SHRP2-NDS data to understand the role of human factors including driving errors, violations, duration of distraction, and instability in driving in SCEs. The final chapter (Chapter 6) aims to understand how pre-crash behaviors of both drivers and pedestrians along with other roadway environment factors relate to pedestrian injury severity (given a crash) while analyzing NCPBCAT data. For a better understanding of the unique aspects of the SHRP2-NDS and NCPBCAT data and how human factors are derived, we provide details in the subsequent paragraphs.

SHRP2-NDS Data: Deriving Human Factors

The SHRP2-NDS data is a unique data source that includes real-world data on pre-crash driver behaviors and performance, secondary tasks (along with their duration), vehicle kinematics (speed, acceleration, and deceleration), and other key roadway and environment variables (Dingus et al. 2006, Hankey *et al.* 2016). In the SHRP2-NDS data, second-to-second driver behavior, performance, eye glance behavior, and other key details about the subject (in instrumented vehicle) drivers were recorded via fitted cameras and recording equipment (Dingus et al. 2006, Hankey *et al.* 2016). The SHRP2-NDS also includes data about instantaneous speed, acceleration, and deceleration (Dingus et al. 2006, Hankey *et al.* 2016) which can be used to derive various indices of instability in driving like the coefficient of variation, standard deviation, and quantile coefficient of variation of speed/acceleration/deceleration/jerk (Arvin *et al.* 2019). To extract and code the data from the video streams, the Virginia Tech Transportation Institute (VTTI) selected a team of data reductionists, through a thorough selection procedure, to ensure quality data extraction (Dingus et al. 2006, Hankey *et al.* 2016). One of the unique aspects of the SHRP2-NDS data is that it includes data on baselines (non-event driving) that could be used to quantify crash risk (Ahmad *et al.* 2021, Khattak *et al.* 2021a). Note that the police crash data do not include data about the baselines and could also be subjective for other variables as the police officer may not be present at the time of the crash (Dingus et al. 2006, Arvin *et al.* 2019, Ahmad *et al.* 2021, Khattak *et al.* 2021a). To have a further understanding of how baseline, near-crash, and crash are defined by the SHRP2-NDS data reductionists, we provide their definitions as per the NDS data dictionary below.

- **Baseline:** “An epoch of data selected for comparison to any of the conflict types listed above rather than due to the presence of conflict. Refer to the unique baseline selection criteria for the set of baselines being analyzed. For SHRP2, these baselines are 21 seconds long and were randomly selected with a goal of 20,000 baselines, a minimum of 1 baseline per driver, and the number of baselines for each driver proportional to the total driving time > 5mph for each driver. Baselines were selected only if vehicle speed did not dip below 5mph for more than 2 consecutive seconds.”

- **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. Near Crashes must meet the following four criteria: 1. Not a Crash. The vehicle must not make contact with any object, moving or fixed, and the maneuver must not result in a road departure. 2. Not pre-meditated. The maneuver performed by the subject must not be pre-meditated. This criterion does not rule out Near Crashes caused by unexpected events experienced during a pre-meditated maneuver (e.g., a premeditated aggressive lane change resulting in a conflict with an unseen vehicle in the adjacent lane that requires a rapid evasive maneuver by one of the vehicles). 3. Evasion required. An evasive maneuver to avoid a crash was required by either the subject or another vehicle, pedestrian, animal, etc. An evasive maneuver is defined as steering, braking, accelerating, or a combination of control inputs that is performed to avoid a potential crash. 4. Rapidity required. The required evasive maneuver must also require rapidity. Rapidity refers to the swiftness of the response required given the amount of time from the beginning of the subject’s reaction and the potential time of impact.”
- **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. Also includes non-premeditated departures of the roadway where at least one tire leaves the paved or intended travel surface of the road.”

As discussed earlier, the first four chapters (Chapters 2-5) in this dissertation mainly analyze the human factors (driving errors, violations, duration of distraction, and instability in driving). To compute the instability (volatility) in driving in the SHRP2-NDS data, the information about instantaneous speed, acceleration, and deceleration can be used. Some of the volatility indices which are derived include the coefficient of variation, quartile coefficient of variation, and standard deviation of speed, acceleration, deceleration, and jerk – for details, please refer to Chapter 4 and Chapter 5 in this dissertation. To understand how this dissertation derives driving errors, violations, and duration of distraction from the original variables in the SHRP2-NDS data, please see below.

Driving Errors and Violations

One of the key human factors analyzed in this dissertation is driving errors and violations which are derived from the two variables including “*driver behavior*” and “*secondary task*” originally coded in the SHRP2-NDS data. A systematic taxonomy based on the theoretical perception-reaction process is developed (see Chapter 2 for details) which brings all pre-crash driver behaviors and secondary tasks (given that a specific secondary task has contributed to crash) into a systematic framework. Note that pre-crash driver behaviors are defined in the SHRP2-NDS data dictionary to have contributed to crashes. To determine whether a

specific secondary task has contributed to a crash, this dissertation accesses the “secondary task outcome” coded in the SHRP2-NDS data. To conclude, this dissertation derives driving errors and violations from the pre-crash driver behavior and secondary task (given that it has contributed by accessing the “secondary task outcome” variable) through the TDEV developed in Chapter 2. To understand the nature and definition of the two key variables used in deriving various driving errors and violations, please see the following definitions from the SHRP2-NDS data dictionary as below.

- **Driver Behavior**

The “Driver Behavior” factor in NDS data is defined by data reductionists as: “Driver behaviors (those that either occurred within seconds before the Precipitating Event or those resulting from the context of the driving environment) that include what the driver did to cause or contribute to the crash or near-crash. Behaviors may be apparent at times other than the time of the Precipitating Event, such as aggressive driving at an earlier moment which led to retaliatory behavior later.”

- **Secondary Task and Secondary Task Outcome**

In some crashes, no contributing driver behavior was reported. In such cases, we checked the “Secondary Task” and “Secondary Task Outcome” factors in the NDS data to determine whether a secondary task contributed to a crash. The “secondary task outcome” in the NDS data is defined by the NDS data reductionists as: “Determination of whether the Secondary Task contributed to the event sequence and severity. (Not whether the factor caused the event but contributed to it.)” These secondary tasks include cell phone use, talking/listening, and using hands-free communication, non-specific internal eye glance, and doing personal hygiene while driving.

Through the TDEV, the data about the pre-crash driver behaviors and secondary tasks coded in the SHRP2-NDS data was then used to develop the TDEV which classify driver behaviors into 1) recognition errors, 2) decision errors, 3) performance errors, 4) physical condition-related errors, 5) experience or exposure errors, and 6) violations. After obtaining various driving errors and violations through the systematic TDEV, safety analyses are conducted in Chapter 2, Chapter 3, Chapter 4, and Chapter 5 to thoroughly investigate the role of these errors and violations in SCEs.

Duration of Distraction

Studies reveal that drivers distracted for a longer duration have a higher chance to involve in a crash. In the SHRP2-NDS data, the information about secondary tasks and their duration is extracted from the videos based on the involvement of the subject driver in distracting activities like talking to fellow passengers, texting or talking via cellphone, eating/drinking. Note that all such secondary activities may distract driver

attention from the primary driving task and may lead to a crash or near-crash. In particular, due to distracted driving, a driver may not be able to timely redirect his/her gaze to the roadway which could lead to SCEs. To see how distracted driving based on the driver gaze was determined in the SHRP2-NDS, please see Figure 1.6 below (Campbell 2012). In the left-side picture (screen shot taken from a video), the driver focuses his gaze on the roadway whereas on the right-side the same driver has taken his gaze off the roadway due to involvement in some distracting activities (Campbell 2012).

In the SHRP2-NDS data, the two variables including “secondary task start time” and “secondary task end time” are coded which could be used in computing the secondary task duration. Note that the “secondary task start time” refers to the time when the subject driver took off his/her gaze of the roadway due to his/her involvement in some distracting activities. The SHRP2-NDS also captures the time instant (in the video) when the subject driver has focused the gaze on the roadway which is termed as “secondary task end time”. For the duration of distraction (secondary task duration), this dissertation computes the difference of “secondary task start time” and “secondary task end time” both of which are coded in the SHRP2-NDS data. For the sake of completeness, please see the definitions of the two times as per the SHRP2-NDS data dictionary below.

- **Secondary Task Start Time**

“The time at which the driver began to engage in the secondary task. This is a specific integer value for the video timestamp in milliseconds from the start of the file. Only secondary tasks that occur during or overlap the period of time starting 5 seconds prior to the Precipitating Event through Event End are included. If the secondary task began more than 5 seconds before the Precipitating Event), then enter the Event Start (Variable 2) timestamp minus 5 seconds (5,000 timestamps).”

- **Secondary Task End Time**

“The time at which the driver disengaged from the secondary task or the driver's attention returned to the driving task or another activity. This is a specific integer value for the video timestamp in milliseconds from the start of the file. Only distractions that occur during or overlap the period of time starting 5 seconds prior to the Precipitating Event through Event End are included. If the secondary task continued after the Event End, then enter the Event End (Variable 5) timestamp.”

Note that some of the subject drivers were involved in more than 1 secondary task (maximum up to 3 secondary tasks); however, the percentage of these drivers was lower. Also, the duration of the first secondary task (Secondary Task 1) is found to be longer for all the drivers who were involved in multiple secondary tasks; therefore, for deriving the secondary task duration, this dissertation considers the start and

end time of the secondary tasks 1. Note that in the analyses, the aforementioned human factors were first computed or computed and then used in the empirical analysis. Regarding the selection of the variables in the empirical analysis, it happened that all the human factors including driver errors, violations, duration of distraction, and instability in driving have shown statistically significant and intuitive correlation with the SCEs (for details, please refer to Chapter 3-5). In summary, the key three human factors selected and included in the analysis (Chapter 2-5) are based on the conceptual design and significance. Also, all of the aforementioned factors showed statistical significance and expected intuitions with the SCEs due to which these factors are included in the analysis. Besides, the key human factors, other key variables include diverse roadway environment, intersection influence, relation to the junction, and roadway configuration are also analyzed in this study to see how these factors may relate to driving errors (Chapter 2-3), instability in driving (Chapter 4), and SCEs (Chapter 2-5) – for details about the definitions and selection of these factors, please refer to the individual chapters.

NCPBCAT Data: Deriving Human Factors

In this study, we analyze a comprehensive database that includes all the necessary crash-related factors (indicating when, where, who, why, and in which conditions) the pedestrian-vehicle crashes have occurred across North Carolina State. The data was derived from the police crash reports. However, crash type and some additional unique attributes were specially coded by professional data experts in the PBCAT tool using the crash narrative in the police crash reports (Levitt). While police-reported pedestrian-involved crash data can have gaps in crash details, the PBCAT provides a more comprehensive high-quality database of pedestrian-vehicle crashes. The PBCAT is a crash typing software that is intended to help local and state bicycle/pedestrian coordinators, engineers, and planners in improving the safety of pedestrians and bicyclists by providing a comprehensive database of pedestrian-vehicle and bicyclist-vehicle crashes occurring across North Carolina (Levitt). The “crash type” variable in the PBCAT data includes details about the risky behaviors of drivers and pedestrians who were involved in pedestrian-vehicle crashes (Levitt). To illustrate some of the pre-crash risky behaviors by pedestrians and drivers in pedestrian-vehicle crashes across North Carolina, please refer to Figure 1.7 and Figure 1.8 respectively (borrowed from the official website of NCPBCAT). Similar to the SHRP2-NDS data, this dissertation uses the “crash type” variable to first identify all the pre-crash risky human behaviors in pedestrian-vehicle crashes in North Carolina which are then systematically classified into various errors and violations by drivers and pedestrians (for details, please refer to Chapter 6).



Figure 1. 6 Illustration of SHRP2-NDS Video: Capturing Driver Distraction and Gaze (Source: (Campbell 2012))

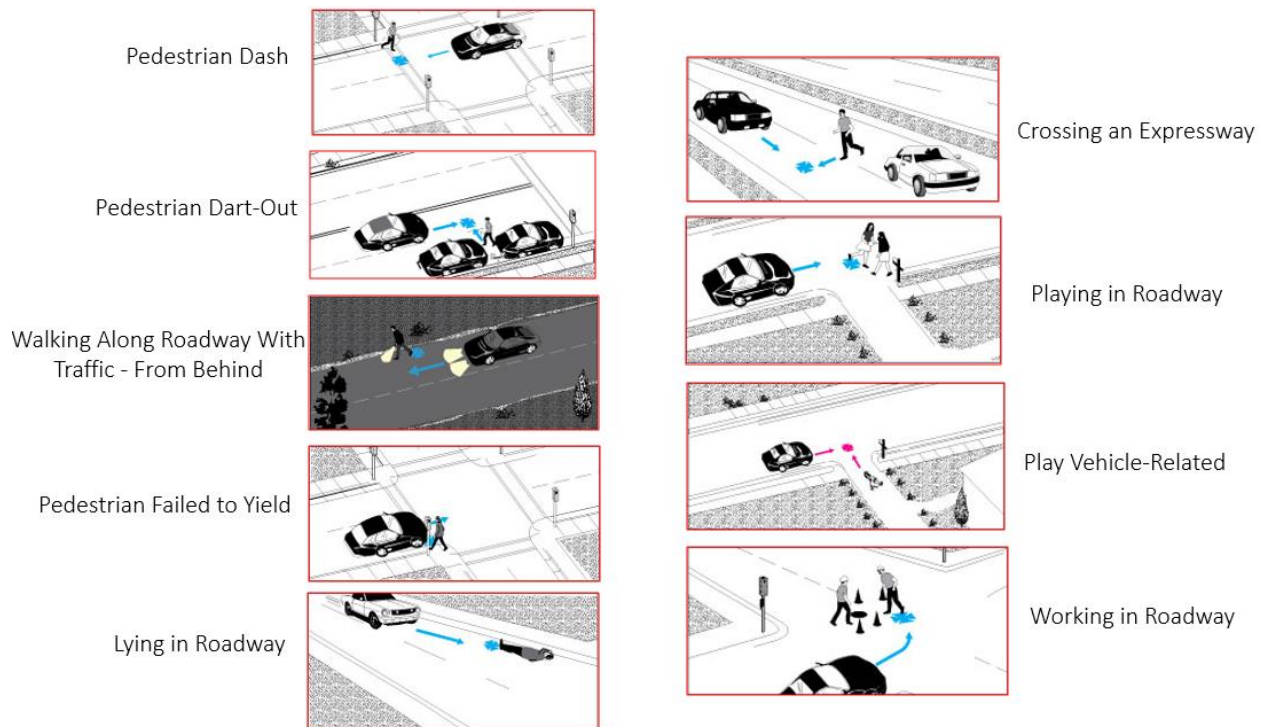


Figure 1. 7 Pre-Crash Risky Behaviors by Pedestrians: Coded in Crash Type Variable (NCPBCAT)

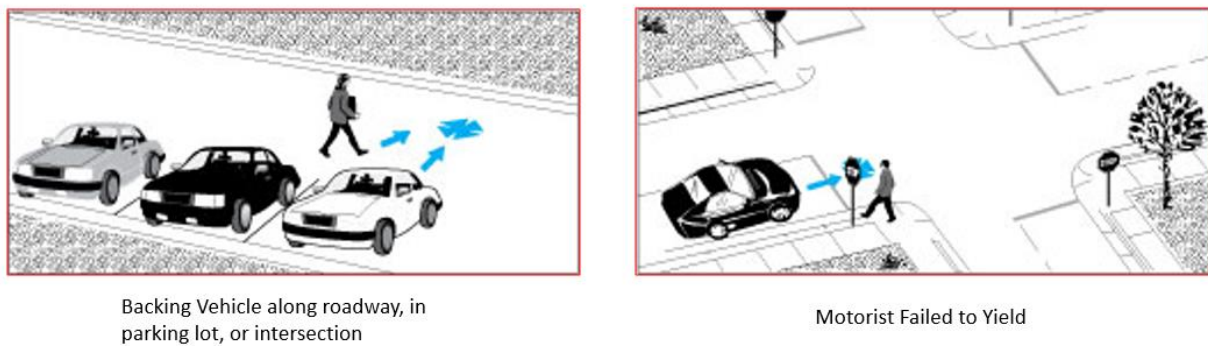


Figure 1. 8 Pre-Crash Risky Behaviors by Drivers: Coded in Crash Type Variable (NCPBCAT)

Other interesting variables in the PBCAT data includes impairment (due to alcohol or drugs) status of pedestrians and drivers, location or position of a pedestrian at the time of crashes, roadway factors (functional classes of roadways, roadway configuration), lighting conditions, speed limit, land use and types of surrounding development, weather conditions, geocoded location of the crash, and demographics details (e.g., age, gender, and race) of pedestrians and drivers who were involved in pedestrian-vehicle crashes (Levitt). We have selected pedestrian-vehicle crashes for the analysis which has occurred between 2009-2019 on roadways across North Carolina. The injury severity of both pedestrians and drivers is reported on a five-level KABCO scale (Levitt). Several other control variables related to the location/position of pedestrians, presence/types of traffic control devices, lighting conditions, and the speed limit are also included in the analysis.

In the future, I plan to extend TDEV (chapter 1) to motorcycle-vehicle crashes and explores how various combinations of errors, violations, and impairment by rider and driver are related to rider injury given a crash. I plan to select and use an appropriate crash database that includes detailed information about pre-crash behaviors of both rider and driver involved in motorcycle-vehicle crashes. The key hypothesis for future research is that the severity of rider (vulnerable road user) injury may significantly increase if both rider and driver are involved in certain errors, violations, or impairment simultaneously. Furthermore, I plan to explore the effect of built-environment, light conditions, and roadway and intersection types on different combinations of errors simultaneously made by rider and driver in motorcycle-vehicle crashes (Figure 1.9). Given the proposed framework, I aim to identify hazardous locations and conditions which encourage or discourage certain types of errors and violations by the two parties which could generate new knowledge and help in identifying effective countermeasures to enhance motorcycle safety. To achieve study objectives, I plan to apply both k-prototype clustering and rigorous path analysis.

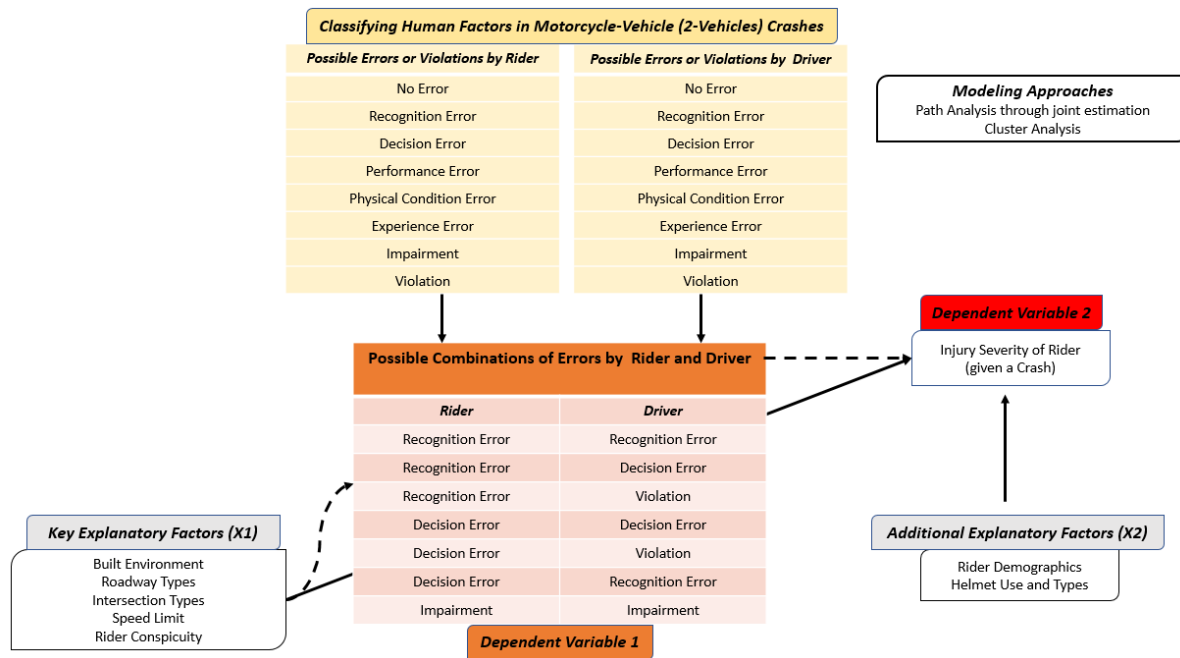


Figure 1. 9 Proposed framework for future study

CHAPTER 2. A TAXONOMY OF DRIVING ERRORS AND VIOLATIONS: EVIDENCE FROM THE NATURALISTIC DRIVING STUDY

A version of this chapter was published in the Journal of Accident Analysis and Prevention. Also, a version of this chapter was presented at the 100th Annual Meeting of the Transportation Research Board, Washington D.C. (Jan 2021).

Khattak, A. J., Ahmad, N., Wali, B., & Dumbaugh, E. (2021). A taxonomy of driving errors and violations: Evidence from the naturalistic driving study. *Accident Analysis & Prevention*, 151, 105873.

Abstract

Driving errors and violations are identified as contributing factors in most crash events. To examine the role of human factors and improve crash investigations, a systematic taxonomy of driver errors and violations (TDEV) is developed. The TDEV classifies driver errors and violations based on their occurrence during the theoretically based perception-reaction process and analyzes their contributions in safety-critical events. To empirically explore errors and violations in diverse built environments, this study harnesses unique and highly detailed pre-crash sensor data collected in the SHRP2 Naturalistic Driving Study (NDS), containing 673 crashes, 1,331 near-crashes, and 7,589 baselines (no-event). Human factors are categorized into recognition errors, decision errors, performance errors, and errors due to the drivers' physical condition or their lack of contextual experience/familiarity, and intentional violations. In the NDS data, built environments (measured by roadway localities) are classified based on roadway functional classification and land uses, e.g., residential areas, school zones, and church zones. Based on the percentage of crashes per percentage of baselines in a specific locality, interstates and open country/open residential (rural and semi-rural settings) may pose lower risks. Contrarily, urban, business/industrial, and school locations seem to have a higher percentage of crashes per percentage of baselines indicating higher crash risk. Human errors and violations contributed to 93% of the observed crashes, while roadway factors contributed to 17%, vehicle factors contributed in 1%, and 4% of crashes contained unknown factors. The most common human errors were recognition and decision errors, which occurred in 39% and 34% of crashes, respectively. These two error types occurred more frequently (each contributing to ~39% of crashes) in business or industrial land use environments (but not in dense urban localities). The findings of this study reveal the continued prevalence of human factors in crashes. The distribution of driving errors and violations across different roadway environments can aid in the implementation of intelligent transportation system (ITS) based driver assistance systems and place-based countermeasures which can potentially reduce these driving errors/violations and crash risk.

Introduction

Human factors are found to be the key contributing factors in crashes (Sabey and Staughton 1975, Treat *et al.* 1979, Frawley and Eisele 2004, Huang *et al.* 2010, Klauer *et al.* 2014, Singh 2015). Studies reveal that human factors contribute to more than 90% of road crashes (Sabey and Staughton 1975, Treat *et al.* 1979, Frawley and Eisele 2004, Huang *et al.* 2010, Klauer *et al.* 2014, Singh 2015). All of the aforementioned studies provide valuable information; however, the studies were based on police crash reports which could be subjective and bias (Hankey *et al.* 2016, Wali *et al.* 2018e). Moreover, there could be other inaccuracies such as underreporting of less severe (property damage only) crashes (Yamamoto *et al.* 2008, Washington *et al.* 2010, Ye and Lord 2011). This study seeks to overcome these limitations by using Naturalistic Driving Survey (NDS) data. NDS data is collected from instrumented vehicles that record the behaviors of drivers, including their speed, breaking performance, and video recordings of a driver's actions, providing the ability to directly observe real-world information from crashes of all severity levels, near-crashes, and baselines (Hankey *et al.* 2016). In addition, the chances of losing important information associated with traditional data protocols are minimized. The NDS data provides information on real-world driving behavior and performance, along with real-world risks and safety consequences (Dingus *et al.* 2006, Carney *et al.* 2015a, Dingus *et al.* 2016, Arvin *et al.* 2019). This facilitates the examination of human factors in greater detail, e.g., errors due to driver inattention, distraction, drowsiness, and judgment-related errors, all of which contribute to crashes.

Previous studies have developed taxonomies for human factors including driving errors and violations through different themes (Treat *et al.* 1979, Reason *et al.* 1990, Petridou and Moustaki 2000, Wierwille *et al.* 2002, Stanton and Salmon 2009). Driver errors were classified as slips, lapses, or mistakes (Reason *et al.* 1990). In another study, behavioral factors affecting driving capabilities were differentiated from those encouraging risky driving (Petridou and Moustaki 2000). From previous taxonomies, key driving errors include recognition errors, decision errors, and performance errors (Treat *et al.* 1979, Wierwille *et al.* 2002). Recognition errors and decision errors were found to be the most prevalent driving errors contributing to crashes (Treat *et al.* 1979, Wierwille *et al.* 2002, Iden and Shappell 2006). Due to variations in driver characteristics and behavioral patterns, different roadway environments can induce different errors and violations (Dumbaugh and Li 2010). To conclude, previous taxonomies provide useful insights; however, none of the studies (to the best of our knowledge) has used real-world naturalistic driving data which provide detailed information on pre-crash driving behaviors in crashes, near-crashes, and baseline driving. Having said this, this study aims to develop a systematic taxonomy of driving errors and violations (TDEV) and then explores how the key driving errors and violations vary across diverse roadway environments. Importantly, this study sheds light on the risks associated with driving errors and violations made during

no-event driving situations (baselines), which can potentially lead to safety-critical events (crash or near-crash) in the future.

Methodology

Data Source

This study uses data from the Naturalistic Driving Study which were collected as part of the 2nd Strategic Highway Research Program (SHRP 2). The NDS SHRP2 is a comprehensive data collection effort in which a variety of volunteer drivers participated. Thousands of vehicles were equipped with advanced technologies (i.e., radars, sensors, and cameras) that continuously monitored and collected data on driver performance, driver behavior, speed, acceleration, lateral and longitudinal positions, and eye glancing behavior (Dingus *et al.* 2016). The NDS data consists of high-frequency and high-resolution information that was collected through an onboard data acquisition system (DAS) and multiple sensors (Hankey *et al.* 2016). The data includes roughly 3,500 drivers from six different states including New York, Washington, Pennsylvania, North Carolina, Indiana, and Florida (Hankey *et al.* 2016). The data were carefully collected in the appropriate study centers located in Bloomington, Indiana; Buffalo, New York; Durham, North Carolina; Seattle, Washington; State College, Pennsylvania; and Tampa, Florida. According to the NDS reports, these study centers encompassed several counties each with more than 21,000 mi² hosting about 7.6 million registered vehicles of all types and had a population of approximately 6.5 million people of driving age (greater than 15 years)-see Appendix Table A0 in (Khattak *et al.* 2021a) for geographic details of counties used for NDS recruitment. Efforts were made to have a representative sample of the population living in the counties surrounding the study areas, which included both rural and urban areas. The study design focused on obtaining a representative sample of the driving population by age and gender. More details are provided in relevant SHRP reports (Blatt *et al.* 2015). A total of 3,247 volunteer participants were recruited for the NDS (SHRP-2) showing a good distribution among the aforementioned six regions (Blatt *et al.* 2015). Out of these 3,247 participants, the distribution was Bloomington 7.82% (N=254), Buffalo 22.79% (N=740), Durham 16.29% (N = 529), Seattle 22.02% (N = 715), State College 8.47% (N = 275), and Tampa 22.61% (N =734) (Blatt *et al.* 2015)—see Table A0 in Appendix in (Khattak *et al.* 2021a). Participants included around 45.80% (N = 1,487) males and 50% (N = 1,624) females, which were fairly distributed among various age groups which include: minor teens (with 16-17 years of age), adult teens (18-20 years), young adult (21-25 years), adult (26-35 years), middle adult (36-50 years), mature adult (51-65 years), older drivers (66-75 years), and older drivers that are 75+ years of age. Details on age distribution are available elsewhere (Blatt *et al.* 2015). Interestingly, there were 4.20% (N = 145) of the vehicles called AVT (Advanced vehicle Technology), equipped with advanced features such as collision

avoidance radar, advanced cruise control, and electronic stability control (Blatt *et al.* 2015). Participants using these vehicles could be of any age or gender; for details, see Blatt *et al.* (2015). While the NDS (SHRP2) data includes detailed information on all the aforementioned participants, these variables (e.g., driver age, driver gender, and detailed information about advanced features of AVT vehicles) were not available in the event sub-sample available for this study.

The data also includes detailed information on vehicular movement, driving behavior, different roadways and environments, traffic conditions, and roadway/environmental factors which were all extracted by a well-trained team of data reductionists following a rigorous protocol. The data reductionists employed by the Virginia Tech Transportation Institute (VTTI) were either students of Virginia Tech or were recruited from other nearby universities, colleges, and the surrounding community (Hankey *et al.* 2016). A major portion of the reductionists had backgrounds in human factors engineering and psychology having good attention to detail, solid work ethic, and sound critical thinking skills (Hankey *et al.* 2016). Good attention capabilities or good attention to details refers to the multidimensional perceptive abilities of individuals (data reductionists in this case) which are important when extracting data from the NDS videos. Data reductionists were evaluated according to such capabilities during the recruitment process. This was important so that minute but important details about the subject driver (e.g., pre-crash driving behavior, and type and duration of secondary task), the subject vehicle, or the surrounding environment were well determined and extracted from NDS videos. Since these qualities in an individual cannot be judged while evaluating his/her written application or resume, therefore appropriate recruitment and retention procedures including interviews, trial periods, reference checks, proficiency tests were conducted and those qualifying these tests were hired as data reductionists (Hankey *et al.* 2016). The minimum qualifications for these positions were a high school diploma, the legal age for signing non-disclosure agreements, having done some college-level coursework, computer proficiency, sound critical thinking capabilities, previous data entry or similar experience, and a U.S. driving license which ensured their understanding of various driving scenarios (Hankey *et al.* 2016). The reductionists were required to qualify for a proficiency test before initiating the NDS data coding (Hankey *et al.* 2016). For complete details on NDS data reductionists and data collection, please refer to (Hankey *et al.* 2016). Overall, the data reductionists were well-qualified to create the variables contained in the NDS data.

This study uses a subset of the original NDS-SHRP2 data, including a total of 9,593 events with 7,589 baselines (20 to 30 seconds each), 1,331 near-crashes, and 673 crashes. The baselines (i.e., driving instances when no safety-critical event happened) in the NDS data provide an opportunity to compare them with crashes and near-crashes that represent safety-critical events. Although not fully representative of exposure, the baselines can provide a sense of crash risk through a comparison of pre-crash behaviors

(Hankey *et al.* 2016). The selection of baselines is one of the key new data sources in NDS. Through baselines, researchers can get a sense of typical driving behaviors across the sample (Hankey *et al.* 2016). In order to select baselines, a baseline sampling method was developed by VTTI. The sampling method was presented for expert review to the Expert Task Group (ETG) who conduct a review for the NDS studies as well as to the NDS (SHRP 2) Technical Coordinating Committee (Hankey *et al.* 2016). In NDS, the VTTI team randomly selected the baselines intending to select at least one baseline for every driver (involved in a safety-critical event) with a typical duration of 20-30 seconds. Note that the number of baselines for a specific driver was proportional to his/her total driving time in the NDS experiment (Hankey *et al.* 2016). Driving speeds of lesser than 5 miles per hour were excluded to disregard the influence of long stopping times and to only consider periods when the subject vehicle had the possibility of an at-fault crash (Hankey *et al.* 2016). Accordingly, the NDS subsample used in the study shows the percentage of baseline (79.10%) compared to the percentages of near-crashes (13.87%) and crashes (7.02%). While this is not representative of the driving and crash proportions in the real-world it can be considered as capturing exposure and non-event data due to the fact that a driver may be rarely involved in crashes compared with the miles driven (i.e., hence baselines represent their normal (non-event) driving).

Notably, there is sufficient variation in the baselines based on roadway and land-use variables. The variables included in the NDS data are generally classified into three classes (Hankey, Perez et al. 2016):

1. Safety-critical event variables, e.g., event nature, event severity, precipitating events, pre-incident maneuvers, and drivers' reactions.
2. Driver variables, e.g., driving behavior, type, and duration of driver distraction/secondary tasks, seatbelt usage, and drivers' steering control.
3. Roadway and Land Use Variables, e.g., locality, roadway alignment, traffic flow, traffic density, and traffic control devices (Hankey *et al.* 2016).

To understand how safety-critical events (crashes and near-crashes) were identified in NDS data, we provide the definitions as per the NDS dictionary in the revised manuscript.

Crash: In NDS data, a crash is considered as any contact of a subject vehicle with an object (fixed or moving) at any speed which results in transfer or dissipation of kinetic energy. It also includes non-premeditated departures of the roadway in which at least one tire of the subject vehicle leaves intended travel or paved surface of the roadway. Furthermore, a crash is considered to have occurred if the subject vehicle strikes any other vehicle, object, roadside barrier, animal, bicyclist, or pedestrian.

Near-Crash: Near-crash is considered as a situation where the subject vehicle, or another vehicle, cyclist, pedestrian, or animal requires a rapid evasive maneuver to escape the crash. Note that rapid evasive maneuver includes accelerating, braking, steering, or a combination of these three maneuvers.

Safety Matrix

A safety matrix was developed to provide insight into the prevalence of safety-critical human, vehicle, and roadway factors. The key objective for developing the safety matrix is to quantify the contributions of key factors in several combinations. Only the factors in the NDS data which data reductionists coded and clearly state that they contribute to crashes were considered. Note that the data reductionists who reviewed and analyzed the driving and video data used specific procedures. For instance, the procedures in the General Estimates System (GES) compiled by the National Highway Traffic Safety Administration were used as a reference for developing variables during data extraction from the NDS videos. This ensured that the key aspects of crashes are considered by the data reductionists and consistency in documenting variables. In this study, we use driver behavior (e.g., driving slowly and below the speed limit) and secondary tasks (if it contributed to crashes) to assess the contribution of human factors in crashes. Similarly, various factors in the NDS data were used to quantify the roadway and environmental factors. A detailed discussion on factors in the NDS data, which were used to derive crash contributing factors related to human, roadway/environment, and vehicle are discussed below.

Human Factors

Driver Behavior

The “Driver Behavior” factor in NDS data is defined by data reductionists as: “Driver behaviors (those that either occurred within seconds before the Precipitating Event or those resulting from the context of the driving environment) that include what the driver did to cause or contribute to the crash or near-crash. Behaviors may be apparent at times other than the time of the Precipitating Event, such as aggressive driving at an earlier moment which led to retaliatory behavior later.” For details on driver behavior in the NDS data, please refer to Figure 2.2. Note that the complete list of all driver behaviors reported in the NDS data (i.e., for all event outcomes including crashes, near-crashes, and baselines) and our derivation of six types of driving errors and violations using these driver behaviors are provided in TABLE A0 (Appendix) in (Khattak *et al.* 2021a). Figure 2.2 only illustrates those driver behaviors which contributed to crashes only.

Secondary Task and Secondary Task Outcome

In some crashes, no contributing driver behavior was reported. In such cases, we checked the “Secondary Task” and “Secondary Task Outcome” factors in the NDS data to determine whether a secondary task contributed to a crash. The “secondary task outcome” in the NDS data is defined by the NDS data reductionists as: “Determination of whether the Secondary Task contributed to the event sequence and severity. (Not whether the factor caused the event, but contributed to it.)” These secondary tasks include cell phone use, talking/listening, and using hands-free communication, non-specific internal eye glance, and doing personal hygiene while driving.

Roadway and Environment Factors

The roadway and environmental factors include all those factors which refer to the roadway, weather, and visual obstructions which data reductionists state as crash contributors in NDS data. These roadway and environmental factors are separate from the “locality” factors in the NDS data which simply indicate in which roadway or land use environment the baseline or safety-critical event (near-crash or crash) was observed. The roadway and environment factors, which contributed to crashes, are derived from various factors in the NDS data including those coded as infrastructure, visual obstructions, surface conditions, and environmental conditions discussed as follows:

- **Infrastructure:** The NDS data dictionary defines these factors as possible contributing causes to the occurrence or severity of the safety-critical event (i.e., in this case, a crash). Examples of such factors include roadway alignment, roadway delineation, traffic control device, and roadway sight distance.
- **Visual obstructions:** Most of the visual obstruction factors relate to blind spots or sight distance issues in the roadway. According to the NDS data dictionary, they had contributed to the occurrence and severity of the safety-critical event (i.e., in this case, crash) or influenced the ability of the subject driver to effectively recognize and respond to the possible safety hazard and precipitating event. Some of the visual obstructions reported in the NDS crashes include curves or hills, inadequate roadway lighting system, parked vehicles, and other obstructions.
- **Surface Conditions:** In some crashes, surface conditions such as snowy, muddy, and oily, were reported to have affected the vehicle’s coefficient of friction at the start of the precipitating event which resulted in crashes.

Vehicle Factors

This study considers the “vehicle contributing factors” available in the NDS data including faulty tires and braking systems as the vehicle-related crash contributing factors. Based on NDS data reductionists, such factors might have contributed to the precipitating event or subject driver’s response which resulted in crashes. Some of the factors in the “Visual Obstructions” include inadequate defrost or defog system, inadequate vehicle headlamps, and faulty headlights, which can be more appropriately categorized as vehicle factors. Hence this study classified such factors as vehicle factors, which might have contributed to the occurrence or severity of the safety-critical events (i.e., a crash).

A Systematic Taxonomy for Driving Errors and Violations

We classify the driving errors and violations using a Spatio-temporal framework consisting of perception, recognition, decision, and reaction (PRDR) (Figure 2.1). During the perception phase, a driver perceives the presence of a potential safety hazard in the roadway through sight. Recognition is the next phase, where a driver realizes the actual identity of the safety hazard. Next comes the decision phase in which a driver decides the course of action to overcome the potential conflict. Finally, a driver initializes the course of action (i.e., start moving the foot to the brake pedal or hand to the steering). In the NDS data, driver behavior (i.e., improper braking, the driver did not see), reported by the reductionists to have contributed to crash, could be classified as perception as well as recognition errors (Figure 2.2). However, after evaluating the narrative, it was more appropriately related to recognition failure as the subject driver was able to view the surrounding, specifically the rear end view, but could not recognize the surrounding (Figure 2.2). According to this study, decision and recognition failures can occur during the decision and recognition phases of the driving. Moreover, a driver, during the reaction phase, can be involved in performance errors or violations (intentional/un-intentional) while executing his/her intended course of action. This study considers physical conditions and experience/exposure-related errors as separate driving errors which can influence driving capabilities during all phases of the PRDR process (Figure 2.1). While we developed and used the TDEV of driving errors and violations for crashes in this study, it can also be used to systematically classify driving behavior in near-crashes and baselines (Figure 2.2). Brief definitions of errors and violations are provided as below.

Recognition Errors

Recognition errors occur when a driver fails to appropriately recognize the real situation, either due to distraction or poor judgment, requiring an evasive maneuver. This variable is coded by data reductionists (by reviewing video data) when proper attention is not maintained immediately before a SCE. Specifically,

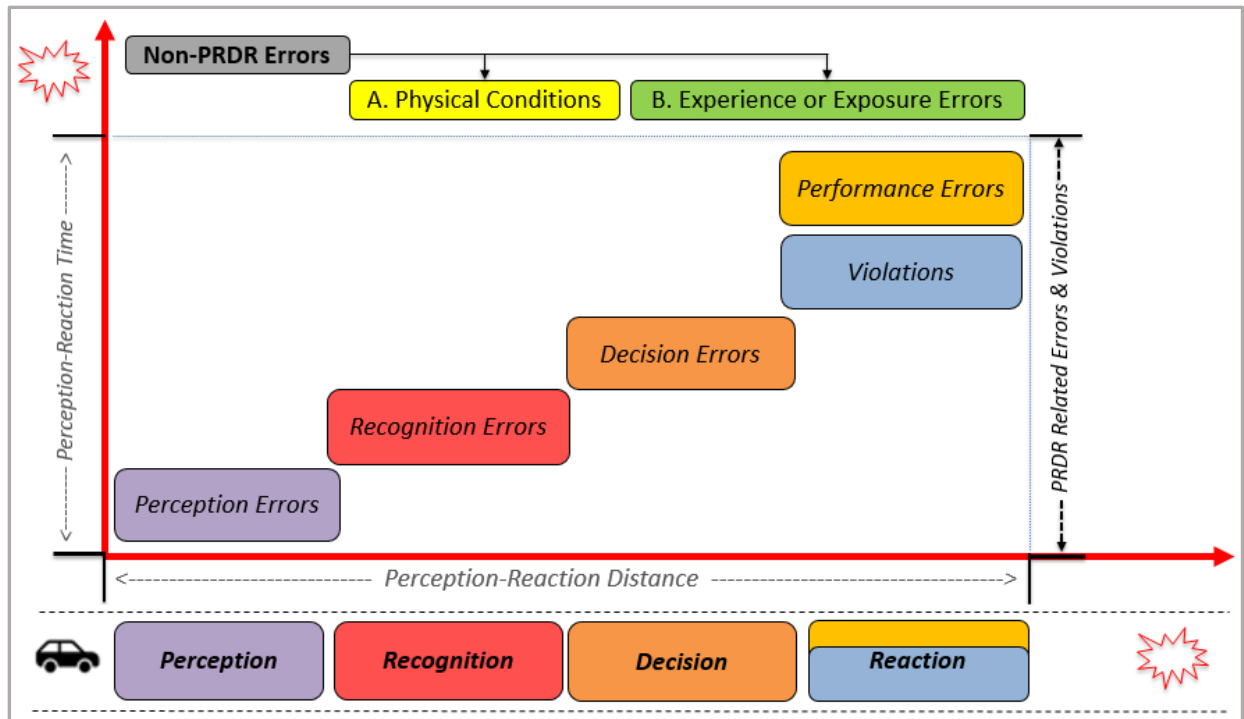


Figure 2. 1 Evolution of Driving Errors and Violations in Space-Time Dimension

Notes: A and B indicate types of errors that can contribute to crash occurrence but may not explicitly belong to any of the perception, recognition, decision, and reaction process categories.

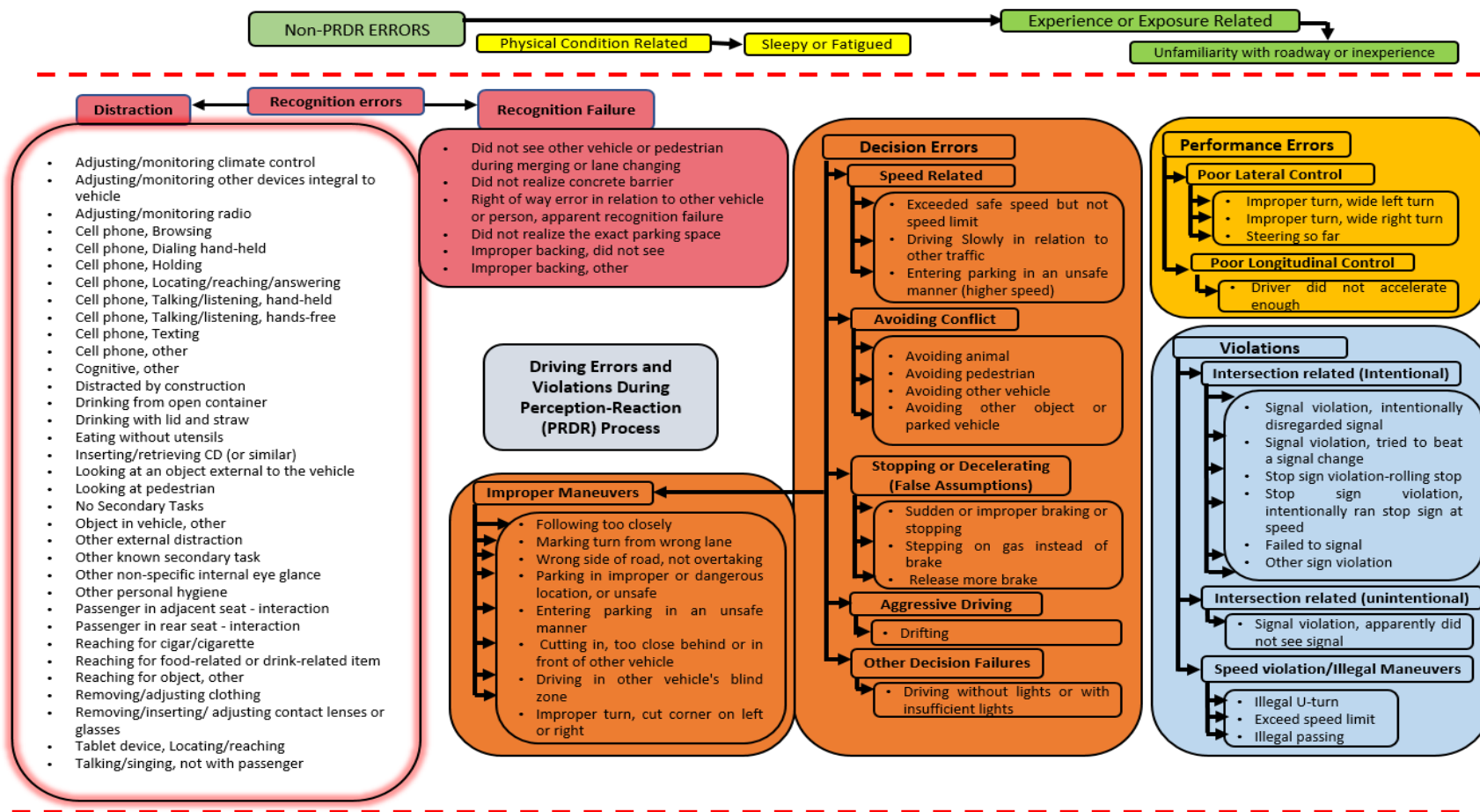


Figure 2. 2 Systematic Taxonomy for Driver Errors and Violations

Notes: PRDR is perception, recognition, decision, and reaction process; the length of each phase of the PRDR may vary in real-time depending on the psychological and physical conditions of the driver, and the complexity of the situations. The length of the decision phase is shown to be longer as more sub-types of decision errors were observed in the NDS crashes than any other driving errors and violations.

according to the NDS data dictionary, in such cases, the subject driver was reported as unable to maintain an appropriate level of attention due to his/her involvement in one or more secondary tasks which occurred within seconds before the start of the precipitating event. The NDS data dictionary defines such situations as “the state of environment or action that began the event sequence under analysis”. The recognition error variable can be indicative of the higher risk associated with distraction compared with if the subject driver was not distracted. If the driver was not distracted, then they could have effectively responded to risky situations. Some examples in the “recognition failure” category, contributing to crashes, include drivers that were unable to recognize the right-of-way to other vehicle or person, drivers not realizing the exact dimensions of a parking space, drivers poorly judging the direction and speed of pedestrians or other vehicles when changing lanes, and inability to realize/see rear-side environment during backing.

Decision Errors

Decision errors happen when a driver is unable to decide or choose the appropriate response or action (e.g., brake or accelerate), having perceived, interpreted, and recognized the information correctly in the perception and recognition phase. For instance, some of the driver behaviors, such as driving slowly below the speed limit and exceeding safe speed but not the speed limit, indicate inappropriate decisions of the subject drivers which the NDS data reductionists considered to have contributed to the crash occurrence or severity, which we classify as decision failures. It indicates that the appropriate decision of the subject driver could be to maintain their speed (below the speed limit) according to the traffic stream and should not drive too slowly. Similarly, subject drivers could have avoided crashes if they would have not exceeded the safe speed limit. For details on relevant driver behaviors related to decision errors, please refer to Figure 2.2. Moreover, this study classifies decision errors into various subclasses which include: decision errors associated with improper maneuvers, speed-related decision errors, decision errors during avoiding conflict, decision errors during decelerating or stopping aggressive driving, and other decision errors (see Figure 2.2).

Performance Errors

Driver performance errors relate to a driver’s response in the reaction phase often reflecting poor lateral or longitudinal control or weak judgment of driving situations. For instance, if a driver intended to accelerate but did not accelerate enough, which resulted in a crash, then this would indicate a performance failure for the subject driver. Likewise, some of the drivers were unable to negotiate turns properly (i.e., taking wide turns or steering too far).

Physical Condition Related Errors

Physical condition errors are motivated by a driver's physical or psychological state, e.g., drowsiness potentially leading to unsafe situations. These physical condition-related errors, due to drowsy, sleepy, asleep, and fatigued, can have negative effects on the driving capabilities of the subject driver during any of the four phases of the PRDR process (Petridou and Moustaki 2000). To avoid crashes occurring due to such physical conditions, subject drivers need proper rest before driving as such conditions reduce a driver's psychological, visual, and audial response level and can also result in short-time negative impacts on their physical performance.

Experience or Exposure Errors

Experience or exposure errors can occur when a driver is either unfamiliar with the roadway and surroundings or lacks driving experience. This variable is determined and reported by the NDS data reductionists. Further detail about the variable can be found in the "Driver Behavior" factor provided in the NDS data dictionary.

Violations

In addition to performance errors, a driver may violate traffic regulations during the reaction phase, thus this study classifies the traffic violations separately from driving errors. Traffic violations, whether intentional or unintentional, are a severe threat to road users. Overall, this study classifies violations into three sub-classes including roadway intersection related (i.e., intentional and/or un-intentional) and roadway segment related (i.e., speed-related/illegal maneuvers). The intersection-related violations are then subcategorized into intentional and unintentional violations. Some drivers were unable to realize the presence of traffic control devices (i.e., signal or yield sign), hence these are classified as unintentional violations. Narratives for all such cases were checked which revealed that in some cases drivers did not realize or see the presence of traffic control devices. Intentional driving violations reported on roadways other than interstates included drivers not complying with the traffic control devices, given that they observed them in time. Some drivers were reported to exceed the posted speed limit on roadway segments, perform illegal maneuvers (illegal passing or illegal U-turns), or perform other illegal/unsafe actions such as violating the work-zone speed limit. All such violations are classified under a separate sub-class noted as "speed-related violations or illegal maneuvers".

The classification in Figure 2.2 is based on the "Driver Behavior" factors in the NDS data explained in the earlier section. There were a few cases where no contributing factor was mentioned, i.e., none of the driver

behavior, secondary task outcome, infrastructure, roadway surface condition, or vehicle obstruction were identified in the NDS data (i.e., marked with an asterisk (*)). In such cases, the research team evaluated the narrative (detailed description) of such cases to determine if any other human, vehicle, and roadway, or environment factors could be identified as contributors. After evaluating the detailed descriptions in the narratives for the few cases, the research team found driver-related contributing factors and classified them per the proposed TDEV (Figure 2.2).

Diverse Roadway Localities: NDS Data

In the NDS data available to the research team, roadway and land-use environments are classified under the variable “locality” into various categories (Table 2.1). The instructions for the data reductionists from the SHRP2 researcher data dictionary for this variable are as follows: “The best description of the surroundings that influence or may influence the flow of traffic at the time of the start of the precipitating event. If there are ANY commercial buildings, indicate as business/industrial or urban area as appropriate (these categories take precedence over others except for church, school, and playground). Indicate school, church, or playground if the driver passes one of these areas (or is imminently approaching one) at the same time as the beginning of the Precipitating Event (these categories take precedence over any other categories except urban, and divided highway).” The final variable includes the following categories: interstate/bypass/divided highway with no traffic signals, bypass/divided highways with traffic signals (principal arterials), open residential (which include few houses, signifying largely undeveloped land use), moderate residential, school, church, playground, business/industrial, urban, and open country (rural) settings (Table 2.1). Importantly, the open country and open residential locations are defined to include only vegetation and one/few housing units respectively. According to the NDS dictionary, open country and open residential areas are regarded as rural and rural/semi-rural locations respectively.

Table 2.1 is presented in original format (source: NDS data dictionary) to understand the coding of various roadway localities. As defined in the table, Open Country indicates rural locations, while Open Residential (including one or few housing units) locations indicate rural or semi-rural locations can be combined and considered as one category. The percentage of all events (i.e., including baselines, near-crashes, and crashes) reported on roadways within the church, playground, and other locations like campground were very low; hence these three categories are collectively considered to have useful insights (also a playground and campground locations somehow include similar activities which can influence driving behaviors in a somehow similar way).

Table 2. 1 Definition of Various Roadway Localities in the NDS Data (Source: NDS Data Dictionary)

Value	Definition
Open country	Other than the roadway, there is nothing but vegetation visible during the time surrounding the Precipitating Event that is described in any of the other categories. Road is not an Interstate or a bypass/divided highway with traffic signals. (Often appears as rural roads, 2 lanes undivided.) Example and Hints: Includes roadways not defined as Interstate or divided highway, when no landmarks mentioned in other categories are visible.
Open Residential	Rural to semi-rural areas where there may be only one or a few houses around (i.e., farmland).
Moderate Residential	An area where multiple houses or apartment buildings are present, but is not as dense as an Urban Locality. Example and Hints: e.g., residential subdivisions
Business/industrial	Any type of business or industrial structure is present, but is not as dense as an Urban Locality. (If there are also houses visible, this category takes precedence over Open residential and Moderate residential).
Church	One or more involved vehicle passes a church building at the time of the Precipitating Event.
Playground	One or more involved vehicle passes any type of playground or children's playing field at the time of the Precipitating Event. Example and Hints: If playground/field is on school grounds, code as School.
School	One or more involved vehicles passes any type of school building or is in a school zone at the time of the Precipitating Event, including adult learning institutions. Example and Hints: Include any training centers, universities, etc. as well as elementary and secondary schools.
Urban	Higher density area where blocks are shorter, streets are a mix of one and two-way, and traffic can include buses and trams. (This category takes precedence over others when either businesses and/or residences are present.)
Interstate/bypass/divided highway (no traffic signals)	Vehicle is travelling on an interstate, bypass, or divided highway with no traffic signals (regardless of what buildings can be seen), at the time of the Precipitating Event.
Bypass/divided highway with traffic signals	Vehicle is travelling on a bypass or divided highway with traffic signals (no other category description is visible) at the time of the Precipitating Event. (Often appears as "Open Country", but with more lanes and/or as a divided road.)
Other	Locality at the time of the Precipitating Event is one not described in other categories. Example and Hints: Ex. In campground.
Unknown	Cannot determine the Locality due to limitations in video views, lighting, visual obstructions, or limited perspective. Example and Hints: Ex. Part of the video is missing or there is insufficient information in the video to make a determination.

Results and Discussion

First, it is important to note that that most of the crashes (N = 673) in the NDS data available to the research team were not severe. Of all the crashes, 8.92% (N = 60), 13.22% (N = 89), 37.59% (N = 253) and 40.27% (N = 271) were severe, police-reportable, minor, and low-risk tire strike crashes, respectively¹. On one hand, with a prevalence of minor and low-risk/tire strike crashes (78%), a wide range of safety-critical events are captured, most of which would not be available through police-reported crashes. On the other hand, only 22.14% of the overall crashes are police reportable, which means that the NDS data are not fully comparable with the widely available police-reported data. Simple cross-tabulations (not reported) showed that recognition errors were more frequently associated with severe crashes (51%) compared with other errors and violations.

Safety Matrix

Using the methodology described earlier, crashes were categorized as having driver, vehicle, and roadway factors that contributed to their occurrence. Furthermore, the PRDR framework was used to characterize human factors (i.e., including driver behavior and few secondary tasks which contributed to crash occurrence when no driver behavior was reported). The safety matrix for the NDS crashes reveals that the

¹ The four severity levels are defined as follows in the NDS dataset. **Severe Crashes:** “Any crash that includes an airbag deployment; any injury of driver, pedal cyclist, or pedestrian; a vehicle roll over; a high Delta V; or that requires vehicle towing. Injury if present should be sufficient to require a doctor's visit, including those self-reported and those apparent from video. A high Delta V is defined as a change in speed of the subject vehicle in any direction during impact greater than 20mph (excluding curb strikes) or acceleration on any axis greater than +/-2g (excluding curb strikes).” **Police-reportable Crashes:** “Police-Reportable Crash. A police-reportable crash that does not meet the requirements for a Level I crash. Includes sufficient property damage that it is police reportable (minimum of \$1500 worth of damage, as estimated from video). Also includes crashes that reach an acceleration on any axis greater than +/-1.3g (excluding curb strikes). If there is a police report this will be noted. Most large animal strikes and sign strikes are included here.” **Minor Crashes:** “Physical Contact with another Object. Most other crashes not included above are Level III crashes, defined as including physical contact with another object but with minimal damage. Includes most road departures (unless criteria for a more severe crash are met), small animal strikes, all curb and tires strikes potentially in conflict with oncoming traffic, and other curb strikes with an increased risk element (e.g., would have resulted in worse had curb not been there, usually related to some kind of driver behavior or state).” **Low-risk Tire Strikes:** “Tire Strike, Low Risk. Tire strike only with little/no risk element (e.g., clipping a curb during a tight turn).”

human factors were the sole contributing factors in 77.56% of crashes (Table 2.2). The statistics indicate that human factors have a contribution to around 93% of crashes (Figure 2.3), this finding is consistent with previous police-report-based studies (Sabey and Staughton 1975, Treat *et al.* 1979, Singh 2015). In the aggregate, our findings regarding the overall contribution of driving errors and violations to crashes are somehow similar to the Tri-level study (in both cases human factors were found to contribute to about 93% of crashes) – despite being 40 years apart. However, at a disaggregate level, the nature of the driving errors (e.g., secondary tasks such as mobile phone use while driving causing distraction) have changed in recent decades. The availability of real-world data on pre-crash driving behavior and secondary tasks reported in safety-critical events (crashes and near-crashes) and baselines (exposure without an event) signifies and validates the need for this research. While previous taxonomies of driving errors and violations provide very useful insights, they were based on police crash reports or on-site human investigations where subjectivity and bias could not be ignored (Mannering and Bhat 2014, Wali *et al.* 2018d). Furthermore, on basis of the sole or definite contribution of human factors to crashes, our findings significantly differ from a Tri-level study (Treat *et al.* 1979). For instance, findings from this study suggest that human factors solely contributed to 77.56% of the crashes (Table 2.2) which is significantly greater than what was found in the Tri-Level Study – where human factors were found to be definite causes in 64.3% and 70.7% of the crashes by the on-site team and in-depth investigation team respectively (Treat *et al.* 1979). While one would expect significant reductions in the sole or definite contribution of human factors to crashes after 40 years due to improvements in vehicle technologies, road design, and enforcement, we noticed a substantial increase (~7% to 13%) in cases where the sole contribution to crashes was human factors. This could be due to the extensive involvement of drivers in secondary tasks and distracting activities like cell phone use etc. Note that in our case the contribution of vehicle factors to crashes is even smaller (~1%), perhaps due to the use of better vehicles in the selected (non-random) sample. Therefore, the proposed taxonomy of driving errors and violations not only contributes by bringing pre-crash driving behavior and secondary tasks (which can contribute to a crash) more systematically into the behavioral framework, but also helps in identifying appropriate driver assistance systems (Intelligent Transportation System “ITS” countermeasures) and other traditional countermeasures which could reduce these driving errors and violations.

Studies suggest that in the future if automated and connected vehicle technologies are deployed only in six countries (i.e., Australia, Canada, India, New Zealand, United Kingdom, and the United States) while disengaging human drivers, the average number of road crashes can be reduced by 3.40 million (Wang *et al.* 2020). Roadway factors were the second most prevailing factors contributing to around 17% of crashes (Figure 2.3). Vehicle factors were found least frequent with a percentage share of around 1%. This can be likely attributed to enhanced and modern vehicle technology in the NDS sample.

Table 2. 2 Safety Matrix: Contribution of Human, Vehicle, and Roadway Factors

Human Factors	Vehicle Factors	Roadway Factors	Crash	
			Freq.	%
Y	Y	Y	2	0.30
Y	Y	N	5	0.74
Y	N	N	522	77.56
N	N	N	30	4.45
N	N	Y	18	2.67
N	Y	Y	0	0.00
N	Y	N	0	0.00
Y	N	Y	96	14.26
Total			673	100.0

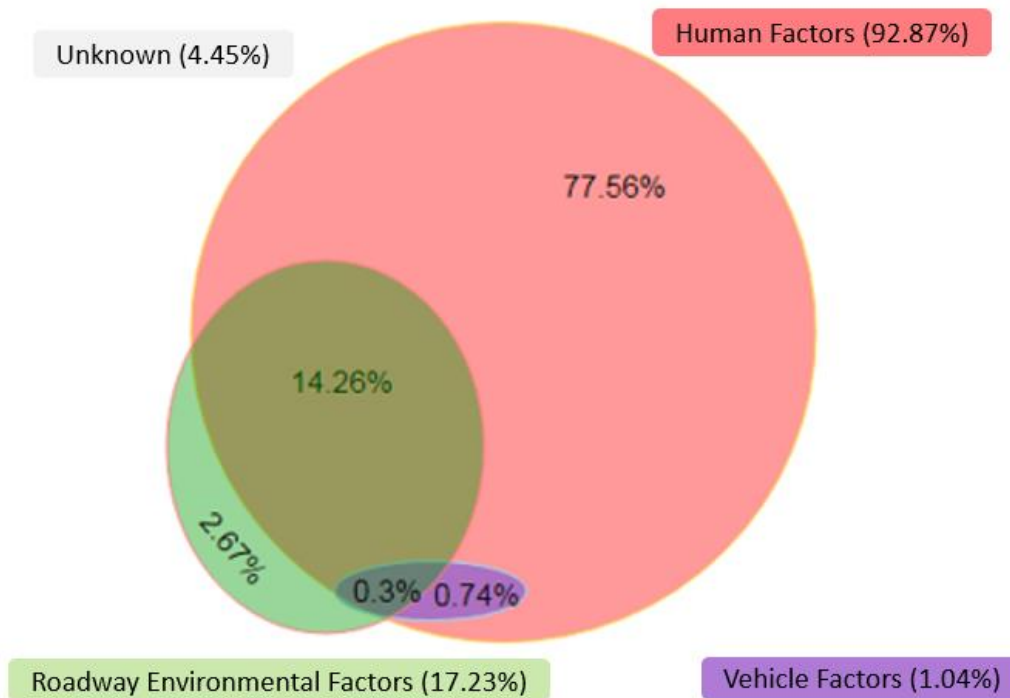


Figure 2. 3 Contribution of Human, Vehicle, and Roadway Factors in Crashes

Prevalence of Driving Errors in Safety-Critical Events: Evidence from NDS Data

Using the NDS data, Table 2.3 provides summary statistics for driving errors and violations for safety-critical events (crashes and near-crashes) and baselines (driving when no safety-critical event occurs). The NDS data statistics show that at least one driving error or violation was reported in 93% of crashes and 61% of near-crashes respectively (Table 2.3). Recognition errors were the predominant driving errors that contributed to crashes and near-crashes with a percentage share of 39% (Table 2.3). Decision errors were reported as the second leading driving error, causing 34% and 14% of crashes and near-crashes respectively (Table 2.3). Drivers were involved in traffic violations in 9% of crashes and 11% of near-crashes. The safety hazard associated with different types of driving errors and violations was quantified using the ratio of the percentage share for each key error type as crashes/near-crashes over percent share in baselines as presented in columns 5 and 6 (Table 2.3). The findings reveal substantial safety risks are associated with recognition errors, as the percent share of recognition errors in near-crashes and crashes was 155 and 176 times of their percent share in baselines (Table 2.3). Decision errors were reported in 2.69% (N = 204) of total baselines, and the percent share of decision errors in near-crashes and crashes were 5 and 13 times their percent share in baselines. If these decision or recognition errors and violations are constantly repeated by drivers, then they can result in a safety-critical event.

It is important to point out that the NDS data captures safety-critical events that did not necessarily result in police-reportable crashes. For example, according to the NDS data dictionary, the “impairments” factor is defined to influence the “driver behavior.” However, in only in about 1% of the crashes (N = 9 out of 673 crashes), the subject drivers were reported to have been impaired with alcohol or other drugs. These impairments were in turn reflected in driver error variables and classified as recognition errors (N = 4 crashes) and decision errors (N = 3 crashes).

Recognition errors contributed to 39% (N=260) crashes with distraction as the most predominant type of recognition errors, constituting around 84% of crashes where hazard recognition failures contributed. This reveals that drivers can be at a higher risk even with a small disengagement from driving tasks. The findings are consistent with previous studies which also report recognition errors as the most frequently contributing driver error, followed by decision and performance errors (Treat *et al.* 1979, Wierwille *et al.* 2002). These previous findings were based on police reports or on-scene investigations, which showed that the contribution of recognition errors to crashes was 9%-13% more than decision errors (Treat *et al.* 1979, Wierwille *et al.* 2002). In comparison, the contribution of recognition errors to crashes in this study is 5% more than that of decision errors (Table 2.3). According to statistics, decision errors resulted in 34% (N = 231) of crashes, which mainly include wrong decision making during sharp or improper turns summing up

Table 2. 3 Prevalence of Driving Errors in Safety-Critical Events

Variable	Event Outcome			Prevalence of Errors in SCEs	
	Baseline (%) (N =7,589)	Near-Crash (%) (N=1,331)	Crash (%) (N=673)	$\frac{\% \text{ in Near Crash}}{\% \text{ in Baseline}}$	$\frac{\% \text{ in Crash}}{\% \text{ in Baseline}}$
Driver's Errors and Violations					
• No error or violation	90.12	38.69	7.13	0.43	0.08
• Recognition error	0.22	34.03	38.63	154.68	175.59
• Decision error	2.69	13.82	34.32	5.14	12.76
• Performance error	0.09	0.68	7.58	7.56	84.22
• Physical condition error	1.25	1.43	1.34	1.14	1.07
• Experience/exposure error	0.07	0.6	1.93	8.57	27.57
• Violation	5.56	10.74	9.06	1.93	1.63
Total	100%	100%	100%	---	---

Notes: SCEs indicate Safety-critical Events (near-crash, or crash). Column 5 and Column 6 indicate the ratio of the percentage share of each error type in safety-critical events (i.e., near-crash or crash) to their respective percentage share in baselines (driving when no safety-critical event occurs).

to 64% (N = 148 crashes) of such crashes. Similarly, 16% (N = 36) and 5% (N = 11) of the overall crashes occurring due to decision errors occurred due to errors in making appropriate decisions to maintain a safe speed and sudden stopping respectively. A total of 48 (8%) crashes were associated with performance errors (Table 2.3). Out of the crashes (N = 48 i.e., 8% of the overall NDS crashes) occurring as a result of performance errors, 63% (N = 32) occurred due to driver's improper performance while making too-wide turns. In another 32% (N=16) of the performance errors that contributed to crashes, drivers failed to maintain appropriate acceleration/deceleration and steering control. Referring to violations that contributed to a total of 9% (N=61) of the overall crashes, 44% (N = 27) and 31% (N = 19) of these crashes occurred due to violations of speed limit regulations and intentional violations of traffic signals at intersections respectively. To control such critical and risky driving errors, research is needed on the deployment of ITS equipment, which can compensate or forgive these driving errors to a significant extent (Stanton and Salmon 2009). Meanwhile, there is a need to research appropriate training programs to mitigate performance and experience-related errors that contributed to roughly 10% of the overall crashes (Wali et al. 2019).

Distribution of Safety-Critical Events across Various Roadways and Environments

This section presents the distribution of crashes and near-crashes across the roadway and environmental variables such as business/industrial and residential locations. Notably, local contexts may be associated with specific driving errors and our analysis sheds light on hotspots of frequent errors and violations potentially resulting in safety-critical events (discussed in section 3.4 earlier). Analysis of frequencies showed that crashes and near-crashes occurred more frequently when a business or industrial structures were present (but not as dense as Urban Locality). For example, crashes and near-crashes occurred more frequently (i.e., 46% and 40% respectively) when business or industrial structures were present (Table 2.4). The next most frequent category was near moderate residential developments. To some extent, these frequencies may reflect the exposure of the NDS sample in residential and business or industrial areas. While direct exposure of the vehicles in the sample is not available to researchers, the baselines provide a coarse surrogate for exposure. Therefore, crash percentage divided by baseline percentage is calculated to get a sense of the rate (Table 2.4) – the best possible but admittedly coarse surrogate for exposure. Also, business or industrial areas may be characterized by the complexity of traffic, diversity of activities, and special roadway and environmental conditions that may lead to driving errors and subsequent crashes.

Table 2.4 provides the distribution of crashes, near-crashes, and baselines across different roadways and environments. The highest percentage of crashes occurred on roadways passing through business or industrial localities and moderate residential locations making up 46% and 20.5% of the overall crashes

respectively (Table 2.4). The percentage of baselines reported in these areas was also greater compared to other locations (Table 2.4). Although the percentage of baselines reported on interstates was the second-highest, the percentage of crashes on Interstates was relatively lower compared to other locations such as urban environments, school zones, and business or industrial.

Based on the percentage of crashes per percentage of baselines in a specific locality, interstates and open country/open residential (rural and semi-rural settings) may pose lower risks (Table 2.4). Since urban, business/industrial, and school locations seem to have experienced a higher percentage of crashes compared to the percentage of baselines reported in these locations, they can be considered high crash risk locations.

Variation of Driving Errors across Various Roadways and Environments

Analysis of frequencies showed that recognition errors contributed to 49%, 39%, and 39% of the overall crashes occurring within school zones, business/industrial, and moderate residential areas, respectively. Moreover, decision errors contributed to the highest percentage of crashes when a business or industrial structures were present. Notably, 6% (N = 41) crashes occurred in localities coded as interstates (Table 2.4). The distribution of various driving errors and violations on interstates, contributing to crashes, is shown in Figure 2.4. Recognition and decision errors contributed to 37% and 32% of the overall crashes when the locality was coded as interstates. The findings are in line with a recent study that indicates that the chance of making decision errors was higher on interstates (Shaon *et al.* 2018). Violations were reported in nearly 5%, while each performance and experience/exposure errors were reported to have contributed to 2% of the crashes. Referring to the distribution of driving errors and violations in crashes occurring on the bypass/divided highways with traffic signals, 35% of the crashes on this roadway occurred due to recognition errors, followed by decision and performance errors which contributed to 24% and 18% respectively. Performance errors were substantially higher on these bypass/divided highways with traffic signals. This finding is in agreement with recent studies (Shaon *et al.* 2018). This could be due to the complex interaction of drivers with other roadway entities specifically at signalized intersections.

The large percentage of crashes occurred in business/industrial areas (46%; N = 311) and moderate residential areas (20.5%; N = 138) (Table 2.4). Figure 2.5 illustrates how drivers were coded in the NDS data to make certain errors and violations in each of the moderate residential areas and business/industrial locations. In areas coded as business/industrial locations, recognition and decision errors each contributed to nearly 39% of the total crashes reported on roadways with such land uses (Figure 2.5). Violations were coded to contribute to only about 6% of the overall crashes occurring on roadways within business/industrial locations. Referring to moderate residential locations, percentage of crashes occurring

Table 2. 4 Distribution of Crashes, Near-Crashes, and Baselines across Different Roadways and Environments

VARIABLE	All Cases (N = 9,593) Percent (S.D)	Baseline (%) (N = 7,589)	Near-Crash (%) (N = 1,331)	Crashes (%) (N = 673)	% of Crashes % of Baselines
ROADWAY LOCALITY					
• Interstate	23.29 (0.4226)	25.6	18.78	6.09	0.2379
• Open Country or Open Residential	8.66 (0.2813)	9.46	5.48	5.94	0.6279
• Moderate Residential	21.21 (0.4088)	22.26	15.55	20.50	0.9209
• School	5.10 (0.2199)	4.51	7.06	7.87	1.7450
• Business/Industrial	34.08 (0.4739)	31.10	40.04	46.21	1.4859
• Urban	2.76 (0.1639)	1.51	7.30	7.88	5.2185
• Bypass or Divided Highways with no Traffic Signals	3.00 (0.1760)	3.04	3.00	2.53	0.8322
• Others (e.g., Church, Playground and Campground)	1.90 (0.1364)	1.66	2.78	2.98	1.7952
Total	100%	100%	100%	100%	

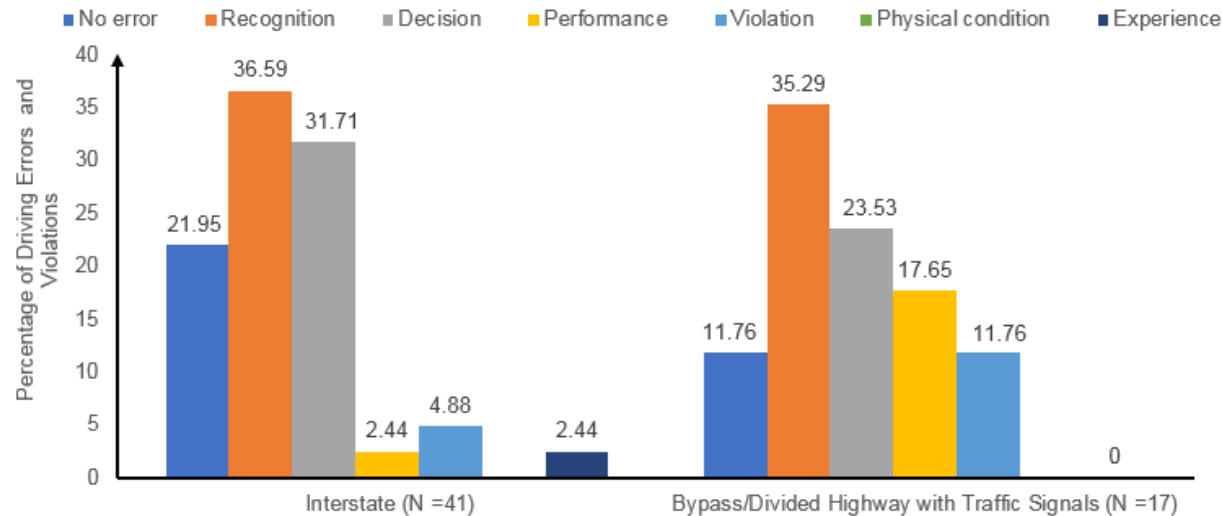


Figure 2. 4 Driving Errors and Violations on Interstates and Bypass/Divided Highways with Traffic Signals (in Crashes)

due to recognition errors (39%) was significantly higher than that of decision errors (26%). A significant percentage of crashes in moderate residential areas (11%) also occurred due to violations.

The analysis of errors further reveals that 8% (N = 53) of crashes were reported on roadways passing within school zones (Table 2.4). Interestingly, nearly 50% of the crashes were reported to have occurred within school zones due to recognition failures (Figure 2.6). Similarly, decision errors and violations contributed to 32% and 9% of the total crashes occurring on roadways passing through school zones, respectively (Figure 2.6). Only about 8% (N = 53) and 6% (N = 40) of the overall crashes (N=673) were coded by data reductionists to occur in urban and rural/semi-rural (open country/open residential) locations. Of the crashes that were coded to have urban areas as a key feature, 96% occurred due to driving errors. Among these, recognition and decision errors were the most prevailing errors contributing to nearly 40% and 34% of the total crashes in urban areas. In rural/semi-rural setups, the decision errors had a large contribution resulting in 35% of the crashes, followed by recognition failures which contributed to 27.5% of crashes in these areas. Violations contributed to around 25% of the crashes in rural/semi-rural locations, perhaps partly due to a comparatively lower level of traffic surveillance than business/industrial and urban areas.

This study explored the distribution of various driving errors and violations using the information on various roadways and environments available in the “locality” factor in the NDS data. In the future, a more holistic approach can be followed where relevant variables are systematically classified to facilitate exploring the correlations of different roadways, activity centers, and land use on driving errors and violations (Figure 2.7). The roadways and environments (coded as a locality in the NDS data by data reductionists) can be divided into roadway functional classification, residential zones, business/commercial zones, industrial zones, and urban, suburban, and rural areas. Perhaps a new activity-based land-use classification system can be used in the NDS context, along with the existing Federal Highway Administration (FHWA) roadway functional classification (FHWA 1989), and locality-based land-use classification (Sorensen 2000, Maret and Dakan 2003). To explain this further, using the locality variable, the categories are marked with (✓) or (✗) indicating whether NDS data does or does not include the category (Figure 2.7). For example, a complete FHWA roadway functional classification includes interstates, arterials (major/minor), collectors (major or minor), and local streets (major/minor) (FHWA 1989). However, the NDS data related to the locality provides information only whether a safety-critical event occurred on interstates or principal arterials (bypass/divided highway with traffic signals). While we acknowledge that residential, industrial, and business localities can be sub-classes of built-up areas (LaGro Jr 2005), the NDS data reductionists considered urban, residential, school, church, playground, business/industrial localities as separate locations (Figure 2.7). In short, the NDS data available to us through the locality variable does not include complete information on all categories which could completely fulfill criteria of either the existing FHWA

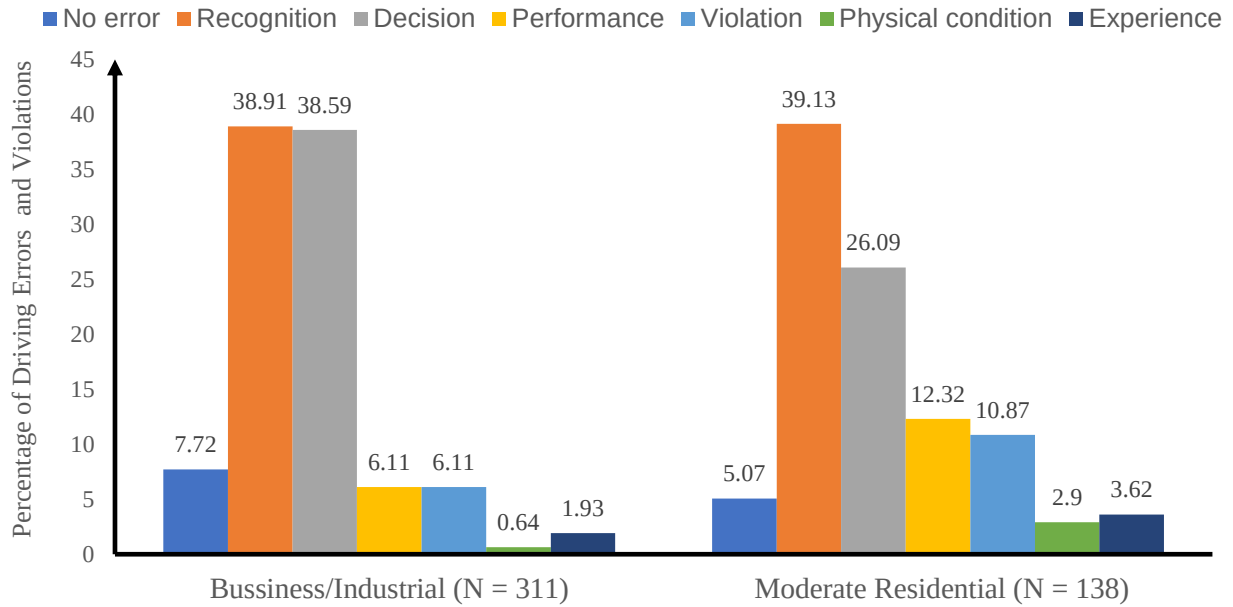


Figure 2. 5 Driving Errors and Violations in Business/Industrial and Moderate Residential Areas (in Crashes)

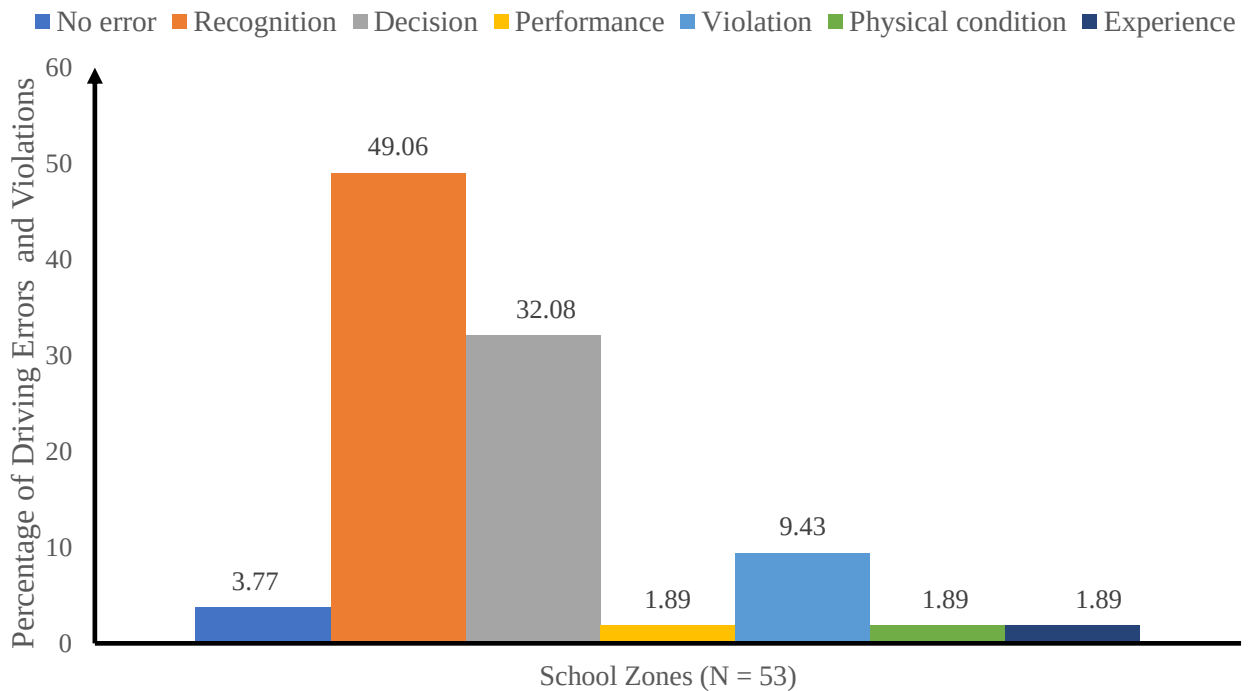


Figure 2. 6 Driving Errors and Violations in School Zones (in Crashes)

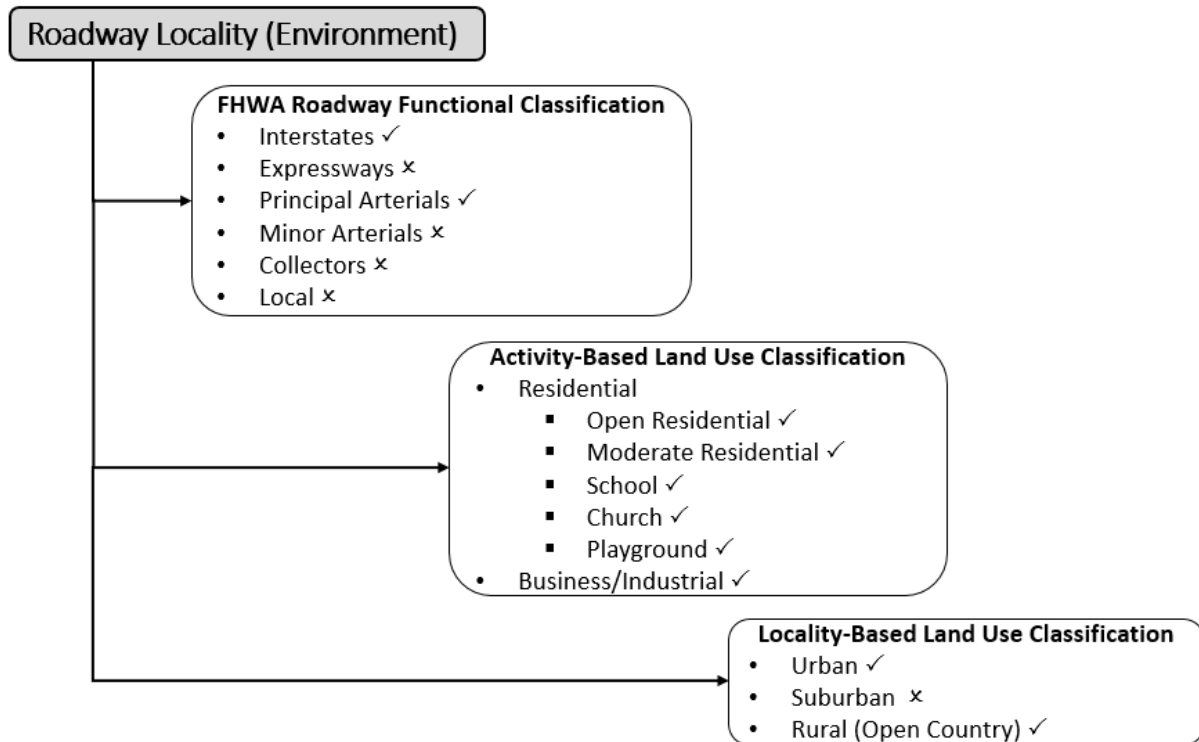


Figure 2. 7 Systematic Classification for Roadway Localities

Note: All three roadway locality classifications use information from a single variable “locality” in the NDS data-hence complete information on all categories within any of the three classifications is not available. The categories marked with ✓ and ✗ indicate factors on which information was available and unavailable, respectively. All known categories within the three roadway classifications add up to 100% (N = 673 crashes, N = 1,331 near-crashes, or N = 9,593 events including baselines, near-crashes, and crashes: whichever is the scope of the future studies).

roadway functional classification or locality-based land-use classification or activity-based land-use classification (Figure 2.7).

Countermeasures to Reduce Driving Errors and Violations

After closely examining crash causation using NDS data, we found that human factors were the most prevalent contributors to 92.87% of crashes (Figure 2.3). Once we developed and applied our systematic taxonomy of driving errors and violations, we found that recognition errors, decision errors, and performance errors have the highest contribution resulting in 38.63%, 34.32%, and 7.58% of the total crashes respectively (Table 2.3). Our findings suggest that a significant percentage (9.06%) of crashes also occurred due to violations which include both intentional and unintentional violations (Table 2.3). Referring to recognition errors, these errors mainly include cases where drivers were either distracted or unable to appropriately recognize other roadway entities (e.g., pedestrian and opposing vehicle) or situations (e.g., during merging or diverging with through traffic, and parking vehicle). Hence, this significant percentage of crashes (38.63%) occurring due to recognition errors can be prevented by improving driver attention and correcting driving habits especially driver involvement in secondary tasks (e.g., cell phone browsing, and talking via cell phone.). Countermeasures that could modify driving behavior include enhancing driver recognition of crash risk (increasing driving recognition about forward safety hazards), improving attention, and enhancing driver awareness about surroundings. From connected and/or automated vehicles (i.e., technology) standpoint, it would be good to equip vehicles with some of the key technological countermeasures (ITS countermeasures) which are briefly discussed in the subsequent section. Note that large-scale implementation of multiple technologies (ITS countermeasures) has yet to be achieved. Therefore, these countermeasures should be considered suggestions (based on predictions) rather than firm recommendations for deployment and implementation (Farmer 2008).

Given police-reported crashes, previous studies have estimated potential reductions in severities and injuries due to technology-based countermeasures (e.g., forward-collision warning system, forward-collision warning system with auto brake, lane-departure system, and blind-spot (side-view assistance) detection system) (Harkey 2020). The aforementioned study also estimated the effects of several technology-based countermeasures on costs associated with personal injury protection (PIP), medical payments to other (MedPay), bodily injury liability, and property damage only liability (Harkey 2020). For instance, large-scale implementation of FCWS may cause the highest reduction (~20%) in MedPay costs, followed by bodily injury liability and PIP costs (Harkey 2020). For more details about the effects of the aforementioned countermeasures on insurance-related costs, please refer to (Harkey 2020). The percentage reductions in overall crashes, as well as injury crashes due to particular technology-based countermeasures,

are discussed in subsequent sections (Harkey 2020). However, note that while the early benefits of these countermeasures are promising, there is a need for further research which relates to in-service evaluations of these countermeasures. Note that more safety improvement can be achieved if a vehicle is equipped with multiple technologies as given below.

- **Forward Collision Warning System:** The forward collision warning system (FCWS) uses radar systems that determine the relative speed and distance between the subject vehicle and the leading vehicle or an object in the roadway. The FCWS provides an audio or visual warning to the subject driver (i.e., driver of the vehicle in which this system is installed) if the subject vehicle is approaching the leading vehicle with a higher speed. This system can be even more beneficial to inattentive or distracted drivers (Ervin 2005, Rakha 2011). Note that this technology typically helps in reducing recognition, decision, and performance errors (Rakha 2011). For instance, in the recognition errors category, this intelligent system can reduce errors related to distraction. In terms of decision errors, the forward-collision warning system can help in avoiding conflicts with animals, pedestrians, and other vehicles that are in front of the subject vehicle (Figure 2.2). Similarly, the technology can help in the appropriate stopping and decelerating of subject vehicles which can assist the subject driver in the decision phase to properly decelerate and/or stop (Figure 2.2). This intelligent system may also forgive the driver's decision of following too closely. This intelligent system may also help the subject driver to have appropriate control of their vehicles specifically in the longitudinal (forward) direction (Figure 2). Studies reveal that the system can also reduce physical condition errors (i.e., drowsy driving) by generating audio-visual warnings to alert subject drivers (Gaspar *et al.* 2019). While enhancing subject driver awareness, the FCWS can reduce the likelihood of fatal crashes (McKeever 1998, Regan *et al.* 2002, Farmer 2008). For instance, studies suggest that with FCWS, 30% of fatal crashes could have instead been non-fatal (Regan *et al.* 2002). One recent study suggests that the large-scale deployment of FCWS could help significantly reduce overall crashes and injury crashes by 20% and more than 26%, respectively (Harkey 2020).
- **Adaptive Cruise Control (ACC) System:** The system controls both speed and headway between the subject and leading vehicles. While the primary benefit of this countermeasure is to reduce recognition errors, it may also help in decreasing decision and performance errors (Ervin 2005, Rakha 2011). For instance, this smart system may maintain a safe distance from the leading vehicles which could be significantly beneficial if the subject driver is distracted (Figure 2.2). In terms of decision errors, adaptive cruise control can help the subject drivers to

reduce their speed-related decision errors (Figure 2.2). For instance, ACC may help the subject driver maintain safe speed in relation to the surrounding traffic and conform to the speed limit (Figure 2.2). Studies reveal that adaptive cruise control systems could help drivers in the recognition phase particularly when subject drivers are distracted or involved in secondary tasks (Kummetha *et al.* 2020). Simulation-based research suggests that adaptive cruise control has the potential to significantly enhance the subject vehicle's longitudinal control (i.e., reduce performance errors) especially when the leading vehicle harshly decelerates (McDonald *et al.* 2001, Dang *et al.* 2015). While assisting subject drivers, implementing the aforementioned system could significantly reduce crash risk (especially rear-end crashes) (Jing-liang and Zhao-du 2009).

- **Lane Tracking System:** This system is designed to measure the distance between the car center and lane markings, determine the curvature of roadway, measure offset between the centerline of the subject vehicle and roadway (lane), and measure marking characteristics which collectively help in determining whether the subject vehicle is within the lane or not. If the driver/vehicle is not in compliance, this system generates an audio/visual warning to the subject driver. This countermeasure can be used to correct and reduce recognition, decision, and performance errors (Ervin 2005, Hanowski *et al.* 2008, Lerner *et al.* 2010, Rakha 2011). In some cases, the lane tracking system may reduce performance errors related to poor-lateral control such as improper turns and wide left or right turns (Figure 2.2). Studies also revealed the effectiveness of lane tracking systems in reducing physical condition errors (i.e., generating warnings for drowsy drivers) (Rimini-Doering *et al.* 2005). While assisting drivers in various driving tasks, this system can help reduce crash risks and associated injuries. One recent study revealed that large-scale implementation of lane tracking or lane departure warning systems may significantly reduce overall crashes and injury crashes by more than 10% and 20%, respectively (Harkey 2020).
- **Blind Spot (Side-View) Assistance System:** This system includes ultrasonic sensors on either side of the vehicle which continuously monitor adjacent lane(s), thus helping the driver view blind spots. Studies suggest that 50,000 crashes could be prevented per year in the U.S. if the "Blind Spot Assistance System" is implemented in all vehicles in the country (Cicchino 2018). Studies reveal that the blind spot assistance system has the potential to prevent 26% (Farmer 2008) and 24% (Jermakian 2011) of lane-change collisions. One recent study suggests that large-scale implementation of a blind-spot assistance system has the potential to assist in driving tasks which in turn could reduce crashes (including all severities) and injury crashes by more than 10% and 20%, respectively (Harkey 2020). The blind spot assistance system has the

- potential to significantly reduce recognition (Liu *et al.* 2017a, Sossa and Zamora 2020) and decision errors (Chovan *et al.* 1994), which we found to be the most prevalent driving errors.
- **Lateral Vehicle Detection System:** Studies reveal that 4-10% of all crashes in the U.S. occur due to intentional lane changes (Lee *et al.* 2004). Lateral vehicle detection systems could potentially reduce crashes associated with lane departures (Tideman *et al.* 2007). This system includes a lane change assistance system, a lateral camera, and a collision warning system which helps the subject driver to perceive and observe the movement of other vehicles in adjacent lanes (Rakha 2011, Blanco *et al.* 2016). This countermeasure can also assist the subject driver to make safe and appropriate maneuvers (Rakha 2011, Blanco *et al.* 2016), indicating the potential to reduce performance errors.
 - **Dilemma Zone Mitigation System:** The installation of this countermeasure in a subject vehicle can reduce crashes occurring due to intentional and unintentional violations at signalized intersections (U.S. DOT 2009, Rakha 2011). For instance, our analysis shows that in several NDS crashes, subject drivers either disregarded signals (intentionally/unintentionally), tried to beat the signals, or did not see the signal (for details, please refer to Figure 2.2). In such cases, this system would generate warnings to subject drivers thus reducing their intentional and unintentional violations. For instance, the Texas Department of Transportation (TxDOT) evaluated and quantified both the operational and safety benefits of the Dilemma Zone Mitigation System (U.S. DOT 2009). Besides other operational benefits, they suggested that overall red-light violations, red-light violations by heavy vehicles, and the frequency of severe crashes reduce by 58%, 80%, and 39%, respectively (U.S. DOT 2009).
 - **Low-speed Automatic Brake System (LSABS):** Studies reveal that a significant percentage of road crashes involve rear-end collisions, which cause a substantial amount of the economic burden on society (Watanabe and Ito 2007, Isaksson-Hellman and Lindman 2012). Most rear-end crashes involve passenger cars and occur at low speeds (Isaksson-Hellman and Lindman 2012). The Low-speed Automatic Brake System uses laser technology, which determines the deceleration rate required to avoid collisions with the leading vehicle after accessing its distance, speed, and acceleration relative to the subject vehicle (Distner *et al.* 2009, Isaksson-Hellman and Lindman 2012). Studies reveal that if 100% market penetration of LSABS is achieved, road traffic fatalities and injuries could be reduced by 13.2% and 9.1%, respectively (Tan *et al.* 2020). One U.S.-based study suggests even larger reductions in overall crashes and injury crashes (Harkey 2020). For instance, both overall crashes (including all types of severities) and injury crashes could be reduced by more than 40% if 100% market penetration of LSABS is achieved (Harkey 2020). This system may assist drivers under the influence,

careless drivers, and inexperienced drivers (Bao and Huang 2015). Studies suggest that LSABS consisting of cameras, radars, and sensors recognize the surrounding environment which may assist drivers in the recognition phase and thus reduce recognition errors (Cicchino 2019, Yang *et al.* 2019). LSABS also has the potential to reduce physical condition errors (drowsy driving) (Bao and Huang 2015) and experience-related errors (Bao and Huang 2015). This intelligent system can also help mitigate a significant percentage of near-crashes. One of the studies suggests that increasing penetration of passenger vehicles equipped with automatic emergency braking (AEB) system is expected to significantly reduce front-to-rear conflicts (Boggs *et al.* 2020). In this regard, twenty automotive industries (collectively exceeding 99% of the US automobile market) have recently asserted to equip all new passenger vehicles and trucks with AEB systems by September 1, 2022, and September 1, 2025, respectively (NHTSA 2018, Boggs *et al.* 2020).

- **Drowsiness Detection System:** Drowsy driving contributes to a significant percentage of both fatal and non-fatal crashes. Studies reveal that drowsy or fatigue driving contributes to around 6% and 21% of the overall crashes and fatal crashes, respectively (Klauer *et al.* 2006, Tefft 2014). Studies suggest implementing intelligent detection systems to assess the subject driver's drowsiness level (Lee *et al.* 2008, Schwarz *et al.* 2019). Such technology-based countermeasures could help generate timely warnings of drowsy driving, which may, in turn, enhance safety. The subject driver's physical conditions (i.e., drowsy driving) can be detected via measuring either physiological signals (i.e., heart rate, brain waves, and/or eyes blinking) or physical changes (i.e., eyes state, mouth state, and/or head orientation), but measuring the former one can be potentially distracting and unrealistic as it requires a sensing electrode physically attached to the subject driver's body (Alioua *et al.* 2011). The nonintrusive detection, which uses a video camera to evaluate physical features like head motion (Smith *et al.* 2000), eyes condition (Juvale *et al.* 2009), and state of mouth (Mohanty *et al.* 2009), is a more suitable method. This system generates warnings after detecting drowsiness symptoms, which can reduce physical condition errors and thus mitigate crash risk (Smith *et al.* 2000, Juvale *et al.* 2009, Mohanty *et al.* 2009).

A summary of the above discussion related to driver assistance systems is provided in Table 2.5. Note that the aforementioned driver assistance systems can be more beneficial in congested and more complex

Table 2. 5 ITS Countermeasures Versus Different Driving Errors and Violations

ITS Countermeasures	Reduction in Different Types of Driving Errors and Violations					
	Recognition	Decision	Performance	Violations	Physical Condition	Experience
Forward Collision Warning System	1 (Ervin 2005, Rakha 2011)	2 (Rakha 2011)	2 (Rakha 2011)	---	2 (Gaspar <i>et al.</i> 2019)	---
Adaptive Cruise Control System	1 (Ervin 2005, Rakha 2011)	3 (Ervin 2005, Rakha 2011)	1 (McDonald <i>et al.</i> 2001, Dang <i>et al.</i> 2015)	---	---	---
Lane Tracking System	1 (Ervin 2005, Hanowski <i>et al.</i> 2008, Lerner <i>et al.</i> 2010, Rakha 2011).	2 (Ervin 2005, Hanowski <i>et al.</i> 2008, Lerner <i>et al.</i> 2010, Rakha 2011).	2 (Ervin 2005, Hanowski <i>et al.</i> 2008, Lerner <i>et al.</i> 2010, Rakha 2011).	---	2 (Rimini-Doering <i>et al.</i> 2005)	---
Lateral Vehicle Detection System	1 (Rakha 2011, Blanco <i>et al.</i> 2016)	2 (Rakha 2011, Blanco <i>et al.</i> 2016)	2 (Rakha 2011, Blanco <i>et al.</i> 2016)	---	---	---
Dilemma Zone Mitigation System	---	2 (Rakha 2011)	---	1 (U.S. DOT 2009, Rakha 2011)	---	---
Blind Spot Assistance System	1 (Liu <i>et al.</i> 2017a, Sossa and Zamora 2020)	1 (Chovan <i>et al.</i> 1994)	---	---	---	---
Low-speed Automatic Brake System	1 (Cicchino 2019, Yang <i>et al.</i> 2019)	---	---	---	2 (Bao and Huang 2015)	2 (Bao and Huang 2015)
Drowsiness Detection System	---	---	---	---	1 (Smith <i>et al.</i> 2000, Juvalé <i>et al.</i> 2009, Mohanty <i>et al.</i> 2009)	---

Notes: The above recommendations about specific driving assistance systems are based on previous studies. Note that 1, 2, and 3 indicate high, medium, and low effectiveness of the particular ITS countermeasure in reducing specific driving error, respectively. Whereas, --- indicates that specific ITS countermeasures may reduce particular driving errors but has not been explored yet (to the best of our knowledge). For instance, most of the aforementioned driving assistance systems may help drowsy drivers without knowing their drowsiness level by generating audio/visual warnings for them. Similarly, most of the aforementioned driving assistance systems may help novice and inexperienced drivers to some extent; however, further research is needed.

surroundings like business/industrial, school locations, and residential areas where recognition errors, decision errors, performance errors, and violations contribute to a significant amount of crashes. The aforementioned recommendations about specific driving assistance systems are based on previous studies to reduce specific driving error(s) and violations. Note that intelligent systems such as lane tracking systems and lateral vehicle detection systems may also help in reducing lane-changing violations while generating audio or visual warnings to subject drivers in case of unsafe movement either by subject vehicle or other vehicles (i.e., indicated via --- in Table 2.5). Furthermore, we consider that all of the aforementioned driver assistance systems can potentially correct or reduce physical condition errors (especially in drowsy or fatigue driving) while generating audio/visual warnings for subject drivers. Additionally, most of the aforementioned driving assistance systems can be helpful for a novice or inexperienced drivers – something which needs further research. Other intelligent countermeasures such as detection systems for road surface conditions, weather, and vulnerable road users (pedestrians and bicyclists) can be tested and implemented to improve driver performance.

Similarly, built-environment changes (road designs) have the potential to reduce errors. For instance, implementing speed calming strategies and providing appropriate infrastructure for vulnerable road users (pedestrians and bicyclists) in areas (school zones, business/industrial, and moderate residential settings) may significantly reduce recognition and decision errors. While built-environment changes (road designs) can potentially reduce driving errors, further research is needed into the effects of built-environment and roadway conditions on crash risk through mediating driving errors/violations.

From a policy standpoint, conducting more research into appropriate training programs for drivers can help in reducing performance and experience errors. Such training programs were found effective in the case of riders (motorcyclists) as one of our studies suggests that riders participating in such training programs had a lower risk of injury crash (Wali *et al.* 2018d).

While the emerging driver assistance systems (as discussed above) could reduce various driving errors and violations, still accessing their safety benefits is not straightforward due to potential complexities which cannot be neglected once vehicles are equipped with these ITS countermeasures on a larger scale. For instance, subject drivers may start relying on these countermeasures (driver assistance systems) which if fail may lead to drastic safety outcomes. Having said that, we consider that driving errors and violations can extensively reduce crashes if the aforementioned countermeasures (driver assistance systems and better roadway design) along with appropriate response and recovery systems are implemented in a systematic manner.

Limitations

This study analyzed NDS data, which includes a finite number of drivers and geographical locations that do not explicitly cover all socio-geographical locations across the U.S. The TDEV proposed in this paper utilizes driving behaviors reported in the NDS data which may vary over time. In the original NDS data, the “locality” factor was classified by the data reductionists, who followed the procedures explained in the report. They included information about the localities where safety-critical events occurred. These could be more systematically categorized, i.e., based on the roadway as well as land-use factor classifications. Unfortunately, the research team did not have access to NDS data that would provide more complete information regarding all categories of roadway and land-use variables. In the future, it would be valuable if the NDS provided independent and complete data on roadway functional classification, activity-based land use, and locality-based land use. Such data would result in a more complete picture of the distribution of driving errors across different roadways, activity centers, and land uses.

Conclusions

This research developed a taxonomy for human errors and violations that lead to crashes, quantifying their contributions to crash occurrence. While it is possible to study driver behavior by examining police-reported crashes, they cannot provide the insights that naturalistic driving data provides. Using sensor-based and video technologies, the NDS data available in this study affords direct observation of pre-crash driver behavior. Using such data, this study applies a systematic taxonomy of driving errors and violations to explore how they vary across roadway and land-use contexts. The taxonomic framework coupled with the data analysis can be used to identify human factors that are most strongly associated with vehicle crashes and highlight the critical ones.

Based on the methodology proposed in the study, human factors contributed to 93% of crashes; roadway factors contributed in 17% of the crashes, and in 14% of the crashes both human and roadway and environmental factors (e.g., roadway conditions, visual obstructions, and weather conditions) were present. Digging deeper into human factors, in crashes, recognition errors (39%) were the most frequently reported driving errors, followed by decision errors (34%), performance errors (8%), and violations (9%). These statistics indicate that in crashes, recognition and decision phases are critical phases. The study further provided insights regarding how various driving errors and violations vary across different roadways and environments in crash events. While recognition errors were the most common resulting in crashes, recognition errors occurred more frequently when roadway environments were school zones (49%), urban areas (40%), moderate residential (39.13%), and business/industrial (39.1%) (but not in dense urban localities). Similarly, the contribution of decision errors to crashes was the greatest within

business/industrial (39%) followed by urban areas and interstate where decision errors resulted in 35% and 32% of the crashes occurring in these areas, respectively. The percentage of violations contributing to crashes was the largest (25% and 12% respectively) on roadways passing through rural/semi-rural areas and bypass/divided highways with traffic signals respectively. The findings reveal that recognition errors can be particularly hazardous, given their prevalence in crashes compared with their share in baselines, and they were also more frequently associated with severe crashes (51%) compared with other errors and violations. Moreover, based on the percentage of crashes per percentage of baselines in a specific locality, interstates and open country/open residential (rural and semi-rural settings) may pose lower risks. On the other hand, urban, business/industrial, and school locations seem to have a higher percentage of crashes per percentage of baselines and they can be considered as high crash risk locations. Other land uses where certain types of errors and safety-critical events are likely to occur are identified. The findings provide a foundation upon which to build a larger transportation safety program. The analysis provides valuable insights that can be aimed at reducing transportation crashes through data-driven strategies.

The prevalence of driving errors (recognition, decision, and performance errors) and violations in crashes reinforce the significance of intelligent driver assistance systems (e.g., forward-collision warning system, adaptive cruise control, lane tracking system, blind-spot detections system, lateral vehicle detection system, dilemma zone mitigation systems), low-speed automatic braking systems, and drowsiness detection systems which if implemented can significantly reduce these errors. Still, there is a need to further investigate their safety benefits as a small fault in these systems may lead to drastic safety outcomes. In the future, connected and automated vehicles (CAVs) have the potential to overcome a large portion of the driving errors and violations which presently contribute to a significant percentage of crashes. More research is needed on such safety intervention programs, e.g., collision warning systems and cooperative adaptive cruise control systems to explore how they may assist during the recognition, decision, and reaction phases, and re-engage the driver in hazardous situations. Similarly, built-environment changes (road design) have the potential to reduce driving errors. Conducting further research before deploying site-specific traditional countermeasures is needed. From a policy standpoint, research into designing and conducting appropriate training programs for drivers may help in reducing performance and experience errors. While connected and automated vehicles can reduce driving errors and violations in the future, still recognition errors, decision errors, performance errors, and unintentional violations can occur due to imperfections and failures of automation and connectivity technologies. The proposed taxonomy of naturalistic driving errors and violations can be extended or modified in the context of connected and/or automated vehicles to examine whether or not (and to what extent) automated and connected vehicles are prone to recognition, decision, and performance failures and unemotional violations even when the human

is fully disengaged. This could be done using test-bed data of connected and/or automated vehicles which would help to identify areas of improvement related to the safe performance/operation of CAVs. Note that the aforementioned suggestions regarding the potential application of our systematic taxonomy of driving errors and violations to CAVs only relate to future research which may help us understand how CAVs will behave and interact with other vehicles and roadway entities.

CHAPTER 3. BUILT ENVIRONMENT, DRIVING ERRORS AND VIOLATIONS, AND CRASHES IN NATURALISTIC DRIVING ENVIRONMENT

A version of this chapter is published in the Journal of Accident Analysis and Prevention. Also, a version of this chapter was presented at the 98th Annual Meeting of the Transportation Research Board, Washington D.C. (Jan 2019).

Ahmad, N., Wali, B., Khattak, A. J., & Dumbaugh, E. (2021). Built environment, driving errors and violations, and crashes in naturalistic driving environment. *Accident Analysis & Prevention*, 157, 106158.

Abstract

Driving errors and violations are highly relevant to the safe systems approach as human errors tend to be a predominant cause of crash occurrence. In this study, we harness highly detailed pre-crash Naturalistic Driving Study (NDS) data 1) to understand errors and violations in a crash, near-crash, and baseline (no event) driving situations, and 2) to explore pathways that lead to crashes in diverse built environments by applying rigorous modeling techniques. The “locality” factor in the NDS data provides information on various types of roadway and environmental surroundings that could influence traffic flow when a precipitating event is observed. Coded by the data reductionists, this variable is used to quantify the associations of diverse environments with crash outcomes both directly and indirectly through mediating driving errors and violations. While the most prevalent errors in crashes were recognition errors such as failing to recognize a situation (39%) and decision errors such as not braking to avoid a hazard (34%), performance errors such as poor lateral or longitudinal control or weak judgment (8%) were most strongly correlated with crash occurrence. Path analysis uncovered direct and indirect relationships between key built-environment factors, errors and violations, and crash propensity. Possibly due to their complexity for drivers, urban environments are associated with higher chances of crashes (by 6.44%). They can also induce more recognition errors which correlate with an even higher chance of crashes (by 2.16% with the “total effect” amounting to 8.60%). Similar statistically significant mediating contributions of recognition errors and decision errors near school zones, business or industrial areas, and moderate residential areas were also observed. From a practical applications standpoint, multiple vehicle technologies (e.g., collision warning systems, cruise control, and lane tracking system) and built-environment (roadway) changes have the potential to reduce driving errors and violations which are discussed in the paper.

Introduction

Road traffic crashes are one of the leading causes of both social and economic losses. Studies reveal that a majority of these crashes occur as a result of human factors including distracted driving and other behavioral factors (Singh 2015). Other crash contributing factors include roadway environmental and vehicle factors.

To mitigate these crashes and the consequent socio-economic losses, there is a need to explore key reasons that contribute to crashes. Fortunately, the advent of cyber-physical systems (CPS) provides an indispensable opportunity to obtain an objective and nuanced understanding of key crash contributing factors (Wali and Khattak 2020). As complex layers of digital information are generated within the urban transportation landscape, new innovative techniques for the study of safety-critical driving errors and violations are needed. With countless terabytes of real-world driving behavior data, cyber-physical systems allow retrofitting or the creation of safer and more resilient transportation systems. When such data are combined with advanced data analytical techniques, the issue of driving errors and violations and their respective contributions to crash occurrence can be objectively examined. To this end, the present study harnesses the SHRP2 Naturalistic Driving Study (NDS) database to develop a deeper understanding of the critical behavioral factors which contribute to safety-critical events (SCEs) – overcoming the substantial limitations of after-the-fact police crash investigations of pre-crash driver behavior. The SHRP2- NDS data is unique because it provides extensive data on real-world driver behaviors and performance, real-world risk and safety-critical outcomes (Dingus et al. 2006, Carney *et al.* 2015a, b, Dingus *et al.* 2016, Arvin *et al.* 2019). This allows examining human behaviors in more detail, e.g., errors as a result of driver distraction, inattention, judgment, and drowsiness, all of which contribute to crashes. Several geo-coded driving parameters like acceleration, speed, secondary tasks/durations, distraction, time to collision, and eye glance behavior of drivers can be used to examine real-world driving behavior and performance (Dingus et al. 2006, Dingus *et al.* 2016). Notably, such a plethora of information is not only available for SCEs but for baselines/normal driving events as well as establishing ground truth regarding prevalent safe and unsafe driving behaviors in normal setups. All of these innovative features make SHRP2-NDS more reliable and unique than conventional police crash reports, which could be biased due to the level of complexity involved in obtaining information after a crash has occurred (Hankey *et al.* 2016). Rather than analyzing individual driving behaviors which contribute to crashes, studies developed taxonomies to classify driving behaviors into various driving errors and violations through different conceptual themes (Treat *et al.* 1979, Reason *et al.* 1990, Petridou and Moustaki 2000, Wierwille *et al.* 2002, Stanton and Salmon 2009, Salmon *et al.* 2010). For details of how previous studies classified driver behaviors, please refer to the summary of studies in Table 3.1. While the studies provide useful insights, they were based used data from police crash reports which could be subjective and bias. Realizing the importance of real-world SHRP2-NDS data, one of the recent studies, developed systematic taxonomy of naturalistic driving errors and violations (Khattak *et al.* 2021a).

Table 3. 1 Summary of Previous Taxonomies, and Types of Driving Errors and Violations

Authors	Classification of Human Errors with Examples	Data Source and Conclusion
(Treat <i>et al.</i> 1979)	Human factors: Direct Causes a) Recognition errors e.g., internal distraction b) Decision errors e.g., excessive speed, tailgating c) Performance errors e.g., improper driving technique d) Critical non-Performance errors e.g., dozing e) Non-performance errors e.g., suicide or anger Human conditions and States - a). Physical/Physiological conditions e.g., fatigue, impairment - b). Mental/Emotional conditions e.g., emotionally upset, in hurry - c). Driver experience e.g., unfamiliarity with road or vehicle	Data Source: ▪ Police Crash Reports ▪ On-Site Investigation Conclusions (In-Depth Investigation): ▪ Human factors contribute to 93% of crashes ▪ Roadway environment cause 33.8% of crashes ▪ Vehicle factors contribute to 12.6% of crashes ▪ Recognition errors and decision errors contribute to 41.4%-56% and 28.6-52.1% crashes respectively
(Reason <i>et al.</i> 1990)	a) Dangerous driving errors e.g., overtake the queue b) Comparatively harmless driving errors (lapses) e.g., miss motorway exit c) Violations e.g., disregard speed at night, race for gap	Data Source: ▪ Behavior Questionnaire Survey of 520 Drivers Conclusions: ▪ Violations reduce with age while errors do not ▪ Women are more prone to harmless lapses
(Reason 2000)	Discusses human error management with two approaches: a) Person approach (focus and blame person for inattention/forgetfulness/weakness) b) System approach (focusing on conditions to mitigate errors)	▪ Psychological conceptualization of human errors from a clinical practice perspective ▪ A thorough understanding of the two can help in mitigating the chance of accidents in clinic practice.
(Petridou and Moustaki 2000)	a) Factors that reduce driving capabilities • In short term e.g., temporary distraction, binge eating, and drowsiness • In long term e.g., disability, aging, inexperience, and drug use b) Factors that encourage risky driving • In short term e.g., suicidal behavior, moderate alcohol use • In long term e.g., intoxication, routine speeding	Reviewed behavioral factors contributing to crashes by search the following keywords: ▪ Outcomes (e.g., road traffic crash/accident/injury) ▪ Risk factors (e.g., stress, risk-taking, experience) ▪ Morbidity etc.
(Stanton and Salmon 2009)	Generic error taxonomy applied to classify driving errors into: a) Action errors (e.g., action too much, action too little, fail to act) b) Cognitive and decision errors (e.g., wrong assumption) c) Observation errors (e.g., failed to observe) d) Information retrieval errors (e.g., misread information) e) Violations (intentional and unintentional violations)	▪ Reviewed literature and proposed generic taxonomy of driving errors using various psychological mechanisms as: • attention, perception, situation assessment, planning & intention, memory & recall, and action execution. ▪ Technological solutions proposed to prevent/reduce errors

The key idea behind the analysis is that specific roadway environmental factors may increase the likelihood of different types of human errors and violations (obtained through TDEV (Khattak *et al.* 2021a)), which may lead to crash events. Thus, while using a systematic Taxonomy of Driving Errors and Violations (TDEV) proposed by the authors (Khattak *et al.* 2021a), we apply rigorous discrete outcome-based path analysis (reflecting systems thinking) to explore how environmental conditions correlate with errors, violations, and crash occurrence. A unique aspect of this study is the utilization of emerging SHRP2-NDS data, which provide more objective information about pre-crash or pre-near-crash driver behaviors in diverse spatio-temporal contexts. In addition to considering crash events, normal driving events are analyzed vis-à-vis near-crash (near-miss) outcomes, as such “close calls” may foreshadow future crashes. The NDS database is ideal for this analysis as it contains human errors in normal driving as well as near-miss (near-crashes) and crashes situations throughout a diverse spectrum of behavioral and roadway environmental conditions.

Research Objectives and Contribution

Given the aforementioned gaps, this study adds to the existing state-of-the-art regarding crash prevention by focusing on the key gaps listed below. While the role of human factors and driver errors is well-recognized in the literature - little to no evidence exists about the role of human errors assessed through comprehensive and objective driving data such as naturalistic driving. First, we use the taxonomy of naturalistic driving errors and violations (Khattak *et al.* 2021a) to classify driver behaviors which are then used in the analysis by focusing on the following two gaps.

1. The role of roadway environments in encouraging or discouraging certain naturalistic risky behaviors is not well-understood, and;
2. The multidimensional connections between roadway environment, driving errors and violations, and the ultimate safety outcomes are not explicitly quantified. Compared to subjective data sources (Anastasopoulos *et al.* 2012, Venkataraman *et al.* 2013, Ahmad *et al.* 2019), the use of objective real-world safety data (such as those used in this study) leads to a more enriched understanding of the key behavioral elements influencing safety - ultimately leading to more effective crash prevention strategies.

The subsequent sections in this paper include literature review, methodology, results, discussion, limitations, and conclusions. The “literature review” section briefly discusses the existing research associated with driving errors, roadway environmental factors, and other key factors which contribute to crashes. The “methodology” section briefly discusses the NDS data source, the taxonomy of driving errors and violations (Khattak *et al.* 2021a), diverse roadway environments in NDS data, and path analysis along

with two constituent models applied in this research. The “results” and “discussion” sections interpret and discuss key results of the path analysis and its constituent models. The limitations of this research are briefly discussed in the “limitations” section, while the conclusion section concludes the key research theme, contribution, and findings of this research.

Methodology

Data Source

The NDS data includes comprehensive information on driving behavior, roadway environment, traffic conditions, vehicular movement, and environmental factors extracted from the videos by well-trained data reductionists, who followed an appropriate data collection protocol to ensure consistent data. For example, the General Estimates Systems (GES) procedures, compiled by National Highways Traffic Safety Administration (NHTSA), were used as a reference to develop variables during SHRP2- NDS data extraction from videos. The data reductionists were recruited by the Virginia Tech Transportation Institute (VTTI) as part-time employees and were students of either Virginia Tech or other neighboring universities, colleges, and communities (Hankey *et al.* 2016). The majority of these reductionists had an educational background in the fields of psychology and human factors engineering and had good attention capabilities contributing to quality data reduction from NDS videos (Hankey *et al.* 2016). Moreover, minimum qualification requirements for the data reductionists included high school diploma, some college-level coursework, having legal age for signing non-disclosure agreements, computer proficiency, prior data entry or similar skill, sound critical thinking capabilities, and a U.S. driving license which could help them to understand driving scenarios (Hankey *et al.* 2016). Before NDS data coding, reductionists were required to pass proficiency tests (Hankey *et al.* 2016). For more details, see (Hankey *et al.* 2016). Overall, the SHRP-NDS data reductionists were well-qualified to code variables in NDS data. These variables can be classified into three key categories (Hankey *et al.* 2016):

- **Safety-Critical Event Variables:** This category includes the variables used to establish sequence and scenarios of events before and during safety-critical events. Variables in this class include event nature, event severity, pre-incident maneuvers, precipitating events, driver’s reaction, post-maneuver control, vehicle configurations, details about other drivers, fault status of the drivers, vehicles involved, and objects involved.
- **Driver Variables:** This class of variables includes information about the subject driver’s behavior, performance, and condition before and during a safety-critical event. These variables include driving behavior, type and duration of secondary tasks, driver impairment, seatbelt use, and the driver’s steering control.

- **Built and Roadway Environment Variables:** In the NDS data, environmental variables including locality type (referred to as roadway environment factors in this paper), traffic density, traffic flow, traffic control devices, number of travel lanes, roadway surface conditions, roadway alignment, lighting, and weather conditions.

Information on crashes, near-crashes, and baselines is provided in the “event data” file. A baseline event is defined as a “normal” driving event in which the driver had no crash and/or near-crash event (Hankey *et al.* 2016). Note that incorporating baselines in the study design help capture the exposure of the population to unsafe events, providing an estimate of a “typical driver behavior.” The crash and near-crash events are defined below:

- Baselines refer to normal driving situations which are not associated with any safety risk and can be regarded as routine (non-event) driving. The baseline drivers were not involved in crashes or near-crash. It should be noted that some drivers in baseline events may be involved in risky driving behavior which could be dangerous as such behaviors may lead to SCEs in the future.
- **Crash:** A crash is an event in which kinetic energy is measurably dissipated or transferred as a result of contact of the subject vehicle with any other fixed or moving vehicle, or object, which can include pedestrians, animals, and barriers. Crashes also include events where at least one tire of the vehicle leaves the intended or paved roadway surface.
- **Near-Crash:** The situation where a rapid evasive maneuver is required by either the subject vehicle or the other vehicle, cyclist, animal to avoid a crash is termed as a near-crash. The evasive maneuvers include braking, steering, accelerating, or some combination of these control inputs.

Note that the majority of the event data file is populated with information extracted from real-world event video files, and as such the information is considerably more reliable than the information in police crash reports (Hankey *et al.* 2016). This study uses a subset of original NDS-SHRP2 data including a total of 9,593 trips which include 7,589 baselines, 1,331 near-crashes, and 673 crashes. The NDS data reductionists term video recording for safety-critical events (crash and near-crash) or baselines (with no SCE) as “epoch” (Papazikou 2019). The epoch for baseline and safety-critical events (crash or near-crash) is generally 20 seconds and 30 seconds respectively (Wali and Khattak 2020). Note that 30 seconds epoch for SCE usually includes 20 seconds before and 10 seconds after the event which makes SCE comparable to baseline (with no SCE) based on the first 20 seconds data used for deriving driving behavior and performance in SCE (Papazikou 2019). It is also important to mention that in few cases of baselines and safety critical events, the duration is less than 20 seconds and 30 seconds respectively (Wali and Khattak 2020, Wali *et al.* 2020). However, this is not threatening to the analysis, details can be found here (Wali and Khattak 2020, Wali *et al.* 2020).

Systematic Classification of Driver Errors and Violations – NDS Data

This study applies a recently developed taxonomy for driving errors and violations (TDEV) to classify human factors reported in crashes and near-crashes (Khattak *et al.* 2021a). Note that human factors were also reported in baselines (where no safety-critical event occurred) and can be classified using TDEV which helps in quantifying the crash risk (Khattak *et al.* 2021a). In the case of safety-critical events, when no “Driver Behavior” was reported, we checked the “Secondary Task” and “Secondary Task Outcome” variables in the NDS subsample to determine if a particular secondary task contributed to a crash or near-crash. This research uses two variables including driver behavior and secondary task (given that it contributed to the occurrence or severity of safety critical event when no driver behavior was reported) to derive six driving errors and violations which are discussed below:

Driver Behavior: The NDS (SHRP2) data reductionists define the “Driver Behavior” as: “Driver behaviors (those that either occurred within seconds before the Precipitating Event or those resulting from the context of the driving environment) that include what the driver did to cause or contribute to the crash or near-crash. Behaviors may be apparent at times other than the time of the Precipitating Event, such as aggressive driving at an earlier moment which led to retaliatory behavior later.” A comprehensive list of driver behaviors reported in SCEs (where driver behaviors contribute to crash/near-crash occurrence) along with their classification using TDEV can be found in (Khattak *et al.* 2021a). Note that some of these driver behaviors were also reported in baselines but no safety-critical event occurred.

Secondary Task and Secondary Task Outcome: In some crash and near-crash events where no driver behavior was reported, drivers were involved in some secondary tasks. Hence, we checked the “Secondary Task” and “Secondary Task Outcome” factors to determine whether a specific secondary task contributed to a crash/near-crash event. The NDS data reductionists define the “Secondary Task Outcome” as: “Determination of whether the Secondary Task contributed to the event sequence and severity. (Not whether the factor caused the event but contributed to it.)” Such secondary tasks contributing to safety-critical events include doing personal hygiene while driving, cell phone use, using hands-free communication, non-specific internal eye glance, and talking/listening. For details, please refer to (Khattak *et al.* 2021a). Finally, the driver behaviors and secondary tasks (which contribute to safety-critical events) were used to systematically derive driving errors and violations using the proposed TDEV crashes (Khattak *et al.* 2021a).

Using information coded in the three variables including driver behavior, secondary task, and secondary task outcome, human factors, which contribute to crashes, were classified into five key driving errors (recognition errors, decision errors, performance errors, physical conditions-related errors, and experience

or exposure errors) and violations. For brief definitions of driving errors and violations, please refer to Chapter 1 or see Khattak et al. 2021).

Diverse Roadway Environment and Localities

The data reductionists classified diverse roadways and environments into various categories via the “locality” variable in the NDS data². The categories included, in the NDS data, under the “locality” variable include a bypass/divided highways with traffic signals (principal arterials), interstate/bypass/divided highway with no traffic signals, open residential (which include few houses, signifying largely undeveloped land use), moderate residential, business/industrial, school, church, playground, open country (rural) settings, urban areas, and other localities (e.g., campgrounds). According to the SHRP-2 NDS dictionary, open country and open residential indicate rural and rural/semi-rural areas respectively. We combine the two categories representing similar environments to obtain useful insights. Some other localities including a playground, other (e.g., campground), and church are combined as their percentages reported in the data were very low. Also, these localities, specifically playgrounds, and campgrounds represent somehow similar activities. Finally, while using the roadway locality as an independent variable in the models, we consider the “interstate” as the base category. For detailed definitions and descriptions of various localities defined by SHRP2-NDS data reductionists, please refer to the study in parenthesis (Khattak *et al.* 2021a).

Path Analysis Framework

Conceptually, we hypothesize that roadway environments (localities) may induce certain driving errors and violations (human errors) which may result in crashes (Figure 3.1). The driving behavior is classified via the taxonomy of driving errors and violations (TDEV) using SHRP2-NDS data (Khattak *et al.* 2021a), and then a two-stage path analysis framework is applied to achieve the study objectives. In the first stage, a discrete outcome model (multinomial logit model) was developed for driver errors and violations with key explanatory variables as shown in Figure 3.1. In the second stage, a discrete outcome model (ordered Probit model) was developed for crash propensity with driver errors and violations (response outcome in Stage 1) and other key factors (e.g., roadway environments) as explanatory variables as shown in Figure 3.1.

² According to the NDS dictionary, the “locality” variable is described as: “Best description of the surroundings that influence or may influence the flow of traffic at the time of the start of the precipitating event. If there are ANY commercial buildings, indicate as business/industrial or urban area as appropriate (these categories take precedence over others except for church, school, and playground). Indicate school, church, or playground if the driver passes one of these areas (or is imminently approaching one) at the same time as the beginning of the Precipitating Event (these categories take precedence over any other categories except urban, and divided highway).”

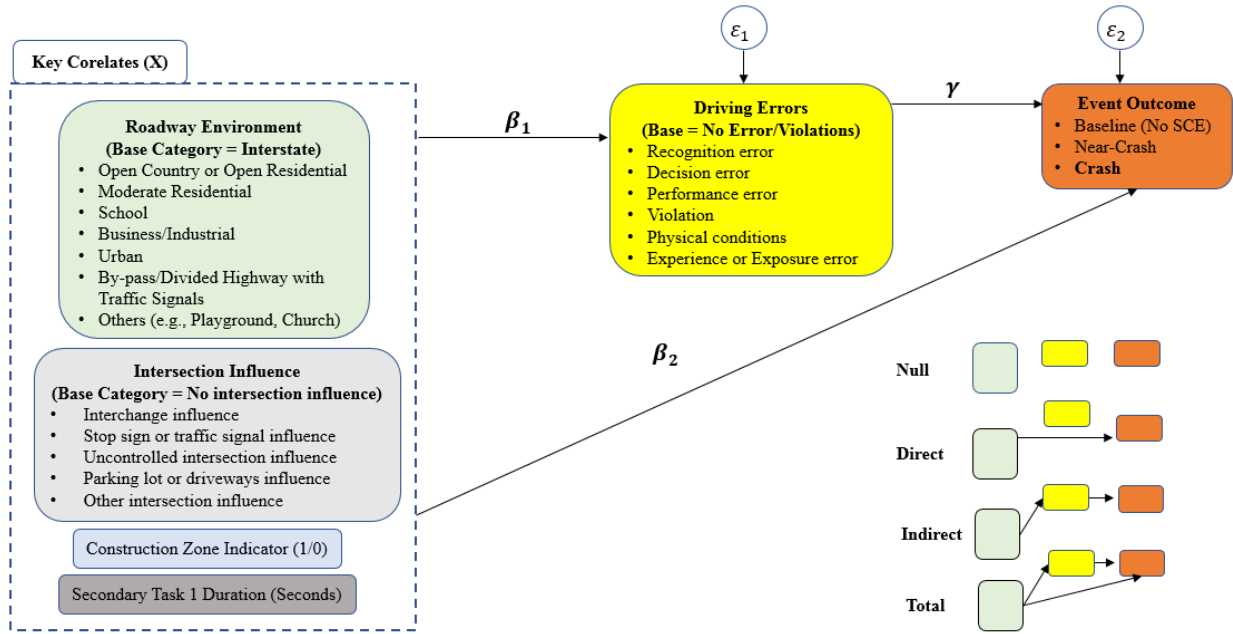


Figure 3. 1 Methodological Framework

Notes: Y_1 (Driving errors and violations) is categorical outcome which is modeled through a multinomial logit framework; Y_2 (event outcomes including baseline (no event), near-crash, and crash) being ordinal is modeled through an ordered probit regression. In the case of the ordered probit model, we check whether one or more of the explanatory variables violate the proportional odds assumption. If one or more of the explanatory variables violate the proportional odds assumption in the event outcome model, a more sophisticated model will be developed which relaxes the proportional odds assumption for those variables violating it.

By using the path structure shown in Figure 3.1, the system's framework allows the decomposition of complex structures embedded in the data and the estimation of relevant direct and indirect effects, as discussed earlier (Liu *et al.* 2015, Zhang *et al.* 2018, Kamrani *et al.* 2019). The two methods are discussed in detail below:

Multinomial Logit Model

Given the categorical nature of the driving errors and violations, a multinomial logit model (driver error model) was estimated which can be formulated as (Gkritza *et al.* 2010, Celik and Oktay 2014, Waseem *et al.* 2019):

$$Y_i = \beta_0 + \beta_i X_i + \varepsilon_{in} \quad (1)$$

where Y_i is a categorical variable with seven levels defining driver errors/violations including no driving error and violation as a base outcome (utility function); β_0 indicates a constant term; β_i is a vector of estimable parameters for associated factors X_i related to i^{th} driver error/violation for the n^{th} observation in the multinomial logit model. For simplicity and understanding, β_i and ε_i are shown as β_1 and ε_1 in Figure 3.1. With the error terms ε_{in} assumed to follow a generalized extreme value distribution, the probability of each driver's error/violation can be calculated as (McFadden 1981, Celik and Oktay 2014):

$$P_n(i) = \frac{\exp(\beta_i X_i)}{\sum_{j=1}^J \exp(\beta_j X_j)} \quad (2)$$

Where X_i indicates a vector of explanatory variables associated with i^{th} driving error/violation (*i.e.*, $i = 0, 1, 2, 3, \dots, I$) for n^{th} observation; and $P_n(i)$ is the likelihood for observing i^{th} driving error/violation for n^{th} observation. In a multinomial logit model setup, the marginal effects (MEs) indicate the change in the probability of each driver's error or violation relative to the base outcome which is no error in this case with a unit change in a specific explanatory variable while keeping all other variables at their constant values.

Ordered Probit Model

As the event outcomes (baselines, near-crashes, and crashes) follow a three-level ordinal scale, an ordered probit model is estimated and shown in Eq. 3 below (Washington *et al.* 2010, Wali *et al.* 2017). To calculate the indirect effects of associated factors on crash outcome mediated by driving errors and violations, the model includes driving errors or violations as an explanatory factor (indicated by Y_1 in Eq. 3) in addition to other roadway environment variables/other controls (indicated by X in Eq. 3). The formulation follows as (Liu *et al.* 2015, Wali *et al.* 2017):

$$Y_2 = \beta_0 + \beta_2 X_i + \gamma Y_1 + \varepsilon_{(2)} \quad (3)$$

Where Y_2 indicates event outcome (baseline, near-crash, crash) following an ordinal scale; β_0 indicates a constant term; β_2 is a vector of coefficients of the associated factors X_i in ordered probit model; γ indicates the association of driver errors/violation (six types of driving errors and violations including no driving error or violation as a base outcome) with the event outcome; and $\varepsilon_{(2)}$ is the associated error term assumed to have a normal distribution.

To relate the latent measure of event outcome (Y_2) with the observed event outcome (Y_2^*), the following parameterization is used (Kockelman *et al.* 2002, Abdel-Aty 2003, Ahmad *et al.* 2019):

$$Y_2^* = \begin{cases} 1 & \text{if } -\infty \leq Y_2 \leq \mu_1 \text{ (baseline)} \\ 2 & \text{if } \mu_1 < Y_2 \leq \mu_2 \text{ (near - crash)} \\ 3 & \text{if } \mu_2 < Y_2 \leq \infty \text{ (crash)} \end{cases} \quad (4)$$

Where Y_2 is the latent and continuous measure of event outcome a driver can be involved in; Y_2^* is the observed event outcome (level); and μ_i indicates the threshold values which are to be estimated along with the coefficients (*i.e.*, β_2 and γ) as shown in equation (3). As the outcome consists of three ordered categories, the estimable probabilities are fully identified by the estimation of two threshold/cut-off points.

Using the estimable parameters from Eq. 3 and 4, the probabilities of each of the three outcomes can be calculated as follows (O'donnell and Connor 1996, Washington *et al.* 2010, Ahmad *et al.* 2019):

$$P(Y_2^* = l) = \Phi(\mu_l - \beta_2 X_i) - \Phi(\mu_{l+1} - \beta_2 X_i) \quad (5)$$

Where l indicates the response alternatives (baseline, near-crash, and crash coded as 1, 2, and 3 respectively), $P(Y_2^* = l)$ indicates the chance that an individual n is involved in an event l , while Φ indicates the standard normal distribution function as shown in equations 8-10. Similarly, the probability of each event outcome can be found with the driving errors and violations (Y_1) included as an explanatory variable in equation (3) of the ordered probit model (Washington *et al.* 2010):

$$P(Y_2^* = l) = \Phi(\mu_l - \gamma Y_1) - \Phi(\mu_{l+1} - \gamma Y_1) \quad (6)$$

$$\Phi(\mu) = \frac{1}{2\pi} \int_{-\infty}^{\mu} e^{\left(-\frac{1}{2}\omega^2\right)} d\omega \quad (7)$$

In case of the ordered probit model, the *MEs* for continuous explanatory variable (e.g., secondary task duration in our case) giving the probability change in the occurrence of an event (e.g., crash) with a unit increase in the value of continuous variable from its mean value keeping all other variables at their mean or constant values can be computed as (Washington *et al.* 2010):

$$\frac{\partial P(Y_2^*=l)}{\partial X} = -[\phi(\mu_l - \beta_2 X_i) - \phi(\mu_{l+1} - \beta_2 X_i)]\beta_2 \quad (8)$$

Similarly, the marginal effects for the indicator variable(s) in an ordered probit setup can be given via equations (9) and (10) for the n th associated factor (X_i) and n th driving errors and violations (Y_{1n}) given below as (Washington *et al.* 2010):

$$\partial P(Y_2^* = l|X) = P(Y_2^* = l|X_i = 1) - P(Y_2^* = l|X_i = 0) \quad (9)$$

$$\partial P(Y_2^* = l|Y_i) = P(Y_2^* = l|Y_i = 1) - P(Y_2^* = l|Y_i = 0) \quad (10)$$

Equation 9 and 10 determines the change in the probability of specific category of the response outcome when indicator variable(s) for the associated factors (e.g., roadway environment variables, intersection influence, presence of construction zone, and secondary task duration) and driving errors and violations switch from zero to one respectively. Whereas, Y_i indicates specific driving errors and violations.

For ordinal data, studies generally use ordered probit or ordered logit models (proportional odds models) (Williams 2016, Wali *et al.* 2017, Ahmad *et al.* 2019). However, proportional odds models are appropriate if all explanatory variables in the model meet proportional odds (or parallel line) assumptions. If the assumption is violated by at least one of the explanatory variables, then partial proportional odds/generalized ordered probit (goprobit) or logit (gologit) models are found to be more appropriate which relate the assumption of variable(s) which violates the assumption (Williams 2005, 2016). While non-ordinal (multinomial logit or probit) models can be used as a remedy when proportional odds assumption is violated by one or more variables; however, these models are less parsimonious and not easily interpretable in the case of ordinal data (Williams 2005, 2016). On other hand, partial proportional odds models including goprobit and gologit are more parsimonious and accurate (Williams 2005, 2016). Furthermore, the goprobit or gologit models can also capture the heterogeneity across different categories of the response variable (Williams 2005, 2016). Having said this, we estimate three alternatives ordered probit models to explore “event outcome” including baseline, near-crash, and crash. The three models are based on the level of imposing proportional odds assumptions (Williams 2005, 2016). Model 2A can be regarded as the proportional odds model which assumes that all explanatory variables meet the proportional odds assumption (Williams 2005, 2016). In Model 2B (fully non-proportional odds model) and Model 2C (partially proportional odds model), the assumption is relaxed fully and partially, respectively (Williams 2005, 2016). The Chi-square statistics (similar to Brant Statistics) can be used to see whether one or more variables violation the parallel line assumption (Williams 2005, 2016). Referring to the interpretation, in Model 2B and Model 2C, the coefficients can be interpreted like binary probit type models with nominal outcomes collapsed into two classes (Williams 2005, 2016). In this study, the three outcomes are baseline,

near-crash, and crash. The coefficients for the first utility function can be interpreted similarly as one would do for a binary probit model for baseline vs. near-crash + crash (Williams 2005, 2016). The second set of coefficients can be interpreted as from a binary logit model for baseline + near-crash vs. crash (Williams 2005, 2016). This being said, a positive parameter estimate for a particular explanatory variable implies that the chance of getting a ‘higher’ value (near-crash + crash for the first utility function and crash for the second utility function) is more (Williams 2005, 2016). Furthermore, marginal effects can be computed for original categories of the response variable (baseline, near-crash, and crash) and are interpreted similar to those obtained for proportional odds ordered probit model (Model 2A) or multinomial logit model (Williams 2005, 2016).

Path Analysis

In the path analysis framework, the marginal effects from both of the constituent models are used in obtaining the direct, indirect, and total effects of the associated factors on the final response variable (i.e. crash outcome in this case) (Liu *et al.* 2015, Zhang *et al.* 2018). The standard path analytic framework, used in this research, assumes the two error terms to be uncorrelated (Liu and Khattak 2018, Arvin *et al.* 2019).

$$\text{Indirect MEs} = (\text{MEs from multinomial logit model}) * (\text{MEs from ordered probit model}) \quad (11)$$

$$\text{Total MEs} = \text{Direct MEs} + \text{Indirect MEs} \quad (12)$$

The indirect effects of various roadway environments (compared to interstate) and other key explanatory variables on the likelihood of a crash through specific driving errors or violations (compared to no driving error and violation) are computed by multiplying MEs obtained from the driver error model (Model 1) with ME obtained from the crash outcome model (Model 2C) (Liu and Khattak 2017, Arvin *et al.* 2019). Note that “crash outcome model” provides the direct effect of specific roadway environment and other key explanatory variables on the likelihood of a crash, and which is added to the indirect effect, provide the total effect of the specific explanatory variable on crash risk (Liu and Khattak 2017, Arvin *et al.* 2019). The equations for determining the indirect and total effects of specific roadway environments or other explanatory variables are given below (Liu and Khattak 2017, Arvin *et al.* 2019).

Results

Descriptive Statistics

Table 3.2 presents summary statistics of the two response variables and other key explanatory variables. The NDS data includes 9,593 observations containing 7,589 baselines (79%), 1,331 near-crashes (14%),

and 673 crashes (7%). The descriptive statistics show that in 90% of the baselines, no driving error or violation was observed (as expected). However, at least one driving error or violation was present in a significant portion of the near-crash and crash events, 61%, and 93%, respectively (Table 3.2). In both crash and near-crash events, recognition and decision errors were the predominant error types, collectively accounting for 48% of the near-crashes and 73% of the crash events (Table 3.2). Hence, recognition errors and decision errors respectively comprise 39% and 34% of crashes. At least one error or violation was present in 10% of the total baseline events (Table 3.2). While driving errors and violations made by almost 10% of the baseline drivers did not result in safety-critical events (crash and near-crash), these errors and violations can be dangerous in the future if drivers continue making these errors and violations. Having said this, strategies to prevent such errors and violations need to be explored from a behavioral standpoint as we found that all the key types of driving errors and violations increase crash risk.

The percentage of drivers traveling on roadways passing through business/industrial locations was 31%, 40%, and 46% for baselines, near-crashes, and crashes respectively (Table 3.2). Furthermore, the percentage of the NDS drivers traveling on roadways within school zones is about 5% for baselines, but 7% for near-crashes and 8% for crashes (Table 3.2). Also, based on the data coded in NDS, 7% and 8% of the total NDS near-crashes and crashes occurred in urban areas (Table 3.2). Around 25% and 20% of the total NDS crashes occurred due to the influence of traffic signal/stop sign and parking lot/driveways respectively (Table 3.2). Of all the crashes, 4% were observed in construction zones compared with about 3% in baselines at construction zones. The mean secondary task duration in both crash and near-crashes is almost twice that in baselines which were expected (Table 3.2). For detailed percentage distribution (both row and column-wise) of baselines, near-crashes, and crashes across different driving errors and violations, diverse roadway environments, and various intersection influences, see Table 3.2.

Modeling Results

For each outcome type, the final models were systematically derived while including the most important variables (based on research objectives), statistical significance, and specification parsimony. In doing so, the likelihood ratio test, Akaike information criterion “AIC” (Akaike 1974), and Bayesian information criterion “BIC” (Schwarz 1978) statistics were examined. In comparing the competing models, the one with lower AIC and BIC statistics indicates superior fit after adjusting for degrees of freedom. Studies also suggest using other performance measuring to access the model performance depending on the scope of the study. For instance, studies show that BIC and Schwarz criterion “SBIC” as more parsimonious criteria; however, SBIC outperforms BIC (Nickelsburg 1985). For the sake of completeness, SBIC (Nickelsburg 1985), Consistent AIC “CAIC” (Bozdogan 1987),

Table 3. 2 Descriptive Statistics (Row and column percentages of Baselines and SCEs)

Explanatory Variables	Overall Cases (N = 9,593)	Event Outcome (9,593)			Event Outcome (N = 9,593)			
		Column-wise (%)			Row-wise (%)			
		*B	*NC	*C	*B	*NC	*C	Sum Row (%)
Driving Errors	% (S.D)	%	%	%	%	%	%	
No Driving Errors	77.16 (0.419)	90.12	38.69	7.13	92.39	6.96	0.65	100
Recognition Errors	7.60 (0.265)	0.22	34.03	38.63	2.33	62.05	35.62	100
Decision Errors	6.45 (0.246)	2.69	13.82	34.32	32.96	29.73	37.32	100
Performance Errors	0.69 (0.083)	0.09	0.68	7.58	10.45	13.43	76.12	100
Violations	6.53 (0.247)	5.56	10.74	9.06	67.41	22.84	9.74	100
Physical Conditions	1.28 (0.113)	1.25	1.43	1.34	77.24	15.45	7.32	100
Experience or Exposure Errors	0.27 (0.052)	0.07	0.60	1.93	19.23	30.77	50.00	100
Sum Column (%)	100%	100	100	100	---	---	---	---
Roadway Environment								
Interstate	23.29 (0.423)	25.60	18.78	6.09	86.97	11.19	1.84	100
Open Country or Open Residential	8.66 (0.281)	9.46	5.48	5.94	86.40	8.78	4.81	100
Moderate Residential	21.21 (0.409)	22.27	15.55	20.51	83.05	10.17	6.78	100
School	5.10 (0.219)	4.51	7.06	7.88	69.94	19.22	10.84	100
Business/Industrial	34.08 (0.474)	31.95	40.05	46.21	74.18	16.30	9.51	100
Urban	2.76 (0.164)	1.52	7.29	7.88	43.40	36.60	20.00	100
Bypass or Divided Highways with no Traffic Signals	3.00 (0.176)	3.04	3.01	2.53	80.21	13.89	5.90	100
Others (e.g., Church, Playground)	1.90 (0.136)	1.65	2.78	2.97	68.68	20.33	10.99	100
Sum Column (%)	100%	100	100	100	---	---	---	---
Intersection Influence								
No Intersection	71.75 (0.449)	81.42	32.83	41.01	89.65	6.34	4.00	100
Interchange	3.22 (0.176)	2.33	8.41	2.97	57.28	36.25	6.47	100
Stop Sign or Traffic Signal	15.78 (0.365)	12.06	32.38	24.96	60.44	28.47	11.10	100
Uncontrolled Intersection	3.38 (0.181)	1.75	10.14	8.32	41.05	41.67	17.28	100
Parking lot or Driveway Entrance/Exit	4.65 (0.211)	1.98	11.95	20.36	33.63	35.65	30.72	100
Other (e.g., crosswalk, roundabouts)	1.22 (0.109)	0.46	4.28	2.38	32.41	52.78	14.81	100
Sum Column (%)	100%	100	100	100	---	---	---	---
Presence of Construction Zone								
Presence of Construction Zone (No)	96.21 (0.191)	96.60	93.99	96.14	79.43	13.56	7.01	100
Presence of Construction Zone (Yes)	3.79 (0.191)	3.40	6.01	3.86	70.88	21.98	7.14	100
Sum Column (%)	100%	100	100	100	---	---	---	---

Note: *B, *NC, and *C indicate baselines (N= 7,589), near-crashes (N = 1,331), and crashes (N = 673) respectively. For Secondary Task 1 Duration (seconds), the mean (standard deviation “S.D”) in Bs, NCs, and Cs are 1.75 (2.16), 3.28 (3.83), and 3.58 (4.19) seconds respectively. % indicates percentage.

and Bozdogan's Information Complexity "ICOMP" (Bozdogan 1990) is considered (where applicable) to select the final model while comparing it to its counterparts. For details about the criteria used to assess the model performance, please refer to the studies in parenthesis (Akaike 1974, Schwarz 1978, Nickelsburg 1985, Bozdogan 1987, 1990). The key explanatory variables in the final models include roadway locality type, intersection influence, secondary task duration, presence of construction zone, and errors/violations. For assessing the statistical significance of the variables, 95% confidence criterion was used for variables in either model, except for locality variables which were left in the model as per 90% confidence criteria given their conceptual importance. The key results obtained through driver error model, event outcome model, and path analysis are briefly discussed in the subsequent subsections.

Driver Error Model

For the sake of completeness, all the explanatory variables were kept in the functions of all driving errors (Table 3.3). However, for the sake of simplicity and understanding, the marginal effects are provided for only those variables which were found to have a significant association with driving errors and violations (Table 3.4). The final driver error model includes the key variables of interest (roadway environment) along with other important variables like intersection influence, construction zone indicator, and secondary task duration. The decision of including additional but most relevant variables (intersection influence, construction zone indicator, and secondary task duration) was based on examining the likelihood ratio test, AIC, and BIC statistics. After including all the four key variables in the driver error model, the values of AIC, BIC, and log-likelihood at convergence were found to be 14789.04, 15434.23, and -7304.5218 respectively. If the "secondary task duration" variable is excluded from the final model, AIC, BIC, CAIC, SBIC, and log-likelihood increased and were found to be 15567.87, 16170.05, and -7699.937 respectively. Note that including additional control and interesting variables did not only improve overall model fit (AIC, BIC) but also shows the statistical significance and intuitive parameter signs. Referring to the roadway environment variable included in the final model, it should be noted that interstate/bypass/divided highway with no traffic signals was used as the base category to explore the associations between all locality-related factors and driving errors and violations. According to the driver error model, secondary task duration was positively correlated with recognition, decision, performance, and experience errors. The marginal effects for all significant variables are presented in Table 3.4. To check for potential multi-collinearity among explanatory variables included in the final model, we computed variance inflation factor (VIF) for each of the explanatory variables. The VIF of all the variables was lower than 2 suggesting no significant evidence of multi-collinearity. The findings (association) in terms of marginal effects of the driver error model are presented in Table 3.4 and briefly discussed in the discussion section of this paper.

Table 3. 3 Driver Error Model Results (Multinomial Logit Model)

Independent Variables	Recognition Err.		Decision Err.		Performance Err.		Violation		Physical Condition		Experience Err.	
	Coeff.	t-stats	Coeff.	t-stats	Coeff.	t-stats	Coeff.	t-stats	Coeff.	t-stats	Coeff.	t-stats
Roadway Locality (Base Category = Interstate)												
Open Country or Open Residential	-0.3225	-1.25	0.5503	3.00	1.6718	1.44	-0.0907	-0.54	0.0456	0.12	0.4386	0.36
Moderate Residential	0.3115	1.80	0.3627	2.34	2.5362	2.45	-0.1947	-1.48	0.4153	1.52	1.1561	1.40
School	0.8722	4.11	0.6745	3.22	1.5164	1.22	-0.3544	-1.64	0.2932	0.63	1.6041	1.59
Business/Industrial	0.4884	3.06	0.3626	2.50	2.0974	2.03	-0.5115	-3.97	0.0222	0.08	1.1802	1.52
Urban	1.3698	5.77	1.1656	4.92	3.1258	2.79	-0.2771	-1.00	0.9352	1.84	2.2329	2.25
Bypass or Divided Highway with traffic signals	0.1797	0.60	0.4141	1.49	2.8566	2.46	-0.1889	-0.74	0.8554	1.97	-	-
Others (e.g., Church, Playground)	0.8005	2.65	0.6259	2.04	2.8218	2.40	-0.3254	-0.98	-0.4948	-0.48	1.7999	1.44
Intersection Influence (Base Category = No Intersection Influence)												
Interchange	1.3505	5.71	1.2575	5.93	-12.2593	-0.01	-0.1622	-0.58	0.3188	0.66	2.4834	3.02
Stop sign or traffic signal	1.3089	11.75	0.3759	2.97	0.7947	2.25	1.0765	9.61	-0.5240	-1.59	0.5408	0.98
Uncontrolled intersection	1.6459	8.96	1.2141	6.49	2.0881	5.18	1.0002	4.66	-0.4838	-0.67	2.2721	4.12
Parking lot or driveways	2.0741	12.86	1.9886	13.72	2.8664	8.75	1.3485	7.07	-0.1514	-0.25	0.6334	0.60
Other (e.g., crosswalk, roundabouts)	2.3620	8.65	1.4822	4.86	1.3483	1.30	1.3335	3.90	-0.0361	-0.04	-	-
Construction zone indicator	0.8073	4.22	0.2459	1.08	-0.3186	-0.31	0.7905	4.48	0.6108	1.52	1.7906	3.15
Secondary Task Duration	0.3407	25.12	0.0678	4.20	0.0853	1.95	-0.0015	-0.09	-0.2763	-4.69	0.1448	2.23
Constant	-4.4577	-30.50	-3.3088	-26.54	-7.5942	-7.54	-2.5622	-26.63	-3.9103	-18.59	-7.5490	-10.22
Summary Statistics (Number of Observations = 9,593)												
Pseudo-R ²	0.1121											
AIC	14789.0400											
BIC	15434.2300											
SBIC	15379.22											
CAIC	15463.22											

Note: For the sake of completeness, we present the estimation results for the driver error model (multinomial logit model) including all variables. However, in Table 3.4, we provide show only significant variables along with their marginal effects. The VIF value each variable included in the driver error model are shown in parenthesis with the corresponding variables: open country or open residential (1.31), moderate residential (1.67), school (1.23), business/industrial (1.96), Urban (1.16), bypass or divided highway with traffic signals (1.11), others (e.g., church, playground, and campground) (1.09), interchange influence (1.09), Stop-sign or traffic signal influence (1.15), uncontrolled intersection influence (1.04), parking lot or driveways influence (1.06), Other (e.g., crosswalk, railroad crossing, roundabouts) (1.02), construction zone indicator (1.01), and secondary task duration (1.01).

Table 3. 4 Driver Error Model Results (Multinomial Logit Model): Marginal Effects for Only Significant Variables

Independent Variables	Type of Driving Error and Violations											
	Recognition		Decision		Performance		Violation		Physical Condition		Experience/ Exposure	
	β_{11}	ME	β_{12}	ME	β_{13}	ME	β_{14}	ME	β_{15}	ME	β_{16}	ME
Roadway Locality (Base Category = Interstate)												
Open Country/Open Residential	---	---	0.5503 ^a	0.0328	---	---	---	---	---	---	---	---
Moderate Residential	0.3115 ^b	0.0130	0.3627 ^a	0.0168	2.5362 ^a	0.0094	---	---	---	---	---	---
School	0.8722 ^a	0.0478	0.6745 ^a	0.0345	---	---	-0.3544 ^b	-0.0324	---	---	---	---
Business/Industrial	0.4884 ^a	0.0251	0.3626 ^a	0.0177	2.0974 ^a	0.0058	-0.5115 ^a	-0.0369	---	---	---	---
Urban	1.3698 ^a	0.0800	1.1656 ^a	0.0665	3.1258 ^a	0.0133	---	---	0.9352 ^b	0.0127	2.2329 ^a	0.0065
Bypass or Divided Highway with traffic signals	---	---	---	---	2.8566 ^a	0.0132	---	---	0.8554 ^a	0.0138	---	---
Others (e.g., church, playground)	0.8005 ^a	0.0417	0.6259 ^a	0.0309	2.8218 ^a	0.0117	---	---	---	---	---	---
Intersection Influence (Base Category = No Intersection Influence)												
Interchange	1.3505 ^a	0.0725	1.2575 ^a	0.0893	---	---	---	---	---	---	2.4834 ^a	0.0152
Stop sign or traffic signal	1.3089 ^a	0.0722	0.3759 ^a	0.0086	0.7947 ^a	0.0030	1.0765 ^a	0.0730	---	---	---	---
Uncontrolled intersection	1.6459 ^a	0.0917	1.2141 ^a	0.0686	2.0881 ^a	0.0176	1.0002 ^a	0.0504	---	---	2.2721 ^a	0.0103
Parking lot or driveways	2.0741 ^a	0.1138	1.9886 ^a	0.1523	2.8664 ^a	0.0324	1.3485 ^a	0.0633	---	---	---	---
Other (e.g., crosswalk, roundabouts)	2.3620 ^a	0.1716	1.4822 ^a	0.0805	---	---	1.3335 ^a	0.0697	---	---	---	---
Construction zone indicator	0.8073 ^a	0.0387	---	---	---	---	0.7905 ^a	0.0419	---	---	1.7906 ^a	0.0043
Secondary Task Duration	0.3407 ^a	0.0184	0.0678 ^a	0.0019	0.0853 ^b	0.0002	---	---	-0.2763 ^a	-0.0037	0.1448 ^a	0.0003

Notes: ME stands for marginal effects; Base outcome in the multinomial logit model is “No error”; and β_{11} , β_{12} , β_{13} , β_{14} , β_{15} , and β_{16} indicates the association of the explanatory variables with recognition, decision, performance, violation, physical condition related, and experience/exposure errors. The superscripts i.e. “a” and “b” indicate that 95% and 90% confidence criteria were met respectively. (-) indicates not applicable as all these variables were tried in the model but did not meet the significance criteria (hence, for simplicity kept out of model). For a model summary and overall goodness-of-fit, please refer to Table 3.3.

Crash Outcome Model

Based on likelihood ratio tests, Model 2C is selected as the best model based on AIC and BIC values (Table 3.5). The significant value of chi-square statistics (Prob. > Chi2 = 0.000), obtained through comparing model 2A versus model 2C, indicate significant evidence that proportional odds assumption is violated by at least one of the explanatory variables in model 2A (Table 3.5). Note that model 2C imposes proportional odds assumption only on those variables which meet the assumptions and can be thus considered as a more appropriate and parsimonious model. The five variables in model 2C, including indicators for decision error, physical conditions, experience error, urban locality, and other localities (church, playground, and campground) do not violate the assumption; hence, the assumption was imposed on these variables (Table 3.6). The value of Chi2 statistics (Prob. > Chi2 = 0.1098) for Model 2C is found to be insignificant (8.98 “Prob. > Chi2 = 0.1098”) which shows that the model does not violate the proportional odds/parallel line assumption (Table 3.6). Having said this, the parameter estimates of these five variables remain the same across the two categories (baseline versus near-crash + crash, and baseline + near-crash versus crash) (Table 3.6). Model 2C outperforms its counterparts; therefore, the results of Model 2C are used in the path analysis. Note that a positive parameter estimate for a specific explanatory variable suggests that the chance of getting a higher value (near-crash + crash for the first utility function and crash for the second utility function) increases. The results in the best-fit model (Model 2C) indicate that compared to no driving error, all five driving errors and violations have positive coefficients which suggest that the chance of crash increases with all five types of driving errors and violations. Note that variables that meet the proportional odds assumption (in Model 2C) have a similar coefficient in the two utility functions (Table 3.6). The marginal effects are computed for all the explanatory variables which show statistically significant association with crash outcome (in the second utility function) (Table 3.6).

Finally, we computed VIF for each explanatory variable or category of the explanatory variable included in the crash outcome model (Model 2C) to see if there could be any potential multi-collinearity issue among the explanatory variables. However, no significant evidence of multi-collinearity among the explanatory variables was noticed – the VIF values associated with all the explanatory variables were found to be lower than 2.

Path Analysis

Finally, Table 3.7 below shows the results from path analysis. Path analysis uncovers direct and indirect relationships between key roadway locality and environment factors, errors and violations, and crash propensity. Direct effects of explanatory factors (including errors and violations) on the crash outcome (obtained from model 2C) are presented in Table 3.6. Next, the effects on human error of all significant

Table 3. 5 Ordered Probit Models with Different levels of Proportional Odds Assumptions

Model	N	DF	AIC	BIC	Prob. > Chi2	SBIC	CAIC	ICOMP
Model 2A (proportional odds)	9,593	22	7677.17	7834.88	---	7816.55	7836.5	7652.604
Model 2B (non- proportional odds)	9,593	42	7146.74	7447.83	0.000	7429.499	7469.499	7101.994
Model 2C (partially proportional odds)	9,593	37	7144.60	7409.85	0.000	7391.515	7426.515	7027.487

Note: N indicates the number of observations, DF indicates degrees of freedom, LR indicates the likelihood ratio test, and prob. indicates probability. The values of Prob. > Chi2 for model 2B and model 2C is obtained while comparing model 2B and model 2C with model 2A respectively – which suggests that in both the cases, proportional odds assumptions are violated in model 2A indicating the use of model 2B and model 2C. Based on AIC, BIC, CAIC, SBIC, and ICOMP, the Model 2C outperforms.

Table 3. 6 Crash Outcome Model Results (Ordered Probit Models: Accounting for Proportional Odds Assumptions)

Independent variables	Model 2A (Proportional Odds)		Model 2B (Non-Proportional Odds)				Model 2C (Partially Proportional Odds)						
			baseline vs. near-crash + crash		baseline + near-crash vs. crash		baseline vs. near-crash + crash		baseline + near-crash vs. crash		*Prop. Odds Assump.	ME for Crash (Model 2C)	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat			
Drivers Errors (Base Category = No Driving Errors)													
Recognition Errors	2.1886	39.63	3.2194	27.17	1.9113	24.72	3.2284	27.26	1.8498	26.59	Relaxed	0.2695	
Decision Errors	1.9964	36.97	1.9188	31.88	2.0294	27.4	1.9497	35.01	1.9497	35.01	Yes	0.3026	
Performance Errors	2.9181	17.83	2.0908	9.42	3.0275	16.71	2.0925	9.42	2.9714	16.68	Relaxed	0.6796	
Violations	0.9973	17.70	0.9260	15.51	1.1403	13.17	0.9315	15.62	1.0800	13.58	Relaxed	0.0832	
Physical Conditions	0.9153	7.30	0.8409	6.27	1.1368	6.49	0.8962	6.91	0.8962	6.91	Yes	0.0573	
Experience or Exposure	2.286	10.10	1.9315	7.05	2.3723	9.58	2.1864	9.46	2.1864	9.46	Yes	0.3871	
Roadway Locality (Base Category = Interstate)													
Open Country or Open Residential	0.0505 ^a	0.64	-0.0137 ^a	-0.16	0.3466	2.60	-0.0109 ^a	-0.13	0.3230	2.65	Relaxed	0.0212	
Moderate Residential	-0.0045 ^a	-0.07	-0.1499	-2.22	0.4151	3.80	-0.1472	-2.19	0.3908	4.15	Relaxed	0.0265	
School	0.2534	3.06	0.1386 ^a	1.48	0.6186	4.62	0.1456 ^a	1.56	0.5806	4.78	Relaxed	0.0434	
Business/Industrial	0.1914	3.39	0.0689 ^a	1.12	0.5768	5.50	0.0782 ^a	1.28	0.5335	6.05	Relaxed	0.0389	
Urban	0.6913	7.48	0.7780	7.05	0.8134	5.65	0.7813	8.12	0.7813	8.12	Yes	0.0644	
Bypass or Divided Highway with traffic signals	0.1832 ^b	1.70	0.1045 ^a	0.88	0.5335	2.95	0.1087 ^a	0.92	0.4980	2.88	Relaxed	0.0357	
Others (e.g., church, playground)	0.2357 ^b	1.92	0.2327 ^a	1.61	0.4704	2.45	0.3110	2.46	0.3110	2.46	Yes	0.0203	
Intersection (Base Category = No intersection influence)													
Interchange	0.8957	10.71	1.1251	11.97	0.2756 ^b	1.77	1.1295	12.00	0.2572 ^b	1.72	Relaxed	0.0221	
Stop sign or traffic signal	0.6471	14.40	0.8411	16.31	0.1075 ^a	1.56	0.8422	16.39	0.1071 ^a	1.58	Relaxed	---	
Uncontrolled intersection	1.0042	13.72	1.3851	15.59	0.2553	2.35	1.3861	15.61	0.2547	2.37	Relaxed	0.0219	
Parking lot or driveways	1.0828	16.54	1.3569	16.49	0.5598	6.49	1.3607	16.58	0.5645	6.62	Relaxed	0.0559	
Others intersection	0.8342	7.01	1.4449	8.93	-0.0901 ^a	-0.49	1.4480	8.95	-0.0860 ^a	-0.47	Relaxed	---	
Construction zone indicator	0.1409 ^b	1.74	0.2447	2.66	-0.0461 ^a	-0.36	0.2459	2.67	-0.0428 ^a	-0.34	Relaxed	---	
Secondary task duration	0.0269	4.66	0.0397	5.45	0.0192	2.53	0.0386	5.31	0.0203	2.70	Relaxed	0.0016	
Constant	-3.1488	-53.39	-1.9461	-37.04	-3.0164	-29.21	-1.9565	-37.68	-2.9242	-35.22	---	---	

Note: ME indicates marginal effects (computed for Model 2C) which predicts the probability of observing a response outcome (Crash) with a unit increase in a continuous explanatory variable (or a switch from 0 to 1 for indicator variable). The base level for drivers' errors is "No error". (a) and (b) indicate that these variables show "no statistical significance" and "partial statistical significance (meeting 90% confidence level criteria)", respectively. In the final path analysis framework, only the variables which showed statistical significance or partial significance were considered. The "gologit2" shows that five variables including indicators for decision error, physical conditions, experience error, urban locality, and other localities (church, playground, and campground) in "Model 2C" did not violate the parallel line (pl) assumption and therefore the assumption was imposed on these variables. All the remaining variables in Model 2C violate the pl assumption and therefore "gologit2" relaxes the assumption for them. Overall, the Chi² statistics for Model 2C is found to be 8.98 (Prob. > Chi2 = 0.1098) – which is insignificant, suggesting that Model 2C does not violate the proportional odds/parallel line assumption (Williams 2005, 2016).

Table 3. 7 Direct, Indirect, and Total Effects of Explanatory Variables (Path Analysis)

Independent Variables	Direct Effect on Crash (τ_1) (%)	Effects on Driving Errors (%)						Indirect Effects on Crash (%)						Total Effect on Crash (%)					
		τ_{21}	τ_{22}	τ_{23}	τ_{24}	τ_{25}	τ_{26}	$\tau_{21} * \tau_{31}$	$\tau_{22} * \tau_{32}$	$\tau_{23} * \tau_{33}$	$\tau_{24} * \tau_{34}$	$\tau_{25} * \tau_{35}$	$\tau_{26} * \tau_{36}$	A	B	C	D	E	F
Driver Errors (Base Category = No driving errors)																			
Recognition Err. (τ_{31})	26.95	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Decision Err. ((τ_{32})	30.26	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Performance Err. (τ_{33})	67.96	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Violations (τ_{34})	8.32	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Physical Condition Err. (τ_{35})	5.73	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Experience/Exposure Err. (τ_{36})	38.71	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
Roadway Locality and Environment (Base Category = Interstate)																			
Open Country or Open Residential	2.12	---	3.28	---	---	---	---	---	0.99	---	---	---	---	2.12	3.11	2.12	2.12	2.12	2.12
Moderate Residential	2.65	1.30	1.68	0.94	---	---	---	0.35	0.51	0.64	---	---	---	3.00	3.16	3.29	2.65	2.65	2.65
School	4.34	4.78	3.45	---	-3.24	---	---	1.29	1.04	---	-0.27	---	---	5.63	5.38	4.34	4.07	4.34	4.34
Business/Industrial	3.89	2.51	1.77	0.58	-3.69	---	---	0.68	0.54	0.39	-0.31	---	---	4.57	4.43	4.28	3.58	3.89	3.89
Urban	6.44	8.00	6.65	1.33	---	1.27	0.65	2.16	2.01	0.90	---	0.07	0.25	8.60	8.45	7.34	6.44	6.51	6.69
Bypass/Divided Highway with traffic signals	3.57	---	---	1.32	---	1.38	---	---	---	0.90	---	0.08	---	3.57	3.57	4.47	3.57	3.65	3.57
Others (church, playground)	2.03	4.17	3.09	1.17	---	---	---	1.12	0.94	0.80	---	---	---	3.15	2.97	2.83	2.03	2.03	2.03
Intersection Influence (Base Category = No Intersection Influence)																			
Interchange	2.21	7.25	8.93	---	---	---	1.52	1.95	2.70	---	---	---	0.59	4.16	4.91	2.21	2.21	2.21	2.80
Stop sign or traffic signal	---	7.22	0.86	0.30	7.30	---	---	1.95	0.26	0.20	0.61	---	---	1.95	0.26	0.20	0.61	---	---
Uncontrolled intersection	2.19	9.17	6.86	1.76	5.04	---	1.03	2.47	2.08	1.20	0.42	---	0.40	4.66	4.27	3.39	2.61	2.19	2.59
Parking lot or driveways	5.59	11.38	15.2 3	3.24	6.33	---	---	3.07	4.61	2.20	0.53	---	---	8.66	10.2 0	7.79	6.12	5.59	5.59
Others intersection	---	17.16	8.05	---	6.97	---	---	4.62	2.44	---	0.58	---	---	4.62	2.44	---	0.58	---	---
Construction zone indicator	---	3.87	---	---	4.19	---	0.43	1.04	---	---	0.35	---	0.17	1.04	---	---	0.35	---	0.17
Secondary Task Duration	0.16	1.84	0.19	0.02	---	- 0.37	0.03	0.50	0.06	0.01	---	- 0.02	0.01	0.66	0.22	0.17	0.16	0.14	0.17

Note: The calculations in Table 3.7 are based on chance (Marginal Effects multiplied by 100%); therefore, we use τ symbol to indicate the chance instead of using “ β ” which indicates parameter estimate (coefficients) for an explanatory variable. τ_1 is the direct effect of explanatory variables on the crash outcome (Model 2C); τ_{21} through τ_{26} indicates the effects (obtained from Multinomial logit model) of explanatory factors on recognition, decision, performance, violations, physical condition related, and experience/exposure errors respectively; whereas A through F indicates the total effects corresponding to each of the explanatory factors, i.e., $A = \tau_1 + \tau_{21} * \tau_{31}$; $B = \tau_1 + \tau_{22} * \tau_{32}$; $C = \tau_1 + \tau_{23} * \tau_{33}$; $D = \tau_1 + \tau_{24} * \tau_{34}$; $E = \tau_1 + \tau_{25} * \tau_{35}$; $F = \tau_1 + \tau_{26} * \tau_{36}$; (---) indicates not applicable.

explanatory factors (including roadway localities and environments, intersection influence, presence of construction zones, and secondary task duration) obtained from the multinomial logit model are presented in Table 3.4. As a final step, the indirect effects of key explanatory factors on crash outcome are shown in Table 3.7. That is, the indirect effect of specific independent variables on the crash outcome through a particular error type was calculated by multiplying the marginal effect of the specific error type, β_{1i} , with the marginal effect of that specific driver error on crash outcome γ_i . Finally, the direct and indirect effects corresponding to each of the explanatory variables are reported (Table 3.7). The direct effects of various roadway environments (localities) are illustrated in Figure 3.2. The effects of driving errors and violations on a crash outcome which are used in computing the indirect effects of roadway and environments on crash outcomes through mediating driving errors and violations are shown (Figure 3.2). Also, the effects of diverse roadways and environments on various driving errors and violations are illustrated in Figure 3.2.

In short, the variables that possess statistically significant parameter estimates in the driver error model are indirectly associated with crash outcome (Table 3.6), a finding that cannot be extracted using the traditional modeling approach. Moreover, all variables which showed a significant direct correlation with crash outcome were also found to have a significant indirect correlation with crash outcome through one or more of the six mediating driving errors and violations (Table 3.7). If the complex path structure is ignored, valuable insights regarding indirect effects could not be obtained. Note that the standard path analytic framework used in this study assumes the two error terms to be uncorrelated; however, there may be some common unobserved factors that may significantly affect both driving errors and event outcome even after conditioning on observed explanatory variables. If such unobserved factors are present, a joint estimation (such as by tying the system of equations through multivariate normal distribution) could lead to more efficient parameter estimates. After doing cross-tabulation of the two observed discrete outcomes, we noticed a strong positive correlation (Goodman and Kruskal's gamma of 0.76 on scale -1 to 1) between observed driving errors and event outcomes suggesting a potential correlation between the residuals associated with the two equations. While joint estimation may help in efficiency gains, the coefficients obtained from the standard path analysis framework (with independent error terms) are still consistent and can be used to explore the key pathways (Liu and Khattak 2018, Arvin *et al.* 2019). Methodologically, the rationale behind the joint estimation is that of potential 'efficiency gains' – and the extent of efficiency depends on the appropriateness of the multivariate normality assumption. This assumption can be severely problematic if the joint distribution of the unobserved factors does not follow multivariate normality. If that is the case (and it typically is most of the time in practice), then the resulting estimates from the joint system are not even consistent. In other words, consistency of parameter estimates is lost to gain efficiency at times when the consistency of parameter estimates always supersede the efficiency. In such cases, more rigorous copula

Association strength (Marginal Effects in %): — 0-4.99% — 5-9.99% — 10-19.99% — 20-29.99% — 30-39.99% — 40-49.99% — 50-59.99% — 60-70%

Note: All the pathways (including direct and indirect) are based on marginal effects (%) from the two constituent models. The red and green color pathway shows increase and decrease in the probability of outcome to which it leads, respectively.



Figure 3. 2 Direct and Indirect (through Driving Errors and Violations) Effects of Roadways and Environments Crash Occurrence

Note: For detailed understanding about the strength (width of the arrows) of the pathways either zoom in Figure 3.2 or refer to Table 3.7.

based methods may be warranted and could be used in the future (Bhat and Eluru 2009, Eluru *et al.* 2010, Wali *et al.* 2018b, Wali *et al.* 2018e).

Discussion

The path analysis framework was used in this study to quantify the indirect effects of all associated factors on crash propensity which are discussed next. The driver error model quantifies the significant increases in the probability of recognition errors and decision errors within the following localities: moderate residential, school zones, business/industrial areas, urban areas, at interchanges, and stop signs or signalized intersections. Several important and new insights follow. The chance of recognition error, compared to no driving error or violation, by a driver increases by 1.30% on roads passing through moderate residential settings as compared to interstates. This is reasonable as interstate/bypass/divided highways with no traffic signals, by definition, is signal-free and high-speed limited-access facilities while moderate residential areas indicate locations with moderate housing density which can result in increased recognition error as compared to interstate (Hankey *et al.* 2016). Similarly, the chance of recognition errors (compared to no driving error or violation) increases by 4.78% in school zones as compared to the interstate. This finding is reasonable in the sense that school zones typically include complex movements and vulnerable road users compared to interstates/bypass/divided highways with no traffic signals. As a result, the chances of a driver failing to recognize the hazardous situation increases (Gregory *et al.* 2014). Furthermore, the estimation results of the driver error model indicate that compared to interstate/bypass/divided highways with no traffic signals, the chance of recognition and decision errors (compared to no driving error or violation) on roadways in urban areas increase by 8.00% and 6.65% respectively which is intuitive due to complex and highly densified traffic conditions (Huang *et al.* 2010). Referring to business/industrial areas, modeling results reveal that the chance of making recognition and decision errors, compared to no driving error or violation, increase by 2.51% and 1.77% respectively on roadways passing through business or industrial areas as compared to interstates. The higher traffic volume, complex activities, and increased presence of commercial drivers can negatively affect drivers' recognition capabilities. According to the driver error model results, compared to no driving error or violation, drivers are more likely to make performance errors on roadways passing through moderate residential areas, business/industrial areas, urban areas, and on bypass/divided highways with traffic signals as compared to the interstate. Note that the estimation results of the driver error model reveal that chance of violation, compared to no driving error/violation, is lower on roadways within school and business/industrial areas as compared to interstate (potentially highlighting higher compliance of drivers in school zones (Wali and Khattak 2020)).

The estimation results of the driver error model show that compared to no driving error/violation the probability of recognition errors, decision errors, performance errors, violations, and experience or exposure-related errors increases at uncontrolled intersections compared to no intersection influence. This finding is reasonable as drivers face relatively more complexity in appropriately recognizing, deciding, and performing within such uncontrolled and irregular traffic flow situations. Also, the non-existence of traffic-controlled devices or traffic wardens may encourage some drivers to violate at such locations. The modeling results indicate that, compared to no intersection influence, the chance of recognition and decision errors increase at interchanges by 7.25% and 8.93% respectively. This is reasonable, as drivers at such diverging areas are required to focus on several driving tasks including recognition, decision, lateral and longitudinal control which can increase their vulnerability to the two error types. According to the modeling results, compared to no driving error/violation, the chances of recognition errors increase by 7.22% at stop signs or signalized intersections. The model estimation results show that the chance of recognition errors, violations, and experience-related errors increase in a construction zone. Additionally, modeling results indicate that chance of recognition errors, decision errors, performance errors, and experience errors increases with an increase in secondary task durations. This finding was expected as involvement in distractions or secondary tasks can induce driving errors and negatively affect driving capabilities.

The crash outcome model (model 2C) quantifies the association of driving errors and other important factors with the crash propensity (crash risk). The model results indicate that all driving errors, compared to no driving errors, are positively correlated with crash propensity. However, performance errors have a strong correlation with a crash propensity, associated with a higher chance of crashes by 67.96%. Similarly, the chance of crash outcome is 8.32% higher if a driver violates traffic laws. Also, the chance of crash (6.44%) is higher in an urban environment which can be due to highly complex and densified traffic conditions. This finding is consistent with previous studies based on police crash reports (Huang *et al.* 2010).

Interestingly, the modeling results of crash outcome indicate that compared to interstates, the chance of a crash increases 4.34% and 3.89% on roadways passing through school and business/industrial locations respectively. These findings are reasonable because of the presence of vulnerable road users and potential distractions in such areas as compared to interstates (Wali and Khattak 2020). Similarly, compared to no-intersection influence, both controlled and un-controlled intersections are positively correlated with crash propensity. The results of the crash outcome model show that the chance of a crash is higher in construction zones as compared to areas with no construction activities. Secondary task duration is also found to be positively correlated with a crash propensity as per the crash outcome model.

As discussed in the methodology section, marginal effects from the constituent models were used to calculate the indirect effects of key correlates on crash propensity. For instance, the urban indicator variable

shows that compared to drivers traveling on the interstate, a driver traveling in an urban locality has an 8.00% higher chance of making recognition errors (compared to no driving error or violation), and as a direct effect, compared to no driving error or violations, drivers involved in recognition errors are associated with a 26.95% higher chance of getting into a crash. As such, the indirect effect of urban locality, compared to interstate, on a crash outcome is $8.00\% \times 26.95\% = 2.16\%$. Therefore, the total effect of urban locality, compared to interstate, on crash outcome becomes 8.60% ($6.44\% + 2.16\%$), which is greater than the direct effect of 6.44% (Table 3.7). This implies that while urban locality, compared to interstate, increases the chance of observing a crash outcome, the actual increase is even more when we simultaneously consider the effect of urban locality on the probability of recognition errors and the effect of the latter on the probability of a crash outcome. The path analysis results indicate that almost all key correlates were both directly and indirectly associated with crash propensity through mediating recognition, decision, or both recognition and decision errors. For instance, the path analysis results indicate that, compared to interstate, the presence of a school zone increases the chance of recognition and decision errors (compared to no driving error or violation) by 4.78% and 3.45% respectively, while compared to no driving error/violation, recognition and decision errors increase the chance of a crash by 26.95% and 30.26% respectively (Table 3.7). Compared to interstate, the presence of a school zone directly increases the chance of a crash by 4.34%. However, path analysis indicates that the overall effect of school zones (compared to interstate) on crash propensity is even larger; the increased chance of crash occurrence through recognition and decision errors (compared to no driving error or violation) is 5.63% and 5.38% respectively (Table 3.7). Construction zones show a direct significant correlation with crash propensity through recognition errors, violations, and experience errors. Similarly, secondary task duration also indirectly correlates all driving errors except violations (Table 3.7). Likewise, almost all roadway localities were not found only in direct correlation with crash propensity but were also indirectly associated with crash risk through either recognition errors, decision errors, or both (Table 3.7). Overall, a rigorous path analysis framework uncovers important information about driving errors that moves us closer to more targeted interventions in complex environments (Liu *et al.* 2015, Zhang *et al.* 2018).

To sum it up, this study discovered some important findings which were previously unexplored. For instance, previous studies have looked at the effects of individual driving behaviors (e.g., aggressive driving, secondary task duration, cell phone use, and unfamiliarity with roadway) on crash risk (Wali and Khattak 2020) (Wali *et al.* 2019) and crash severity (Arvin *et al.* 2019) using SHRP2-NDS data; however, none of the studies (to the best of our knowledge) provides a holistic picture of how key driving errors and violations (obtained through systematic taxonomy of driving behaviors) contribute to crashes in naturalistic environment. More importantly, this study explored and quantified key pathways which lead from diverse roadway environments (e.g., school zones, business/industrial, and urban locations) to crashes both directly

and indirectly through one or more naturalistic driving errors (including recognition errors, and decision errors) and/or violations which have not been explored earlier. Using a safe system approach, this study considered the problem in a broader sagacity to see the links of roadway environment with crashes through driving errors and violations. From a practical implication standpoint, the findings generate new knowledge and foundation which could be used to identify more effective countermeasures to reduce crash risk. For instance, based on the direct pathways leading from diverse roadway environments to crashes, traditional countermeasures (roadway “built-environment” changes) could be selected. However, after understanding the significant mediating role of driving errors and violations (as evident through this study), we consider that multiple vehicle technologies (including collision warning system, adaptive cruise control, and dilemma zone mitigation system) could also reduce crash risk by eliminating or reducing a significant percentage of one or more driving errors and violations. To conclude, findings from this study highlight the significance of the safe system approach and path analysis to understand and solve the safety problem in a broader context. For a detailed discussion about the potential countermeasures, please refer to the conclusion section of this study.

Limitations and Future Directions

This study analyzed the NDS data containing extensive information on driver performance and behavior, roadway environment factors, and pre-crash maneuvers, still there could be other factors (e.g., drivers’ characteristics, and local demographics) which might be significantly correlated with crash propensity. Hence, it would be interesting to account for all such unobserved factors in the modeling framework, which could have a significant influence on crash propensity. The specific number of drivers and geographical regions captured in this study may not represent all of the socio-geographical regions across the U.S. Similarly, TDEV—the human error taxonomy uses driver behaviors while classifying driving errors and violations which may vary with time.

To categorize roadway and land-use factors, a more holistic approach would be to apply appropriate roadway-locality related classifications, e.g., roadway functional classes (interstate, expressways, arterials (major/minor), collectors (major/minor), and local roads), activity centers (specifically separately categorizing business from industrial), and locality (urban, suburban, and rural). The information available from a single variable called “locality” in the NDS data only partially fulfills the research needs. In the future, it will be appropriate to either collect separate and complete information on roadway functional classes (interstate, expressways, arterials (major/minor), collectors (major/minor), and local roads), activity centers (specifically separately categorizing business from industrial), and locality (urban, suburban, and rural) or provide geo-codes so that such information can be obtained. Such efforts can help with an even deeper understanding of the variations in driving errors and violations and crash risks across contexts. In

the future, it would be important to apply the proposed TDEV (human error taxonomy) framework to other vulnerable locations such as school zones (Wali and Khattak 2020). Merging the theoretical richness of the TDEV framework with kinematic measures of driving behavior can provide deeper insights at vulnerable locations (Wali and Khattak 2020). Also, information on important variables like driver demographics (e.g., driver age, gender, driving experience, and education) is not available in the SHRP2-NDS subsample used in this research. In the future, with more data, it would be interesting to explore how driver demographics relate to driving errors and violations, and crash risk.

Conclusions

This research contributes to the literature by conceptualizing the direct relationships between key roadway and environmental (roadway locality) factors and crash propensity, as well as the indirect relationships between the roadway and environmental factors and crash propensity through mediating errors and violations. To achieve these objectives, the study capitalizes on emerging sensor-based technologies, telematics, video, and radar surveillance to obtain an objective and nuanced understanding of key crash contributing factors. Along these lines, the study harnesses data from the Naturalistic Driving Study where driving behaviors in thousands of baseline, crash, and near-crash events in diverse Spatio-temporal contexts are analyzed and compared. In addition to considering crash events, no-event driving is analyzed vis-à-vis near-crash outcomes, as such “close calls” may foretell actual future crashes.

A taxonomy based on the perception-reaction to hazardous situations is used to classify driving errors and violations and analyzed them. Next, this study uses information from the “locality” factor in the NDS data to quantify the direct as well as indirect effects of the different roadway and environmental variables on a crash outcome through different types of driving errors and violations.

The results of the driver error model indicate that several roadway environmental factors are statistically significantly associated with driver errors. The most prevalent types of errors are recognition errors and decision errors, which are correlated with moderate residential, school zones, urban areas, business or industrial areas, at interchanges stop signs, and signalized intersections. Modeling results indicate that the duration of secondary tasks is correlated with recognition errors, decision errors, performance errors, and violations.

To model roadway safety outcomes, a discrete outcome model was estimated. The model examines crash propensity as a function of mediating error types/violations and roadway-environmental contexts. Some of the explanatory variables in the crash outcome model violate the proportional odds assumption due to which we develop a more sophisticated (partially proportional odds) model which relaxes the assumption only for those explanatory variables which violate it. Note that the partially proportional odds model showed

superior fit to the data and its results were used in computing direct and indirect effects of key explanatory variables in the final path analysis. Importantly, while urban locality associates with higher chances of observing a crash by 6.44%, path analysis showed that they can induce recognition errors and the effect of the latter on the chance of crash outcome, i.e., the “total effect” of urban locality amounts to 8.60% on the chances of having a crash through recognition errors. Additionally, school zones and business (commercial) or industrial land uses were found to have a substantial correlation with crash propensity both directly and indirectly through mediating recognition errors and decision errors. By examining the detailed behavior of drivers, this study is perhaps among the first ones to uncover business/commercial areas as hazardous in a substantive way. Indeed, the analysis indicates that complex path structures should be explored in line with the systems approach. This avoids the possibility of concluding that key factors have an only direct association with crash outcomes, while in reality almost all factors (e.g., roadways and environments, intersection influence, presence of construction zone, and secondary task duration) can and were found to have significant indirect correlations with crash outcome through driving errors and violations.

Referring to the practical applications of this research, we consider that multiple vehicle technologies can help address the errors and violations explored in this study. Collision warning systems and the adaptive cruise control can help reduce recognition errors, decision errors, performance errors, and/or physical condition errors. With regards to the physical condition and violation-related errors, dilemma zone mitigation systems and lateral vehicle detection systems can be useful. Likewise, drowsiness monitoring and detection systems can help reduce physical condition-related errors associated with drowsiness. We recommend conducting further research before implementing these technologies as accessing their safety benefits is not straightforward until these technologies are implemented on a large scale. For instance, a fault in these technologies may lead to drastic safety outcomes. For a detailed discussion on the potential role of multiple vehicle technologies in reducing one or more driving errors and violations, please refer to another study (Khattak *et al.* 2021a). Besides, changes in the built environment (roadway designs) can potentially reduce driving errors and violations. Moreover, conducting awareness campaigns or webinars for drivers while highlighting the safety risk associated with distractive driving, being involved in secondary tasks, and drowsiness can help to reduce a significant portion of recognition and physical condition errors. Considering formal training programs for drivers can be beneficial for novice and inexperienced drivers and those frequently committing performance errors. While there is a need for further research on implementing vehicle technologies and traditional countermeasures, we consider that more safety benefits can be achieved if the two countermeasures are systematically implemented along with the policy-related interventions. The study does not advocate for specific government policies but provides information that can be used for future formulation of safety policies and research. The data and methods discussed in the

paper should allow for replication of the study. As a part of future work, it will be interesting to investigate how driver errors and violations may change with some control of the driving task being given to connected and automated vehicles. Similarly, it would be interesting to explore the effects of the roadway environment on injury outcomes (given a crash).

**CHAPTER 4. EXPLORING PATHWAYS FROM DRIVING ERRORS, VIOLATIONS,
AND DURATION OF DISTRACTION TO CRASHES: THE ROLE OF INSTABILITY
IN DRIVING**

The key findings from this chapter have led to two papers one of which was presented at the 100th Annual Meeting of the Transportation Research Board, Washington D.C. (Jan 2021) while the other one is accepted for presentation at the Transportation Research Board 101st Annual Meeting, Washington D.C (Jan 2022). The two papers are also in review with the journal of Accident Analysis and Prevention. For the titles of the two papers, please see below.

Ahmad, N., Arvin, R., & Khattak, A. Exploring Pathways from Driving Errors and Violations to Crashes: The Role of Speed Volatility (Presented in TRB2021 and under review with AAP)

Ahmad, N., Arvin, R., & Khattak, A. How is the duration of distraction-related to safety-critical events? Harnessing naturalistic driving data to explore the role of driving instability (Presented in TRB2021 and under review with AAP)

Abstract

This study explores how different driving errors, violations, duration of distraction, and roadway environments contribute to safety-critical events through speed volatility. We harness a subsample (N = 9,237) of the naturalistic driving study (NDS) data collected through the Second Strategic Highway Research Program (SHRP2). From a methodological standpoint, we use the safe systems approach relying on path analysis to jointly model outcomes. This accounts for the potential correlation between unobserved factors associated with both speed volatility and event outcomes, i.e., baseline or event-free driving, near-crashes, and crashes. Tobit and ordered Probit regressions are estimated to model the coefficient of variation (COV) of speed and event outcomes, respectively. Results from the Tobit model indicate that driving errors, violations, and duration of distraction are positively associated with instability in driving (COV of speed). The Probit model reveals that driving errors, violations, duration of distraction, and instability in driving are associated with higher chances of crashes and near-crashes. Our key finding is that driving errors, violations, and duration of distraction not only induce event risk directly but also indirectly through instability in driving. For instance, recognition errors associate with higher crash risk by 6.54% but this error is accompanied by instability in driving, which further increases event risk by 5.01%, bringing the total increase in risk to 11.55%. Referring to the total effects of distraction duration on SCEs obtained through path analysis (via predictive margins), the chance of both SCEs significantly increases non-linearly with an increase in distraction duration beyond 4 seconds. For instance, the chance of a crash is 16% for a driver distracted for 4 seconds, which increases to 22% if a driver is distracted for 10 seconds. Moreover, significant correlations were found between unobserved factors reflected in the error terms of the two models. Ignoring such correlations can lead to inefficient parameter estimates. Based on the findings,

practical implications are discussed, which can lead to effective countermeasures that effectively reduce crash risk.

Introduction

Driving behaviors (human factors) contribute to more than 90% of crashes (Treat *et al.* 1979, Singh 2015). Given the prevalence of human factors in crashes, there is a dire need to have a deeper understanding of these factors. To understand the prevalence of human factors systematically, studies developed taxonomies of driving behaviors that contribute to crashes (Treat *et al.* 1979, Reason *et al.* 1990, Wierwille *et al.* 2002, Stanton and Salmon 2009). Rather than looking at individual driving behaviors in crashes, the taxonomies provide a more holistic picture while bringing driving behaviors into a systematic framework using different cognitive philosophies (Treat *et al.* 1979, Reason *et al.* 1990, Wierwille *et al.* 2002, Stanton and Salmon 2009). The previous taxonomies were based on police crash reports which could be subjective and biased when it comes to pre-crash driving behavior and performance (Dingus *et al.* 2006, Yamamoto *et al.* 2008, Khattak *et al.* 2021a). On the other hand, the availability of high-frequency naturalistic driving study (NDS) from the Second Strategic Highway Research Program (SHRP2) provides extensive real-world data on driving behavior and performance (Dingus *et al.* 2006, Blatt *et al.* 2015, Carney *et al.* 2015b, Papazikou *et al.* 2017, Papazikou *et al.* 2019, Richard *et al.* 2020). The large-scale SHRP2-NDS data also includes second-to-second information on vehicle kinematics, roadway environments, and traffic conditions which were continuously captured via advanced equipment like video cameras, radars, and ambient sensors fitted in the subject vehicles (Dingus *et al.* 2006, Blatt *et al.* 2015, Carney *et al.* 2015b, Arvin *et al.* 2019b). Also, real-world microscopic data on driving parameters like speed, acceleration/deceleration, and secondary tasks and their durations make SHRP2-NDS data a more unique and realistic data source (Arvin *et al.* 2019b, Wali *et al.* 2019, Wali and Khattak 2020). Note that SHRP2-NDS data includes data on both spatial and temporal variations in the movement of subject vehicles before SCEs which could be used to compute driving volatility (driving instability).

First, this study uses the recently developed taxonomy of driving errors and violations (TDEV) to categorize driving behaviors using objective SHRP2-NDS data (Khattak *et al.* 2021a). Next, the study aims to achieve the following objectives:

- Identify the potential pathways which can lead from errors and violations in diverse roadway environments to SCEs through instability in driving.
- Investigate whether the duration of distraction significantly relates to instability in driving and SCEs and explore if there could non-linear association of duration of distraction with driving instability and SCEs.

- Jointly estimate models that account for potential correlation between the unobserved factors associated with both event outcomes (baseline, near-crash, and crash) and coefficient of variation “COV” of speed (instability in driving).

Literature Review

Driving Errors and Violations, Driving Instability, and SCEs

Studies have proposed systematic classification systems of human factors based on different themes which provide a more inclusive and systematic understanding about their contribution to crashes (Treat *et al.* 1979a, Reason *et al.* 1990, Petridou and Moustaki 2000, Wierwille *et al.* 2002, Stanton and Salmon 2009). For instance, human factors in crashes were classified into recognition errors, decision errors, performance errors, and violations (Treat *et al.* 1979). While previous taxonomies provide useful insights about how human factors contribute to crashes, they were based on police crash reports which could be subjective and biased (Yamamoto *et al.* 2008, Khattak *et al.* 2021a). Moreover, police crash reports use indirect observations because it is unrealistic for a police officer to be present at the crash scene before the crash occurrence (Hankey *et al.* 2016). Besides potential issues of underreporting and bias in police crash data, police crash reports can be prone to underreporting issues such as property damage and minor crashes (Yamamoto *et al.* 2008, Washington *et al.* 2010, Ye and Lord 2011).

To bridge the aforementioned research gaps, one of the recent studies used the comprehensive and objective SHRP2-NDS data to classify driving behaviors that contributed to crashes (Khattak *et al.* 2021a). Using the theoretically-based perception-reaction process, naturalistic driving behaviors were systematically categorized into recognition errors, decision errors, performance errors, physical condition-related errors, experience errors, and violations where the later ones include both intentional and unintentional violations (Khattak *et al.* 2021a). Note that all naturalistic driving behaviors (e.g., distraction due to cell phone browsing, cell phone texting, adjusting/monitoring radio, and looking at objects external to the vehicle) due to which a subject driver failed to appropriately recognize the true situation (which consequently led to a crash) were classified as recognition errors (Khattak *et al.* 2021a). Decision errors and performance errors refer to the failure of subject drivers in the decision and reaction phases of the perception-reaction process, respectively (Khattak *et al.* 2021a). Note that in the performance phase, besides performance errors, a subject driver could be involved in violations (Khattak *et al.* 2021a). The two driving errors of physical condition errors and experience errors, which are not directly relevant to the perception-reaction process, were classified into separate categories (Khattak *et al.* 2021a). For further details about the systematic taxonomy of driving errors and violations (TDEV) and constituent driving errors and violations, please see (Khattak *et al.* 2021a). The TDEV can be considered as an important contribution to the previous

taxonomies (based on police crashes) as driving behaviors, especially sources of distraction (e.g., cell phone use, cell phone, browsing), have significantly changed compared to the past (Khattak *et al.* 2021a). The present study takes an advantage of the TDEV to bring naturalistic driving behaviors into a systematic framework using SHRP2-NDS data which are then used to compute pathways which leads from these driving errors and violations to SCEs both directly and through instability in driving.

Besides pre-crash driving behaviors, studies reveal that driving volatility (instability) significantly influences crash risk (Wali *et al.* 2019, Wali and Khattak 2020) and crash severity (Arvin *et al.* 2019b). Driving volatility can be defined as deviations from the normal values of driving parameters like speed, acceleration or deceleration, and jerk (Arvin *et al.* 2019b, Wali *et al.* 2019, Wali and Khattak 2020). The SHRP2-NDS data includes information on spatiotemporal variations in vehicle parameters like speed, acceleration, deceleration, and jerk, which can be used to derive various volatility indices including standard deviation, coefficients of variation (COV), and quartile coefficient of variation of driving speed, acceleration, and deceleration (Arvin *et al.* 2019b, Wali *et al.* 2019, Wali and Khattak 2020). To understand the relationship of driving volatility indices with driving violations and crash occurrence, studies suggest that driving volatility increases the likelihood of violation and crash (Yeo *et al.* 2020). Referring to the association of volatility and crash severity, studies suggest that driving volatility positively correlates with crash severity (Arvin *et al.* 2019b).

Driving volatility leads to unsafe outcomes. For details about how various volatility indices relate to crash risk, and crash severity, please refer to Table 4.1. It can be seen that one of the previous studies focused on how diverse roadway environments relate to “*crash severity*” through driving volatility (Arvin *et al.* 2019b). However, the underlying pathways from errors and violations in diverse roadway environments leading to “*crash risk*” remain unexplored.

Duration of Distraction, Driving Instability, and SCEs

Due to the increase in road crashes that occur due to driving distractions, numerous research efforts have explored distracted driving (Choudhary and Velaga 2017b, a, Oviedo-Trespalcacios *et al.* 2017, Qin *et al.* 2019, Ahmad *et al.* 2021). Note that distracted driving increases crash risk and leads to valuable life losses. For instance, in the U.S., 3,142 people were killed in road crashes that occurred due to distracted driving in 2019 alone (NHTSA). Due to involvement in distracting activities, a driver usually may take his or her eyes off the roadway or may not be mentally active about the surrounding traffic and roadway environment (Dingus *et al.* 2006c, Harbluk *et al.* 2007), which may not only reduce visual scan (Recarte and Nunes 2003) but also increase the cognitive load (Lee *et al.* 2001). Realizing driving distraction as a key crash contributing factor, studies have explored and revealed that the chance of crash increases if a driver is involved in distracting activities (Young *et al.* 2008, Yannis *et al.* 2010, Guo *et al.* 2017). While quantifying

Table 4. 1 Relevant Literature: The Role of Driving Volatility in Roadway Safety

Authors	Focus	Key driver behaviors	Data and Methodology	Key Conclusions
(Arvin <i>et al.</i> 2019b)	To quantify the effects of driver behaviors, roadway environments, and intersection influence on crash intensity directly and through driving volatility (standard deviation of speed and COV of deceleration)	Distracted driving; Aggressive driving	Data: SHRP2-NDS data Methods: Standard path analysis including two constituent models (Tobit and ordered Probit regression)	Effects on the chance of most severe crash: Aggressive driving (+): indirectly through each of the two volatility measures. Distracted driving (+): both directly and indirectly through the two volatility measures.
(Yeo <i>et al.</i> 2020)	To investigate changes in driving behaviors due to rainfall intensity and explore how volatility relates to traffic violations and crashes	Various measures of driving volatility including variations in acceleration and jerks were used to measure instability in driving behaviors	Data: Driving records of 620 taxis in Seoul (South Korea) Methods: Mixed-effects logistic regression	Driving volatility increases with rainfall intensity. Chance of violation and crash increase with an increase in volatility.
(Wali <i>et al.</i> 2019)	To quantify the effects of driver behaviors, event-based driving volatility, and secondary task duration on safety-critical events	Unfamiliar/inexperience with roadway; Aggressive Driving; Secondary task duration; Various measures of driving volatility	Data: SHRP2-NDS data Methods: Fixed- and random parameter discrete choice models	Risk of both crash and near-crash increases with intentional volatility.
(Wali and Khattak 2020)	To explore the relationship of driver behaviors, driving volatility, and key variables with a crash- and near-crash risk in a school zone.	Eating without utensils; A driver is using cell phone/texting; A driver is distracted; Violations; Various measures of driving volatility	Data: SHRP2-NDS data Methods: Generalized mixed logit models	Crash risk (in a school zone) increases with driver eating without utensils, being distracted with a cell phone, and volatility.

the crash risk for distracted drivers (or those involved in a specific secondary task) may be important, it is also imperative to understand how distraction duration relates to the crash and near-crash outcome (Klauer *et al.* 2010, Liang *et al.* 2012, Gershon *et al.* 2019). Utilizing the instrumented vehicle data (SHRP2-NDS), studies have found that an increase in the duration of distraction during driving significantly increases the chance of a crash and/or near-crash (Klauer *et al.* 2010, Liang *et al.* 2012, Tian *et al.* 2013, Simons-Morton *et al.* 2014, Gershon *et al.* 2019, Arvin and Khattak 2020, Wali and Khattak 2020, Ahmad *et al.* 2021). The abovementioned studies provide useful insights about how the duration of distraction relates to crashes and/or near-crashes; however, most of these studies have not accounted for the possible non-linear relationship of distraction duration with SCEs. As discussed in the previous subsection, previous studies have also found that instability in driving significantly induces the risk of both crash and near-crash (Wali *et al.* 2019, Yeo *et al.* 2020, Fu *et al.* 2021, Khattak *et al.* 2021b). To sum it up, previous studies have investigated 1) how the duration of distraction relates to SCEs directly, and 2) how instability in driving relates to SCEs. This study considers that duration of distraction may not only relate to SCEs directly but also through instability in driving – something which remains unexplored (to the best of authors' knowledge). To capture the overall effects of duration of distraction on SCEs through instability in driving, studies suggest using path analysis which could reveal and quantify the important pathways from the duration of distraction to SCEs through instability in driving.

Study Scope and Contribution

Considering the aforementioned research gaps, we aim to test a hypothesis which is that the subject driver, after making certain errors and violations, may take corrective actions which may induce driving volatility, thus increasing crash risk – something which needs to be investigated. Besides, another hypothesis in this study is that duration of distraction may both directly and indirectly (through instability in driving) relate to SCEs in a non-linear manner which needs to be taken into account in the analysis. From a methodological standpoint, we estimate the two endogenous variables (including COV of speed and event outcome) jointly, assuming that there could be some unobserved factors that influence both instability in driving and event outcome (including baselines, near-crashes, and crashes). Note that most of the previous relevant studies have used standard path analysis which considers the unobserved factors (residuals) to be uncorrelated and their results could be still consistent. However, if a correlation between the two sets of unobserved factors exists but is not taken into account, the estimates could be biased and inefficient. To bridge the identified research and methodological gaps, this study applies path analytic framework via joint estimation procedure to understand how driving errors and violations, and duration of distraction relate to SCEs both directly and through driving instability while accounting for potential correlation between the unobserved factors associated with the driving instability and event outcome. Furthermore, unlike most of the studies which

rely on the average marginal effects (suggesting a constant change in the response variable or probability of specific category of the response variable like in Tobit and ordered Probit model respectively), this study accounts for the non-linear relationship of distraction duration with SCEs both directly and through driving instability in the joint path analytic framework.

Methodology

Conceptual Framework

This study hypothesizes that driving errors and violations may influence driving volatility (speed volatility) which in turn may significantly influence SCEs (Figure 4.1). The key idea is that drivers, after making an error or violation and realizing it, may take corrective actions which could induce driving volatility. Furthermore, this study hypothesizes that a longer duration of distraction may lead to higher instability in driving and a higher chance of SCEs. Furthermore, the effects of duration of distraction on both instability in driving and SCEs could be not linear but rather non-linear which need to be captured in the analysis. To achieve the study objectives, this study applies a rigorous path analytic framework (Liu 2015, Arvin *et al.* 2019b). From a methodological standpoint, we assume that there could be some unobserved factors that significantly influence both COV of speed and event outcome (baseline, near-crash, and crash). To account for potential correlation between the two sets of unobserved factors, this study estimates the two constituent models including the Tobit model (COV of speed) and the ordered Probit model (event outcome) in a joint fashion which could bring efficiency gains. Note that most of the relevant studies have used standard path analysis which assumes the unobserved factors to be uncorrelated (Liu 2015, Arvin *et al.* 2019b). From this point onward in the paper, driving instability refers to COV of speed. Note that both the terms are interchangeably used and imply the same situations.

Modeling Framework

This study estimates two models including the Tobit model (for instability in driving) and the ordered Probit model (for event outcome) in a joint fashion. The joint estimation (system of equations) is used to account for potential correlation between the unobserved factors which could significantly affect the two response outcomes. For each outcome type (and for the joint model), the final models were systematically derived while including the most important variables statistical significance, and specification parsimony. In doing so, the likelihood ratio test, Akaike information criterion “AIC” (Akaike 1974), and Bayesian information criterion “BIC” (Schwarz 1978) statistics were examined. In comparing the competing models, the one with lower AIC and BIC statistics is selected as the best fit model after adjusting for degrees of freedom. Studies also suggest using other performance measures for accessing the model

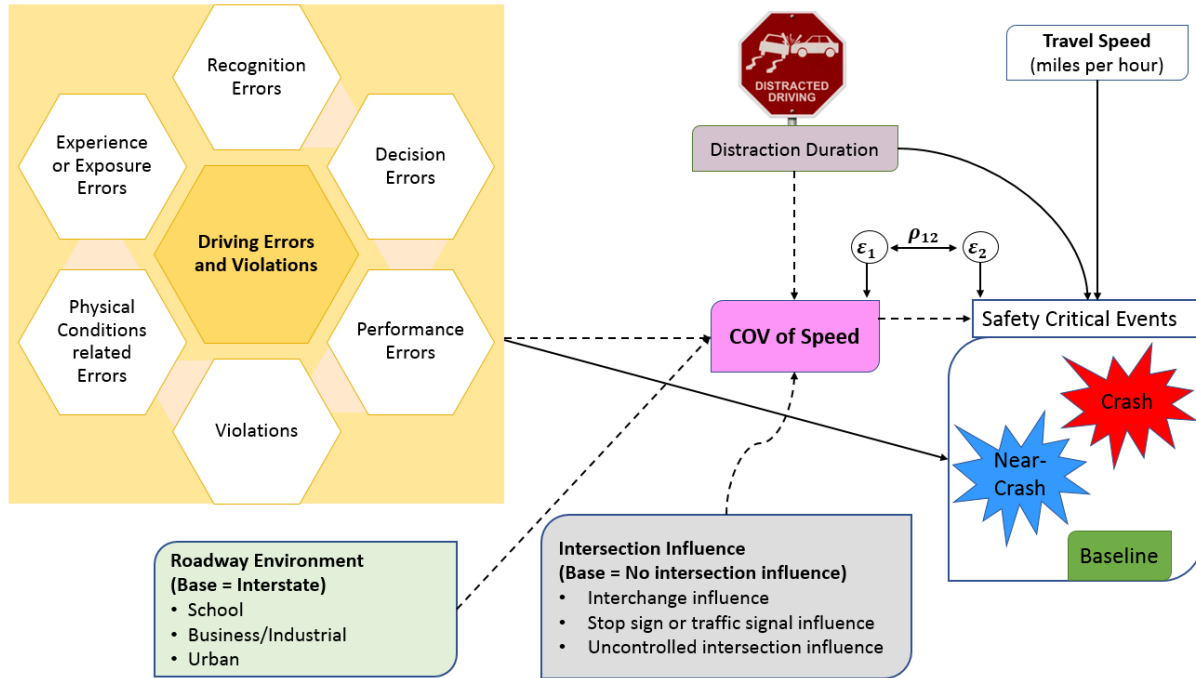


Figure 4. 1 Conceptual Framework: Path Analysis through Joint Estimation

Note: COV of Speed (instability in driving) in SHRP2-NDS subsample exhibits corner-solution setup and is modeled via Tobit regression. Please note that the three possible outcomes in the event outcome model include baseline, near-crash, and crash. The solid lines and dashed lines in Figure 4.1 refer to direct and indirect pathways, respectively. In the first stage, the dependent variable is COV of speed, whereas, in the second stage, the dependent variable is event outcome (baseline, near-crash, and crash) which are modeled through Tobit and ordered Probit model, respectively.

performance depending on the study scope. For instance, BIC and Schwarz criterion “SBIC” are considered as more parsimonious criteria; however, SBIC outperforms the BIC (Nickelsburg 1985). For the sake of completeness, SBIC (Nickelsburg 1985), Consistent AIC “CAIC” (Bozdogan 1987), and Bozdogan’s Information Complexity “ICOMP” (Bozdogan 1990) is considered (where applicable) to select the final model while comparing it to its counterparts. For details about various criteria used to assess the model performance, please refer to the studies in parenthesis (Akaike 1974, Schwarz 1978, Nickelsburg 1985, Bozdogan 1987, 1990). The two models are briefly discussed below.

Tobit Regression

Tobit regression is usually used to model censored or truncated data (Anastasopoulos *et al.* 2008, Washington *et al.* 2010, Ma *et al.* 2015). Truncation and censoring in the data are considered as two different data observability issues (Anastasopoulos *et al.* 2008, Washington *et al.* 2010, Ma *et al.* 2015). On other hand, in corner-solution problems, there is no observability issue or defect in the data; however, it happens that significant observations are reported at lower thresholds (e.g., instability in driving exhibits spikes at 0-1).

The general formulation of the Tobit regression to model speed volatility in a corner-solution setup is given below (Anastasopoulos *et al.* 2008, Washington *et al.* 2010, Debnath *et al.* 2014, Ma *et al.* 2015, Anastasopoulos 2016).

$$\begin{aligned}
 Y_i^* &= \beta_1 X_i + \varepsilon_i, \quad (i = 1, 2, 3, \dots, N) \\
 Y_i &= Y_i^* && \text{(if } Y_i^* > \tau) \\
 Y_i &= \tau_Y && \text{(if } Y_i^* \leq \tau)
 \end{aligned} \tag{1}$$

Where Y_i^* and Y_i indicate latent variable (stochastic index) and dependent variable respectively, N indicates the number of observations, β_1 indicates a set of parameters to be estimated for a set of explanatory variables (X_i), and ε_i indicates independent error term following a normal distribution with $N(0, \sigma_1^2)$. The likelihood function for the Tobit model can be given below (Anastasopoulos *et al.* 2008, Washington *et al.* 2010, Ma *et al.* 2015).

$$L = \prod_i^N \left[\frac{1}{\sigma} \phi \left(\frac{Y_i - \mu}{\sigma} \right) \right]^{d_i} \left[1 - \phi \left(\frac{\mu - \tau}{\sigma} \right) \right]^{1-d_i} \tag{2}$$

Where τ indicates the point where the values exhibit left-spike (corner solution), ϕ is standard normal density function, and Φ is standard cumulative normal distribution function. The value of τ is usually set to zero ($\tau = 0$), whereas μ is parameterized as a function of observed variables and parameters to be estimated $\beta_1 X_i$. Having said this, the likelihood function becomes as given below.

$$L = \prod_i^N \left[\frac{1}{\sigma} \phi \left(\frac{Y_i - \beta_1 X_i}{\sigma} \right) \right]^{d_i} \left[1 - \phi \left(\frac{\beta_1 X_i}{\sigma} \right) \right]^{1-d_i} \quad (3)$$

After expanding Equation (3), the log-likelihood can be given below.

$$\ln L = \sum_i^N \left\{ d_i \left(-\ln \sigma + \ln \phi \left(\frac{Y_i - \beta_1 X_i}{\sigma} \right) \right) + (1 - d_i) \ln \left(1 - \phi \left(\frac{\beta_1 X_i}{\sigma} \right) \right) \right\} \quad (4)$$

Equation (5) includes two parts: 1) The first part relates to a conventional regression part for uncensored observations; 2) The second part relates to the estimable probabilities that particular observations are censored (Sigelman and Zeng 1999). To understand the change in a response variable with a unit increase in a specific continuous explanatory variable (or when the indicator variable switches from 0 to 1), marginal effects (MEs) can be computed. In Tobit regression, three types of MEs can be computed which relate to various expected values of the response variable (Sigelman and Zeng 1999). One or more of these MEs can be used depending on the context and research scope (Sigelman and Zeng 1999). We compute all three types of MEs in this paper; however, we use only one of the MEs based on our research scope which is briefly discussed later in the paper. MEs for the actual response variable based on observations that are not censored, the following equation can be used (Sigelman and Zeng 1999).

$$\frac{\partial E[Y_i | Y_i > 0]}{\partial X_i} = \beta_i \left\{ 1 - \lambda(\alpha) \left[\frac{\beta_1 X_i}{\sigma} + \lambda(\alpha) \right] \right\} \quad (5)$$

Where,

$$\lambda(\alpha) = \frac{\phi \left(\frac{\beta_1 X_i}{\sigma} \right)}{\Phi \left(\frac{\beta_1 X_i}{\sigma} \right)} \quad (6)$$

Equations (5) and (6) become more useful in “*true censoring*”. In our case, the instability in driving exhibits left spikes (corner-solution); however, its values at lower limits are true and are of interest. Having said this, we compute MEs using all observations (including both censored and uncensored) through Equation (7).

$$\frac{\partial E[Y_i]}{\partial X_i} = \phi \left(\frac{\beta_1 X_i}{\sigma} \right) \beta_1 \quad (7)$$

Importantly, to determine the probability of an observation being uncensored, the following equation can be used.

$$\frac{\partial \Pr[Y_i > 1 | X_i]}{\partial X_i} = \varphi\left(\frac{\beta_1 X_i}{\sigma}\right) \frac{\beta_1}{\sigma} \quad (8)$$

Note that MEs obtained via Equation (7) provides a more useful interpretation for change in instability in driving (exhibiting corner-solution) with a unit increase in a specific covariate; therefore, Equation (7) can be preferably used in our case (Greene 2003, Wooldridge 2010). To conclude, from this point onward in the paper, MEs in the case of Tobit regression would refer to Equation (7).

Ordered Probit Model

For the event outcome, the study estimates an ordered Probit model as the event outcome (baseline, near-crash, and crash) follows a clear ordinal pattern (Kockelman and Kweon 2002, Quddus *et al.* 2002, Rifaat and Chin 2007, Washington *et al.* 2010, Arvin *et al.* 2019). Note that COV of speed (response outcome in the first stage) is included as an explanatory variable in the event outcome model. A general expression for the ordered Probit model is given below (Kockelman and Kweon 2002, Abdel-Aty 2003, Liu *et al.* 2015, Arvin *et al.* 2019b).

$$Y_2 = \beta_0 + \beta_2 X_i + \gamma Y_1 + \varepsilon_2 \quad (9)$$

Where Y_2 is a latent variable for event outcome (baseline, near-crash, and crash); β_0 and β_2 show constant terms and vectors of estimable parameters of the associated factors $\mathbf{X}_i$ respectively; γ shows association of instability in driving with event outcome; and ε_2 is an independent error term with $\mathbf{N}(\mathbf{0}, \sigma_2^2)$. The observed event outcome (Y_2^*) relates to the latent measure of event outcome (Y_2) as shown in the equation below (Kockelman and Kweon 2002, Washington *et al.* 2010). For instance, the probability of any category of the response variable with driving volatility (Y_1) used as an independent variable in the Ordered Probit model can be determined as below (Kockelman and Kweon 2002, Washington *et al.* 2010).

$$P(Y_2^* = l) = \Phi(\mu_l - \gamma Y_1) - \Phi(\mu_{l+1} - \gamma Y_1) \quad (10)$$

$$\Phi(\mu) = \frac{1}{2\pi} \int_{-\infty}^{\mu} e^{\left(-\frac{1}{2}\omega^2\right)} d\omega \quad (11)$$

Where l indicates different outcomes or categories of the response variable (baseline, near-crash, and crash), $P(Y_2^* = l)$ is the chance of an individual " n " to be involved in event l , and Φ is a standard normal distribution function. MEs are computed for continuous variables (COV of speed, and secondary task duration), which quantify probability changes in a specific category of the response variable with a unit

increase in a specific continuous variable when all other variables are kept at their constant or mean values, see below (Washington *et al.* 2010, Arvin *et al.* 2019b).

$$\frac{\partial P(Y_2^*=l)}{\partial X} = -[\Phi(\mu_l - \beta_2 X_i) - \Phi(\mu_{l+1} - \beta_2 X_i)]\beta_2 \quad (12)$$

MEs for a dichotomous variable (switching from 0 to 1) can be computed as (Washington *et al.* 2010, Arvin *et al.* 2019b):

$$\partial P(Y_2^* = l|X) = P(Y_2^* = l|X_i = 1) - P(Y_2^* = l|X_i = 0) \quad (13)$$

Joint Estimation (System of Equations)

Most of the previous studies consider that unobserved factors (residuals) associated with the two response outcomes are uncorrelated (Liu 2015, Arvin *et al.* 2019b). However, there could be some unobserved factors that may significantly influence the two response outcomes (O'Higgins 2009). It should be noted that results obtained from standard path analysis are still consistent (Liu 2015, Arvin *et al.* 2019b). However, if a correlation between the unobserved factors exists, then using joint estimation (system of equations) becomes more relevant, which could bring efficiency gains. Furthermore, ignoring this methodological concern may lead to biased estimates (underestimation or overestimation) (Ahmad *et al.* 2021). To account for the potential correlation between the unobserved factors which affect both instability in driving and event outcome, we use the Limited Information Maximum Likelihood (LIML) procedure to estimate the two equations in a joint fashion (Roodman 2009). In Stata, the two equations can be jointly estimated using the Conditional Mixed-Process (CMP) routine (Roodman 2009). Using the CMP routine to jointly estimate the two equations (Equation 1 and Equation 9), the joint distribution of the two error terms (ε_1 and ε_2) are assumed to follow the multivariate normal distribution as below (O'Higgins 2009, Roodman 2009).

$$\varepsilon = (\varepsilon_1, \varepsilon_2) \sim N(0, \sigma^2) \quad (14)$$

Note that correlation between unobserved factors can be captured via the correlation coefficient (ρ).

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 \\ \rho_2 & 1 \end{pmatrix} \right] \quad (15)$$

Path Analysis: Computing Indirect and Total Effects

Once MEs from the two models are obtained, we use rigorous path analysis to quantify direct, indirect, and total effects of driving errors, violations, and roadway environments on crashes and near-crashes (Liu *et al.* 2015, Arvin *et al.* 2019b). Note that the indirect effects of a specific variable on crashes or near-crashes through instability in driving are obtained by multiplying MEs of a particular variable from the first model (Tobit model) with MEs of instability in driving from the second model (ordered Probit model), where the latter one links COV of speed to crashes or near-crashes (Liu *et al.* 2015, Arvin *et al.* 2019b). The following equations can be used to compute the indirect and total effects of specific variables on SCEs through instability in driving (Liu *et al.* 2015, Arvin *et al.* 2019b).

$$\text{Indirect MEs} = (\text{MEs from Tobit Model}) * (\text{MEs from Ordered Probit Model}) \quad (16)$$

$$\text{Total MEs} = \text{Direct MEs} + \text{Indirect MEs} \quad (17)$$

Data

Data Source and Systematic Classification of Driving Errors

This study uses the SHRP2-NDS subsample that includes real-world data on pre-crash driving behaviors, secondary tasks, and their duration, vehicle kinematics, and diverse roadway environments (Hankey *et al.* 2016). The SHRP2-NDS data was collected via instrumental vehicles (fitted with unobtrusive cameras and recording equipment) to record second-to-second driving behavior, performance, eye glance behavior, and other necessary details about the subject drivers in instrumented vehicles (Dingus *et al.* 2006). Data on instantaneous speed, acceleration, and deceleration is also available in SHRP2-NDS data (Hankey *et al.* 2016), which can be used to derive various volatility indices (Arvin *et al.* 2019b, Wali and Khattak 2020). Studies used various measures of driving volatility, which include the standard deviation of speed/acceleration/deceleration, the COV of speed/acceleration/deceleration, and the quartile coefficient of variation of speed/acceleration and deceleration (Arvin *et al.* 2019b). Besides the study objectives, it is important to mention that among all the volatility indices (mentioned above), COV of speed showed the strongest correlation through pairwise correlation. One advantage of SHRP2-NDS data is that it also includes data on near-crashes and baselines where the latter one could be used to compute crash risk and near-crash risk which could not be possible in police-reported crashes (Dingus *et al.* 2016, Arvin *et al.* 2019b, Khattak *et al.* 2021a). A team of data reductionists was selected via a comprehensive selection procedure for data quality assurance who was assigned the task of extracting data from the videos (Dingus *et al.* 2016, Hankey *et al.* 2016).

In the SHRP-NDS data, several variables including driving behaviors and secondary tasks are defined as contributing factors in SCEs (Dingus *et al.* 2016, Hankey *et al.* 2016, Khattak *et al.* 2021a). Using the aforementioned variables, this study applies a recently developed TDEV to systematically classify human factors that contributed to crashes (Khattak *et al.* 2021a). Through TDEV, driving behaviors by subject drivers can be classified into recognition errors, decision errors, performance errors, physical condition-related errors, experience errors, and violations (Khattak *et al.* 2021a). For more details about TDEV and its constituent driving errors and violations, please refer to (Khattak *et al.* 2021a). Figure 4.2 illustrates the TDEV using SHRP-NDS data (Khattak *et al.* 2021a). Note that compared to the past, both types and sources of distraction have significantly changed, which justifies the need for TDEV based on SHRP2-NDS data (Khattak *et al.* 2021a). For instance, types of recognition errors including cell phone texting, cell phone browsing, adjusting/monitoring radio, and other types of distractions in SCEs could only be available through large-scale SHRP2-NDS data (Khattak *et al.* 2021a). Likewise, decision errors including exceeding a safe speed but not the speed limit and following too closely are some of the driving behaviors which could only be captured through SHRP2-NDS data efforts (Khattak *et al.* 2021a).

For understanding, based on the SHRP2-NDS data dictionary (SHRP2-NDS), baseline, near-crash, and crash can be defined below.

- **Baseline:** The baselines refer to non-event (routine) driving where the subject drivers were not involved in safety-critical events (Hankey *et al.* 2016, Ahmad *et al.* 2021, Khattak *et al.* 2021a). In the SHRP2-NDS, the Virginia Tech Transportation Institute (VTTI) has randomly selected the baselines intending to select at least one baseline per driver (who was involved in SCE). The typical duration of the baseline was 20-30 seconds (Hankey *et al.* 2016). Note that the number of baselines for a specific driver was proportional to his/her overall driving time in the SHRP2-NDS experiment (Hankey *et al.* 2016, Ahmad *et al.* 2021, Khattak *et al.* 2021a). Furthermore, one of the criteria for selecting baseline was to make sure that the vehicle speed was not lower than 5 miles per hour (MPH) for more than 2 consecutive seconds so that the long time stopping effects could be ignored and only those drivers could be considered who could have a risk of involving in an at-fault crash (Hankey *et al.* 2016, Ahmad *et al.* 2021, Khattak *et al.* 2021a).
- **Near-Crash:** Following the SHRP2-NDS data dictionary, a near-crash refers to a situation where a subject (in instrumented vehicle) driver or any other driver, cyclist, pedestrian, or animal was required to take an evasive maneuver to avoid a crash (Hankey *et al.* 2016). The evasive maneuvers refer to steering, braking, accelerating, or a combination of two or more of these maneuvers (Hankey *et al.* 2016). In near-crashes, the subject vehicle should have not contacted any moving or

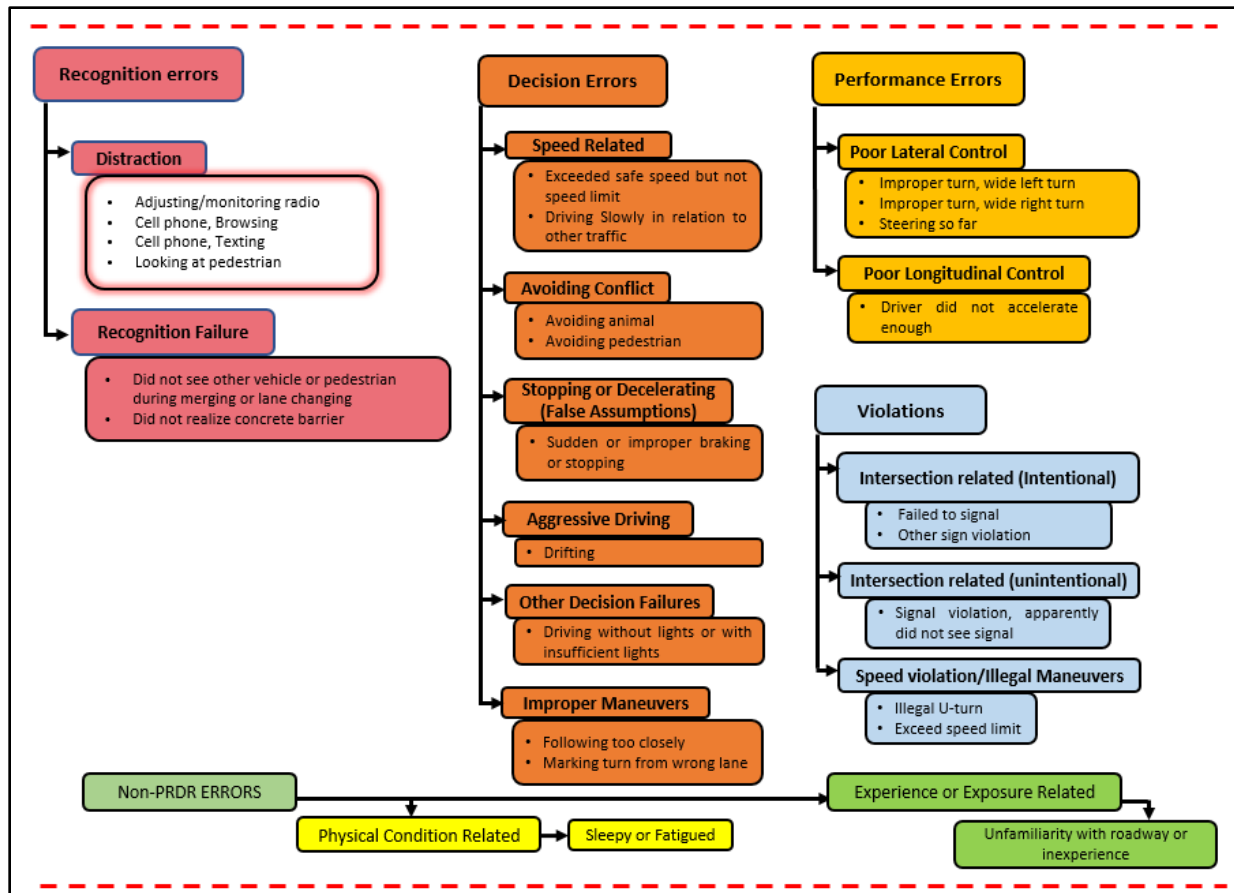


Figure 4. 2 Modified Version of Systematic TDEV “Source: (Khattak *et al.* 2021a)”

Note: Figure 4.2 shows a modified and brief version of TDEV (Khattak *et al.* 2021a). For the sake of brevity, a few examples of each driving error and violations from SHRP2-NDS data are shown here. For details, please refer to (Khattak *et al.* 2021a).

fixed object and there should be no roadway departure maneuver by the vehicle as per the SHRP2-NDS dictionary (Hankey *et al.* 2016).

- **Crash:** The situation in which a subject vehicle could have come in contact with either a fixed or moving vehicle, or object (barriers, pedestrians, animals) which resulted in dissipation or transfer of kinetic energy was considered to be a crash in the SHRP2-NDS data (Hankey *et al.* 2016). Note that according to the SHRP2-NDS dictionary, crashes could also include non-premeditated roadway departures of a subject vehicle where at least one tire left the intended or paved travel roadway surface (Hankey *et al.* 2016).

Descriptive Statistics

The SHRP2-NDS sub-sample used in this study includes 9,237 events which consist of 7,393(80.04%) baselines, 1,237 (13.39%) near-crashes, and 607 (6.57%) crashes (Table 4.2). To understand the distribution of instability in driving (COV of speed), please refer to Figure 4.3. The COV of speed (for all baselines, near-crashes, and crashes) exhibits a corner-solution set-up and hence can be modeled through Tobit regression – for details, see the methodology section. Statistics reveal that human factors (driving errors and violations) still prevail, contributing to 92.48% of overall crashes (Table 4.2). In both crashes and near-crashes, recognition errors were the predominant errors, followed by decision errors (Table 4.2). Furthermore, statistics reveal that on average, drivers in crashes and near-crashes had a higher mean COV of speed (0.65 and 0.55, respectively) compared to baselines (no SCE) where the mean COV of speed was found to be 0.14 (Table 4.2). For better understanding, we show the distribution of COV of speed for baselines, drivers in SCEs (Figure 4.3).

Referring to the distribution of distraction duration in baselines and SCEs, Figure 4.4 illustrates that more than 50% of the baseline drivers were not distracted (distraction duration was 0 seconds). It should be noticed that almost 25% of the baseline drivers were distracted for 5 seconds but they were not involved in any SCEs. However, such baseline drivers, if distracted for a longer duration (5 seconds), can get into a crash or near-crash in the future. Studies show similar insights where 10% of the baseline drivers were involved in driving errors or violations; however, it does not justify those wrong driving behaviors but rather reveals a potential safety risk in the future (Khattak *et al.* 2021a). Compared to baselines drivers, those who were involved in SCEs show longer distraction duration as expected.

Before conducting in-depth analysis, we first checked pairwise correlation of eight volatility indices with the final response outcome (including baseline, near-crash, and crash) which include: standard deviation of speed, COV of speed, quartile COV of speed, time-varying stochastic volatility for speed, COV of acceleration, quartile COV of acceleration, COV of deceleration, and quartile COV of deceleration – for

details about how these indices can be derived, please refer to (Arvin *et al.* 2019b, Wali and Khattak 2020). All volatility indices showed a significant correlation with event outcome; however, COV of speed had the strongest correlation with event outcome which is also the key focus in this study.

Results

This section discusses key results obtained from the two constituent models including the Tobit regression and the ordered Probit model, which were developed to estimate COV of speed and event outcome respectively. Each of these models was systematically derived based on statistical significance, specification parsimony, and theoretical justification to include the most important variables. The decision to include variables in our final models was evaluated based on 95% confidence level criteria. For the sake of completeness and variable importance, we have kept some of the variables which showed partial significance (an indicator of physical conditions-related errors and secondary task duration in the Tobit model). To understand the change in the response variable (COV of Speed) or probability of a specific category of the response variable (event outcome) with a unit increase in a specific continuous explanatory variable (or if an indicator variable switches from 0 to 1), MEs are computed. Note that MEs obtained from individual models are then used to compute the indirect, direct, and total effects of a specific explanatory variable on the final response outcome. In the case of Tobit regression, we present all three types of MEs (as discussed in Section 3.2.1). However, we use and discuss only the most relevant and useful MEs (ME-1 in our case) which refers to a change in the COV of speed (based on both censored and uncensored observations) with a unit increase in a specific explanatory variable.

Separate Estimation: Speed Volatility Model and Event Outcome Model

First, this study applies the standard path analysis (including the two independent models) which assumes that the residuals (unobserved factors) in the two equations are uncorrelated (Table 4.3). After comparing the performance criteria (AIC, BIC, CAIC, SBIC, and ICOMP) of the competing models, the final models with the lowest values of these performance measures were considered as the final models (Table 4.3). Results of the Tobit model indicate that compared to no driving error and violation, all types of driving errors and violations positively influence instability in driving. For instance, our findings indicate that COV of speed increases by 0.2556 and 0.1530 units if a subject driver makes a recognition error and decision error, respectively as compared to no driving error and violation. Results also indicate that compared to the interstate, COV of speed increases in urban areas, school zones, and business/industrial locations. Referring to the event outcome model, our findings indicate that the chance of both crash and near-crash increases if a driver makes an error or a violation compared to no driving error or violation. Note that COV of speed

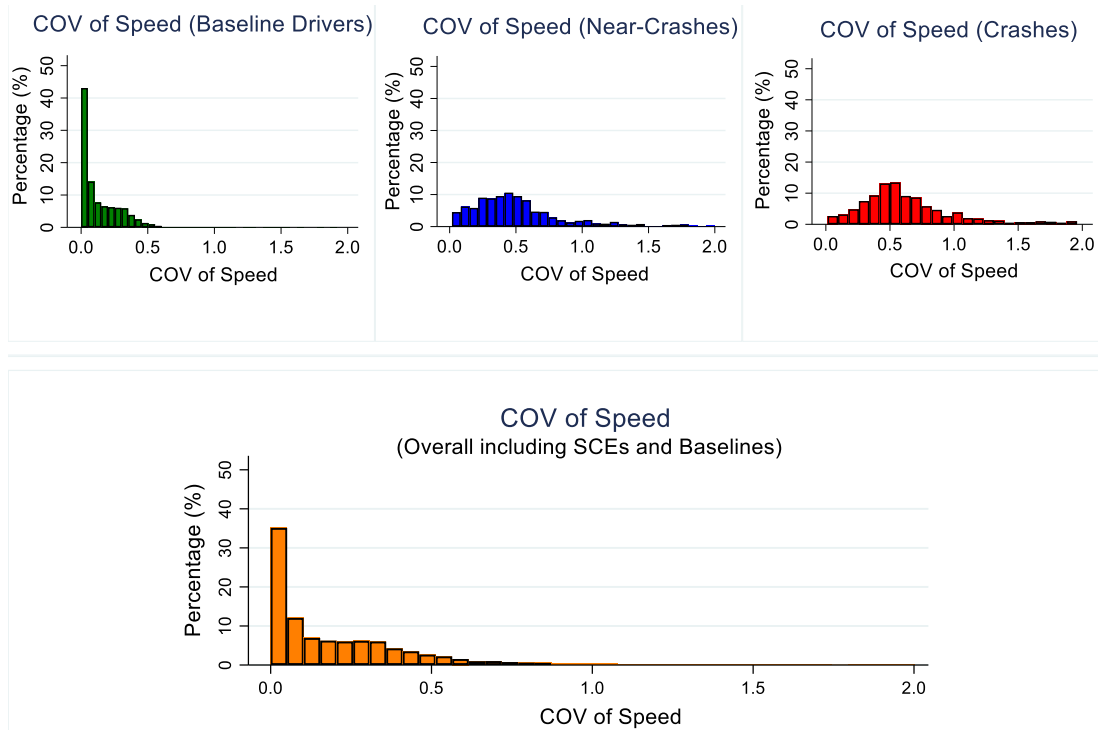


Figure 4.3 Distribution of Drivers based on COV of Speed in Baselines, Near-Crashes, and Crashes

Note: The maximum COV of speed in the SHRP2-NDS subsample is 3.048 (Table 4.2); however, for the sake of brevity, we show COV of speed in baselines, near-crashes, and crashes till 2.0.

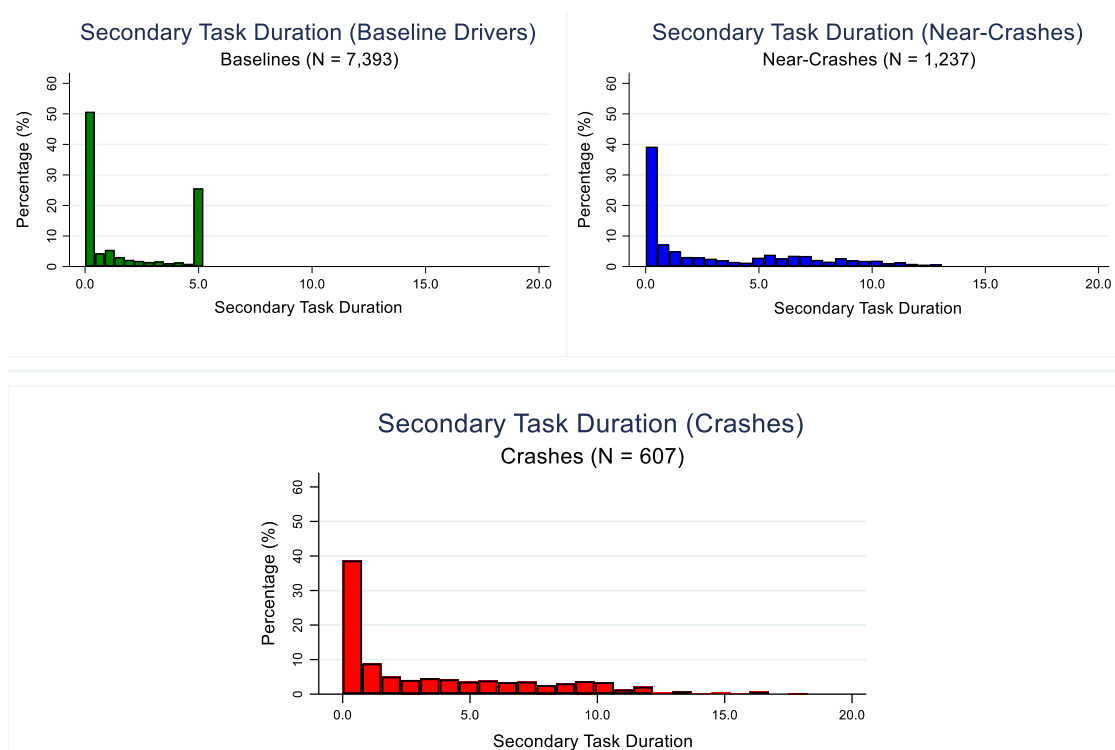


Figure 4.4 Distribution of Baselines and SCEs on Secondary Task Duration (seconds)

Table 4. 2 Descriptive Statistics of Key Variables

Explanatory Variables	Overall (N = 9,237)		Baselines (N = 7,393)		Near-Crashes (N = 1,237)		Crashes (N = 607)	
Continuous	Mean (S.D.)	(Min., Max.)	Mean (S.D.)	(Min., Max.)	Mean (S.D.)	(Min., Max.)	Mean (S.D.)	(Min., Max.)
COV of Speed	0.23 (0.29)	(0, 3.05)	0.14 (0.16)	(0, 2.60)	0.55 (0.42)	(0.2, 3.048)	0.65 (0.40)	(0.014, 2.67)
Secondary Task Duration (sec)	2.08 (2.69)	(0, 18.26)	1.75 (2.15)	(0, 16.51)	3.33 (3.87)	(0, 15.68)	3.61 (4.13)	(0, 18.26)
Speed (miles per hour)	16.29 (8.55)	(0.002, 99.93)	17.72 (8.24)	(0.002, 99.93)	11.67 (7.46)	(0.12, 57.28)	8.28 (6.19)	(0.13, 34.40)
Categorical	Freq. (%)	(Min., Max.)	Freq. (%)	(Min., Max.)	Freq. (%)	(Min., Max.)	Freq. (%)	(Min., Max.)
Driving Errors and Violations								
No Driving Errors	7182 (77.75)	(0, 1)	6659 (90.07)	(0, 1)	477 (38.56)	(0, 1)	46 (7.58)	(0, 1)
Recognition Error	667 (7.22)	(0, 1)	17 (0.23)	(0, 1)	420 (33.95)	(0, 1)	230 (37.89)	(0, 1)
Decision Error	574 (6.21)	(0, 1)	197 (2.66)	(0, 1)	172 (13.90)	(0, 1)	205 (33.77)	(0, 1)
Performance Error	62 (0.67)	(0, 1)	7 (0.09)	(0, 1)	8 (0.65)	(0, 1)	47 (7.74)	(0, 1)
Violation	607 (6.57)	(0, 1)	415 (5.61)	(0, 1)	135 (10.91)	(0, 1)	57 (9.39)	(0, 1)
Physical Condition-related error	119 (1.29)	(0, 1)	93 (1.26)	(0, 1)	17 (1.37)	(0, 1)	9 (1.48)	(0, 1)
Experience or Exposure Error	26 (0.28)	(0, 1)	5 (0.07)	(0, 1)	8 (0.65)	(0, 1)	13 (2.14)	(0, 1)
Roadway Locality								
Open Country or Open Residential	807 (8.74)	(0, 1)	697 (9.43)	(0, 1)	72 (5.82)	(0, 1)	38 (6.26)	(0, 1)
Moderate Residential	1949 (21.10)	(0, 1)	1654 (22.37)	(0, 1)	174 (14.07)	(0, 1)	121 (19.93)	(0, 1)
School	471 (5.10)	(0, 1)	335 (4.53)	(0, 1)	90 (7.28)	(0, 1)	46 (7.58)	(0, 1)
Business/Industrial	3130 (33.89)	(0, 1)	2344 (31.37)	(0, 1)	502 (40.58)	(0, 1)	284 (46.79)	(0, 1)
Interstate	2182 (23.62)	(0, 1)	1902 (25.73)	(0, 1)	239 (19.32)	(0, 1)	41 (6.75)	(0, 1)
Urban	244 (2.64)	(0, 1)	114 (1.54)	(0, 1)	87 (7.03)	(0, 1)	43 (7.08)	(0, 1)
Bypass/Divided Highway with traffic signals	277 (3.00)	(0, 1)	223 (3.02)	(0, 1)	38 (3.07)	(0, 1)	16 (2.64)	(0, 1)
Others (e.g., church, playground)	177 (1.92)	(0, 1)	124 (1.68)	(0, 1)	35 (2.83)	(0, 1)	18 (2.97)	(0, 1)
Intersection Influence								
No intersection influence	6652 (72.01)	(0, 1)	6005 (81.23)	(0, 1)	396 (32.01)	(0, 1)	251 (41.35)	(0, 1)
Interchange influence	295 (3.19)	(0, 1)	165 (2.23)	(0, 1)	110 (8.89)	(0, 1)	20 (3.29)	(0, 1)
Stop sign or traffic signal influence	1463 (15.84)	(0, 1)	909 (12.30)	(0, 1)	405 (32.74)	(0, 1)	149 (24.55)	(0, 1)
Uncontrolled intersection influence	308 (3.33)	(0, 1)	130 (1.76)	(0, 1)	125 (10.11)	(0, 1)	53 (8.73)	(0, 1)
Parking lot or driveways influence	415 (4.49)	(0, 1)	149 (2.02)	(0, 1)	146 (11.80)	(0, 1)	120 (19.77)	(0, 1)
Others intersection influence	104 (1.13)	(0, 1)	35 (0.47)	(0, 1)	55 (4.45)	(0, 1)	14 (2.31)	(0, 1)
Level of Service								
LOS C/D/F	873 (9.45)	(0, 1)	495 (6.70)	(0, 1)	327 (26.43)	(0, 1)	51 (8.40)	(0, 1)
LOS A: Free flow traffic condition	6501 (70.38)	(0, 1)	5531 (74.81)	(0, 1)	525 (42.44)	(0, 1)	445 (73.31)	(0, 1)
LOS B: Traffic Flow with some restriction	1863 (20.17)	(0, 1)	1367 (18.49)	(0, 1)	385 (31.12)	(0, 1)	111 (18.29)	(0, 1)
Traffic Flow or Roadway Configuration								
Divided (median strip or barrier)	3702 (40.08)	(0, 1)	3060 (41.39)	(0, 1)	504 (40.74)	(0, 1)	138 (22.73)	(0, 1)
Simply undivided – or with 2WLT	4809 (52.06)	(0, 1)	3862 (52.24)	(0, 1)	620 (50.12)	(0, 1)	327 (53.87)	(0, 1)
One-way traffic	361 (3.91)	(0, 1)	247 (3.34)	(0, 1)	75 (6.06)	(0, 1)	39 (6.43)	(0, 1)
No lanes	365 (3.95)	(0, 1)	224 (3.03)	(0, 1)	38 (3.07)	(0, 1)	103 (16.97)	(0, 1)

Table 4. 3 Results of Speed Volatility Model and Event Outcome Model (Separate Estimation)

Independent variables	Tobit Model (COV of Speed)					Ordered Probit Model (Event Outcome)				
	Coeff.	t-stat	ME-1	ME-2	ME-3	Coeff.	t-stat	MEs		
								Baseline	Near-Crash	Crash
COV of Speed	---	---	---	---	---	1.8473	23.41	-0.2713	0.1674	0.1039
Drivers Errors (Base Outcome = No Driving Error or Violation)										
Recognition Error	0.3274	31.45	0.2556	36.24	0.2010	1.8963	31.44	-0.2784	0.1718	0.1066
Decision Error	0.1960	19.39	0.1530	21.70	0.1204	1.8438	32.2	-0.2707	0.1671	0.1037
Performance Error	0.3169	10.76	0.2474	35.07	0.1946	2.7291	15.74	-0.4007	0.2473	0.1534
Violation	0.0623	6.40	0.0486	6.89	0.0382	1.0356	17.47	-0.1521	0.0938	0.0582
Physical Condition-related Error	0.0406	1.92	0.0317	4.49	0.0249	0.7931	6.02	-0.1165	0.0719	0.0446
Experience or Exposure Error	0.2101	4.67	0.1640	23.25	0.1290	2.2750	9.60	-0.3341	0.2062	0.1279
Roadway Locality (Base = Interstate)										
Open Country or Open Residential	0.0427	3.70	0.0333	4.72	0.0262	---	---	---	---	---
Moderate Residential	0.0972	9.95	0.0759	10.76	0.0597	---	---	---	---	---
School	0.1070	8.16	0.0835	11.84	0.0657	---	---	---	---	---
Business/Industrial	0.1211	14.51	0.0945	13.41	0.0744	---	---	---	---	---
Urban	0.2041	12.02	0.1593	22.59	0.1253	---	---	---	---	---
Bypass or Divided Highway with traffic signals	0.0703	4.78	0.0549	7.78	0.0431	---	---	---	---	---
Others (e.g., church, playground)	0.1315	6.92	0.1026	14.55	0.0807	---	---	---	---	---
Intersection Influence (Base = No intersection influence)										
Interchange influence	0.0724	4.93	0.0565	8.02	0.0445	---	---	---	---	---
Stop sign or traffic signal influence	0.2033	28.14	0.1587	22.50	0.1248	---	---	---	---	---
Uncontrolled intersection influence	0.1365	10.03	0.1066	15.11	0.0838	---	---	---	---	---
Parking lot or driveways influence	0.1937	15.77	0.1512	21.44	0.1189	---	---	---	---	---
Others (e.g., crosswalk, railroad crossing)	0.1765	7.68	0.1378	19.54	0.1084	---	---	---	---	---
Level of Service (Base Category = C to F)										
LOS A: Free flow traffic condition	-0.1218	-14.01	-0.0951	-13.48	-0.0748	---	---	---	---	---
LOS B: Traffic Flow with some restriction	-0.1130	-11.93	-0.0882	-12.51	-0.0694	---	---	---	---	---
Traffic Flow or Roadway Configuration (Base = Divided "median strip or barrier")										
Simply undivided – or undivided with 2WLTL	0.0276	3.86	0.0214	3.15	0.0167	---	---	---	---	---
One-way traffic	0.0324	2.42	0.0252	3.67	0.0197	---	---	---	---	---
No lanes	0.1927	13.85	0.1646	16.71	0.1335	---	---	---	---	---
Secondary Task Duration (second)	-0.0036	-1.83	-0.0006	-0.13	-0.0004	-0.1065	-7.10	0.0070	-0.0067	-0.0003
(Secondary Task Duration) ²	0.0007	2.95	-0.0001	-0.01	-0.0001	0.0147	9.86	-0.0004	0.0002	0.0002

Table 4.3 Results of Speed Volatility Model and Event Outcome Model (Separate Estimation) (Continued)

Independent variables	Tobit Model (COV of Speed)					Ordered Probit Model (Event Outcome)				
	Coeff.	t-stat	ME-1	ME-2	ME-3	Coeff.	t-stat	MEs		
								Baseline	Near-Crash	Crash
Speed (miles per hour)	---	---	---	---	---	0.0476	5.04	-0.0010	0.0000	0.0010
(Speed “miles per hour”)²	---	---	---	---	---	-0.0015	-5.87	0.0002	-0.0001	-0.0001
Constant	0.1417	15.32		---	---	---	---	---	---	---
Var (e. Coefficient of Variation of Speed)	0.0521	67.96	---	---	---	---	---	---	---	---
Thresholds										
μ_1	---	---	---	---	---	2.1622	22.22	---	---	---
μ_2	---	---	---	---	---	3.5598	33.88	---	---	---
Model Summary										
Number of observations	9,239					9,239				
AIC	-1032.011					6663.051				
BIC	-839.475					6755.754				
SBIC	-857.737					6737.492				
CAIC	-832.737					6748.492				
ICOMP	-1049.223					6678.131				

Note: Our discussion related to Tobit regression is based on ME-1 which refers to both censored and uncensored observations. ME-2 indicates the probability of being uncensored while ME-3 is similar to ME-1 but is based on only uncensored observations. Variables that showed full or partial significance (at 95% and 90% confidence level criteria, respectively) were used in the final analysis. Note that AIC, BIC, SBIC, CAIC, and ICOMP values for the Tobit model are found to be negative; however, it is not an issue as for the sake of comparison of the competing models, the model with the lowest values of these performance parameters are found to be better.

shows a significant and positive correlation with the event outcome. Importantly, secondary task duration and its squared term show partial significant and significant association with the instability in driving respectively which provide statistical evidence about the non-linearity in its effects on instability in driving. Also, in the ordered Probit model, secondary task duration and its squared term are significantly associated with the event outcome which also reveals that the effects of distraction duration on SCEs are non-linear (Table 4.3).

Joint Estimation (Systems of Equations) for Speed Volatility and Event Outcome

The joint estimation results show significant evidence that the two sets of unobserved factors (which significantly contribute to both the instability in driving and event outcome) are correlated ($\rho = -0.4112$; $t\text{-stats} = -12.97$). Through joint estimation, the parameter signs for all variables remain the same; however, their magnitudes vary compared to those in the separately estimated models (see Table 4.3 and Table 4.4). This indicates that if a correlation between the unobserved factors were not captured through the joint estimation, parameters would be biased (under- or over-estimated).

The Tobit model results reveal that all driving errors and violations (compared to no driving error or violation) are positively correlated with instability in driving. For instance, results of the Tobit model (in the joint system of equations) suggest that instability in driving increases by 0.2524 and 0.1493 units if a driver makes a recognition error and a decision error, respectively. The two MEs are lower than the corresponding MEs obtained when the Tobit model was independently estimated, thus showing the overestimation of parameter estimates. Similar underestimation (COV of speed in the event outcome model) can be noticed in the parameters of other variables. Note that the VIF values of all variables in each of the two models were lower than 5, suggesting no significant evidence of multicollinearity issue (Menard 2002, James *et al.* 2013).

Path Analysis Results: Joint Estimation

The existence of a significant correlation between the unobserved factors associated with the two response variables validates the use of joint estimation over the independent estimation of the two models. Based on joint-estimation results (Table 4.4), a unit increase in instability in driving (COV of speed) increases the probability of crashes and near-crashes by 0.1985 and 0.3304 units respectively (Table 4.4). The indirect effects of factors associated with a crash through instability in driving are computed by multiplying MEs obtained in the first stage (Tobit regression) with 0.1985 (Table 4.5). A similar procedure is used for the near-crash (Table 4.5). Results of the path analysis suggest that recognition error (compared to no driving error and violation) directly increases the chance of a crash by 6.54% (ME multiplied with 100%), whereas

Table 4. 4 Results of Speed Volatility Model and Event Outcome Model (Jointly Estimated)

Independent variables	Tobit Model (COV of Speed)					Ordered Probit Model (Event Outcome)				
	Coeff.	t-stat	ME-1	ME-2	ME-3	Coeff.	t-stat	MEs		
								Baseline	Near-Crash	Crash
COV of Speed	---	---	---	---	---	3.3834	26.29	-0.5288	0.3304	0.1985
Drivers Errors (Base Outcome = No Driving Error or Violation)										
Recognition Error	0.3233	31.09	0.2524	35.84	0.1984	1.1157	11.90	-0.1744	0.1089	0.0654
Decision Error	0.1913	18.92	0.1493	21.20	0.1174	1.3549	18.04	-0.2118	0.1323	0.0795
Performance Error	0.3112	10.57	0.2429	34.49	0.1909	1.8916	10.15	-0.2957	0.1847	0.1110
Violation	0.0617	6.33	0.0481	6.84	0.0378	0.7943	12.68	-0.1242	0.0776	0.0466
Physical Condition-related Error	0.0401	1.89	0.0313	4.44	0.0246	0.7105	5.60	-0.1111	0.0694	0.0417
Experience or Exposure Error	0.2025	4.50	0.1581	22.44	0.1242	1.6979	7.16	-0.2654	0.1658	0.0996
Roadway Locality (Base = Interstate)										
Open Country or Open Residential	0.0418	3.72	0.0326	4.63	0.0256	---	---	---	---	---
Moderate Residential	0.0891	9.35	0.0696	9.88	0.0547	---	---	---	---	---
School	0.1050	8.28	0.0820	11.64	0.0644	---	---	---	---	---
Business/Industrial	0.1143	13.99	0.0892	12.67	0.0701	---	---	---	---	---
Urban	0.2116	13.05	0.1652	23.46	0.1298	---	---	---	---	---
Bypass or Divided Highway with traffic signals	0.0714	4.99	0.0558	7.92	0.0438	---	---	---	---	---
Others (e.g., church, playground)	0.1254	6.83	0.0979	13.90	0.0770	---	---	---	---	---
Intersection Influence (Base = No intersection influence)										
Interchange influence	0.1030	7.23	0.0804	11.42	0.0632	---	---	---	---	---
Stop sign or traffic signal influence	0.1960	27.31	0.1530	21.73	0.1203	---	---	---	---	---
Uncontrolled intersection influence	0.1667	12.85	0.1301	18.48	0.1023	---	---	---	---	---
Parking lot or driveways influence	0.2151	18.50	0.1679	23.84	0.1320	---	---	---	---	---
Others (e.g., crosswalk, roundabouts)	0.1877	8.67	0.1465	20.80	0.1152	---	---	---	---	---
Level of Service (Base Category = C to F)										
LOS A: Free flow traffic condition	-0.1231	-14.75	-0.0961	-13.65	-0.0755	---	---	---	---	---
LOS B: Traffic Flow with some restriction	-0.1113	-12.23	-0.0869	-12.34	-0.0683	---	---	---	---	---
Traffic Flow or Roadway Configuration (Base = Divided "median strip or barrier")										
Simply undivided – or undivided with 2WLT	0.0272	3.94	0.0211	3.10	0.0164	---	---	---	---	---
One-way traffic	0.0363	2.82	0.0283	4.09	0.0221	---	---	---	---	---
No lanes	0.1943	14.50	0.1660	16.81	0.1347	---	---	---	---	---
Secondary Task Duration	-0.0034	-1.71	-0.0005	-0.12	-0.0003	-0.0851	-5.91	0.0058	-0.0051	-0.0007
(Secondary Task Duration) ²	0.0006	2.81	-0.0001	-0.01%	-0.0001	0.0115	7.86	-0.0004	0.0002	0.0001

Table 4. 4 Results of Speed Volatility Model and Event Outcome Model (Jointly Estimated) (Continued)

Independent variables	Tobit Model (COV of Speed)					Ordered Probit Model (Event Outcome)				
	Coeff.	t-stat	ME-1	ME-2	ME-3	Coeff.	t-stat	MEs		
								Baseline	Near-Crash	Crash
Speed “miles per hour”	---	---	---	---	---	0.0510	5.88	-0.0032	0.0015	0.0017
(Speed “miles per hour”) ²	---	---	---	---	---	-0.0011	-4.76	0.0002	-0.0001	-0.0001
Constant	0.1450	16.22		---	---	---	---	---	---	---
Thresholds										
μ ₁	---	---	---	---	---	2.5149	27.63	---	---	---
μ ₂	---	---	---	---	---	3.8050	39.05	---	---	---
ρ (Rho)	-0.4112 (t-stats = -12.97)									
Model Summary										
Number of observations	9,239									
AIC	5,516.989									
BIC	5,809.359									
SBIC	5763.704									
CAIC	5799.704									
ICOMP	---									

Note: Our discussion related to Tobit regression is based on ME-1 which refers to both censored and uncensored observations. ME-2 indicates the probability of being uncensored while ME-3 is like ME-1 but is based on only uncensored observations. Variables that showed full or partial significance (at 95% and 90% confidence level criteria, respectively) were used in the final analysis.

Table 4. 5 Path Analysis Results (Based on Joint Estimation)

Independent variables	Direct Effects on Crash (%)	Indirect Effects on Crash through COV of Speed (%)	Total Effects (Direct + Indirect through COV of Speed) on Crash (%)	Direct Effects on Near-Crash (%)	Indirect Effects on Near-Crash through COV of Speed (%)	Total Effects (Direct + Indirect through COV of Speed) on Near-Crash (%)
COV of Speed	19.85	---	19.85	33.04	---	33.04
Drivers Errors (Base Outcome = No Driving Error or Violation)						
Recognition Error	6.54	5.01	11.55	10.89	8.34	19.23
Decision Error	7.95	2.96	10.91	13.23	4.93	18.16
Performance Error	11.10	4.82	15.92	18.47	8.02	26.49
Violation	4.66	0.96	5.62	7.76	1.59	9.35
Physical Condition-related Error	4.17	0.62	4.79	6.94	1.03	7.97
Experience or Exposure Error	9.96	3.14	13.10	16.58	5.22	21.80
Roadway Locality (Base = Interstate)						
Open Country or Open Residential	---	0.65	0.65	---	1.08	1.08
Moderate Residential	---	1.38	1.38	---	2.30	2.30
School	---	1.63	1.63	---	2.71	2.71
Business/Industrial	---	1.77	1.77	---	2.95	2.95
Urban	---	3.28	3.28	---	5.46	5.46
Bypass or Divided Highway with traffic signals	---	1.11	1.11	---	1.84	1.84
Others (e.g., church, playground)	---	1.94	1.94	---	3.23	3.23
Intersection Influence (Base = No intersection influence)						
Interchange influence	---	1.60	1.60	---	2.66	2.66
Stop sign or traffic signal influence	---	3.04	3.04	---	5.05	5.05
Uncontrolled intersection influence	---	2.58	2.58	---	4.30	4.30
Parking lot or driveways influence	---	3.33	3.33	---	5.55	5.55
Others (e.g., crosswalk, roundabouts)	---	2.91	2.91	---	4.84	4.84
Level of Service (Base Category = C to F)						
LOS A: Free flow traffic condition	---	-1.91	-1.91	---	-3.17	-3.17
LOS B: Traffic Flow with some restriction	---	-1.73	-1.73	---	-2.87	-2.87
Traffic Flow or Roadway Configuration (Base = Divided "median strip or barrier")						
Simply undivided – or undivided with 2WLTL	---	0.42	0.42	---	0.70	0.70
One-way traffic	---	0.56	0.56	---	0.94	0.94
No lanes	---	3.30	3.30	---	5.49	5.49

it further increases the chance of crash indirectly by 5.01% ($= (0.2525 \times 0.1985) \times 100\%$) through COV of speed. Thus, the total increase in the chance of a crash, due to recognition error, is 11.55% ($= 6.54\% + 5.01\%$) which is significantly greater than the direct increase in the chance of crash due to recognition error. A similar procedure was used to compute indirect effects of other driving errors and roadway environments which indirectly link to crashes and near-crashes through instability in driving. To conclude this, several useful findings surfaced from the path analysis. For example, our findings indicate that all driving errors and violations (compared to no driving error and violation) significantly induce crash risk and near-crash risk not only directly but through instability in driving (Table 4.5). Such findings can only be obtained using rigorous path analysis. For details about how various driving errors in roadway environments and other key variables relate to crashes and near-crashes indirectly through instability in driving, please refer to Table 4.5 and Figure 4.5. To understand how the duration of distraction relates to SCEs both directly and through instability in driving, please refer to Table 4.6. Note that the results show significant statistical evidence about the non-linear association of duration of distraction with both driving instability and SCEs; therefore, we use predictive margins for the path analysis rather than relying on average MEs.

Discussion

The paper contributes by jointly estimating models that explore potential pathways leading from errors, violations, and duration of distraction in diverse roadway environments to safety-critical events through instability in driving. The joints estimation showed that unobserved factors in the two models estimated for path-analysis are significantly correlated. We found that compared to no driving errors, all five driving errors and violations associate with higher instability in driving (COV of speed). For instance, compared to no driving error and violations, a recognition error made by a subject driver increases the instability in driving by 0.2524 units. The finding of recognition error was expected as a driver after making recognition error and realizing it, may take certain corrective actions which could increase driving instability (Arvin *et al.* 2019b). Likewise, we found that compared to no driving error or violations, decision error and performance error by the subject driver increase instability in driving by 0.1493 and 0.2429 units, respectively. Note that among all driving errors and violations, performance error is found to have the highest impact on instability in driving, which makes sense as performance errors are more closely related to driving performance during the execution of intended actions. Furthermore, we also found that inexperienced drivers or those who were new to a specific roadway environment exhibit greater driving instability. On the other hand, driving violation and physical condition error also show a significant and positive correlation with instability in driving; however, their impact is lower compared to the other four driving errors. Furthermore, instability in driving is significantly higher in urban areas (by 0.1652 units), business/industrial locations (by 0.0892 units), and school zones (by 0.0820 units) compared to driving on

Association strength with SCEs (Marginal Effects in %): — 0-4.99% — 5-9.99% — 10-19.99% — 20-29.99% — >30%
 Association strength with Instability in Driving (Marginal Effects): — < 0.0999 — 0.1000-0.1999 — 0.2000-0.2999

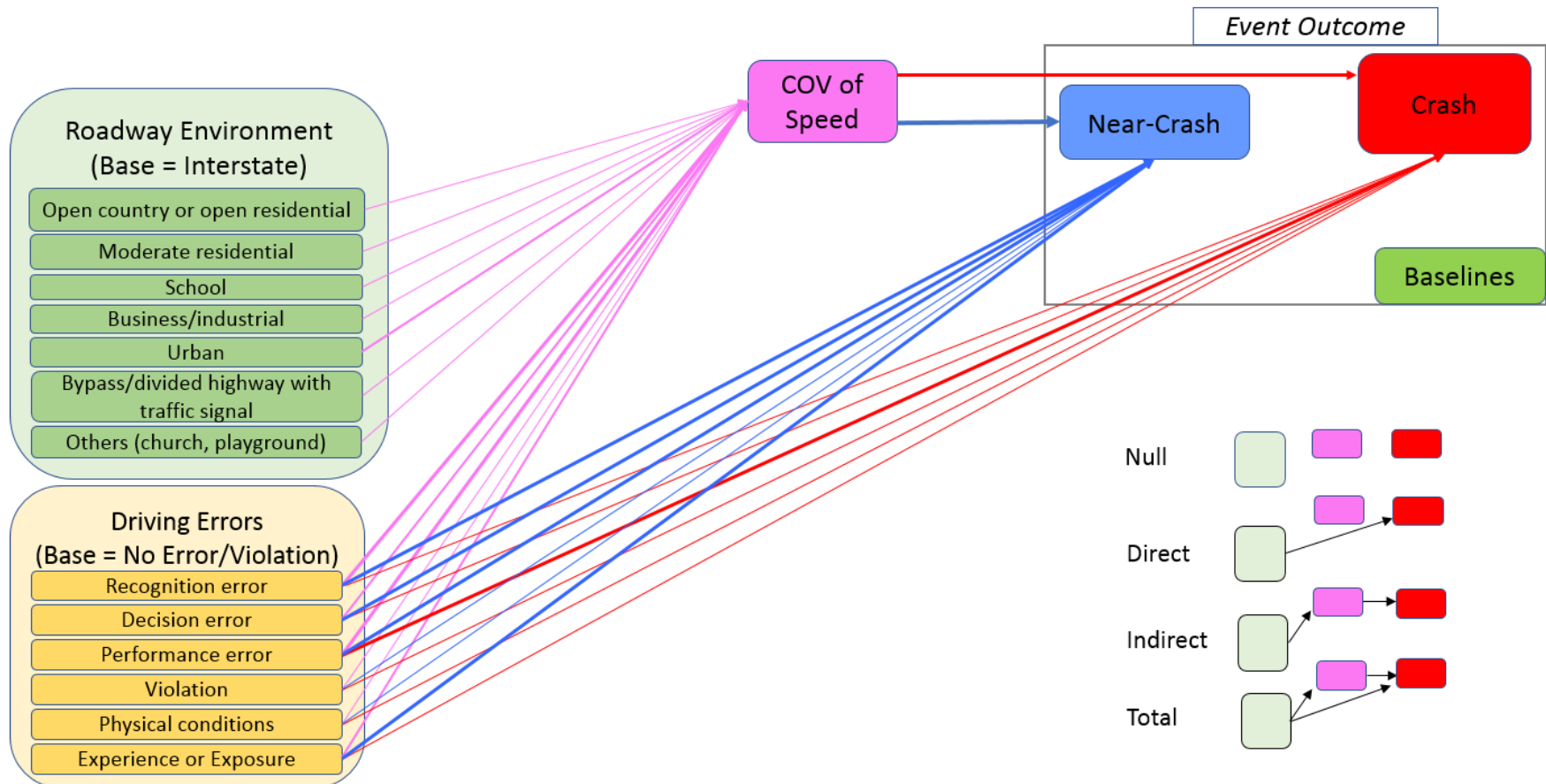


Figure 4. 5 Pathways to SCEs through COV of Speed (based on Joint Estimation Analysis)

Note: Tobit regression is used to model instability in driving while Ordered Probit regression is used to model Event Outcome. The two models are estimated in a joint fashion, where $\rho_{12} = -0.4112$ ($t - stat = -12.97$) indicates that residuals in the two equations have a significant negative correlation. All the pathways (including direct and indirect) are based on marginal effects (%) from the two constituent models. The color of a pathway is similar to the box color of the outcome to which the pathway leads.

Table 4. 6 Path Analysis Results (Based on Joint Estimation): Effects of Duration of Distraction on SCEs

Secondary Task Duration (seconds)	Effects of Distraction Duration on Crash				Effects of Distraction Duration on Near-Crash			
	Direct Effects of Secondary Task Duration on Crash	Effects of Secondary Task Duration on Instability in Driving	Indirect Effects of Secondary Task Duration on Crash through Instability in Driving (%)	Total Effects of Secondary Task Duration on Crash	Direct Effects of Secondary Task Duration on Near-Crash	Effects of Secondary Task Duration on Instability in Driving	Indirect Effects of Secondary Task on Near-Crash through Instability in Driving (%)	Total Effects of Secondary Task on Near-Crash
	Predictive MEs (%)	Predictive MEs			Predictive MEs (%)	Predictive MEs		
0	13.51	0.230	4.56	18.08	7.29	0.230	7.60	14.89
1	12.76	0.227	4.51	17.27	6.86	0.227	7.51	14.37
2	12.27	0.226	4.48	16.75	6.58	0.226	7.46	14.04
3	12.01	0.226	4.48	16.48	6.43	0.226	7.45	13.88
4	11.96	0.227	4.50	16.46	6.40	0.227	7.49	13.89
5	12.13	0.229	4.55	16.67	6.50	0.229	7.57	14.06
6	12.52	0.233	4.62	17.14	6.72	0.233	7.68	14.41
7	13.16	0.237	4.71	17.87	7.09	0.237	7.85	14.93
8	14.06	0.244	4.84	18.90	7.61	0.244	8.05	15.66
9	15.29	0.251	4.98	20.27	8.34	0.251	8.29	16.63
10	16.88	0.260	5.15	22.03	9.30	0.260	8.58	17.88
11	18.90	0.270	5.35	24.25	10.56	0.270	8.91	19.47
12	21.39	0.281	5.57	26.96	12.22	0.281	9.28	21.49
13	24.37	0.293	5.82	30.19	14.38	0.293	9.69	24.07
14	27.77	0.307	6.09	33.86	17.23	0.307	10.14	27.37
15	31.36	0.322	6.39	37.75	20.98	0.322	10.64	31.62
16	34.68	0.338	6.72	41.39	25.90	0.338	11.18	37.08
17	37.02	0.356	7.06	44.09	32.29	0.356	11.76	44.05
18	37.51	0.375	7.44	44.95	40.37	0.375	12.38	52.75

interstates. The Tobit model also provides useful insights about how other factors related to traffic flow conditions (roadway configurations), intersection influence, and level of service affect the instability in driving. All these parameters showed intuitive signs. For instance, we found that compared to physically separated (through median strip or barrier) divided traffic conditions, the instability in driving increases by 0.0211, 0.0283, and 0.1660 if a subject driver is traveling in undivided traffic or roadway segments (which is simply undivided or with two-way-left-turn lane), a roadway with one-way traffic, and roadway with no lane marking respectively. All these findings were expected because compared to physically divided traffic flow, undivided, roadway segments with one-way traffic, and roadway segments with no lane marking can be more complicated. Furthermore, the Tobit model results suggest that compared to no intersection influence, the instability in driving increases by 0.1530, 0.1301, and 0.0804 at stop sign/traffic signal, uncontrolled intersection, and interchange respectively.

Referring to the event outcome model, several useful insights were found. Results from the event outcome model reveal that compared to no driving errors and violations, all types of driving errors and violations increase the probability of both crashes and near-crashes. Performance errors showed the strongest positive correlation with crash risk and near-crash risk, followed by experience errors, decision errors, and recognition errors. For instance, performance error was found to have the highest impact on both crashes and near-crashes which increases their probabilities by 0.1110 and 0.1847 respectively. Furthermore, results indicate that probability of crash increase by 0.0996, 0.0795, and 0.0654 units if a subject driver makes experience error, decision error, and recognition error, respectively. Violations are also found to induce crash risk. All these findings are intuitive, suggesting that a subject driver that makes various errors and violations has a higher risk of getting involved in SCEs (Khattak *et al.* 2021a). It is interesting to mention that results of the event outcome model suggest that chance of crash increases by 0.1985 units with a unit increase in instability in driving. Similar findings are revealed in previous studies that quantify the potential safety risk associated with various volatility indices (Wali and Khattak 2020, Yeo *et al.* 2020).

To interpret the non-linear relationship, this study discusses the predictive margins of the duration of distraction obtained in the two models which were jointly estimated. Results of the Tobit model reveal that instability in driving increases with an increase in distraction duration. The results indicate that while the driver is distracted from 0-4 seconds, the predicted instability in driving is almost within a similar range (0.226 to 0.227) which increases in a non-linear way with an increase in distraction duration, reaching a value of 0.375 if the subject driver is distracted for 18 seconds (Figure 4.6). Previous studies also suggest a positive association of distraction duration with instability in driving (Arvin and Khattak 2020).

Referring to the effects of distraction duration on SCEs, Figure 4.7 illustrates how distraction duration relates to the chance of SCEs using the predictive margins (MEs, when multiplied with 100 can be interpreted as chance) obtained for the distraction duration in the ordered Probit model. We found that the chance of a crash increases exponentially when distraction duration exceeds 8 seconds (for details, please refer to Table 4. 6). These findings suggest a higher chance of crash if the subject driver is distracted for a longer duration, which was possible after accounting for the non-linear relationship of distraction duration with SCEs. For the effects of distraction duration on the chance of SCEs, the ordered Probit model suggests that the chance of a crash and near-crash increase by 19.85% and 33.04% respectively with a unit increase in instability in driving (Table 4. 4). Referring to the predictive margins, we found that the predicted chance of a crash is 0.38% when COV of speed is 0 which exponentially increases with an increase in the instability in driving reaching a value of 98.22% when instability in driving is 1.6 beyond which the chance of crash follows a constant trend (Figure 4.8). This finding highlights the higher crash risk at a higher instability in driving.

Referring to the results of path analysis which quantifies the effects of distraction duration on SCEs, we found that the total chance of crash and near-crash for a specific distraction duration could be significantly higher if its indirect effects on crash and near-crash through instability in driving are considered (Table 4.6). Based on the direct effects of distraction duration on a crash, the chance of crash exponentially increases with distraction duration if the latter exceeds 4 seconds (Table 4.6). For instance, the results from the path analysis reveal that the predicted chance of a crash is 12.52% if a subject driver is distracted for 6 seconds; however, if the indirect effects of distraction duration (through instability in driving) on the crash are considered, there is an additional 4.62% ($= 0.233 \times 0.1985 \times 100$) chance of a crash (Table 4.4). This indicates that the crash risk becomes significantly higher if the indirect effects of distraction duration on crashes through driving instability are considered. Moreover, based on the total effects of distraction duration on a crash, the probability of crash significantly increases with an increase in distraction duration. Similarly, results can be seen for the effects of distraction duration on near-crash through instability in driving. The results suggest significant statistical evidence ($\rho = -0.4112$; $t\text{-stats} = -12.97$) about unobserved factors which could significantly influence both the instability in driving and event outcome. The negative sign of correlation parameter (ρ) shows that there are unobserved factors that significantly increase instability in driving and decrease crash risk or vice versa. One example of such unobserved factors could be driver age – compared to older drivers, younger drivers could have higher instability in driving; however, their chance of involving in a crash could be lower than older drivers (possibly due to better mental and physical capabilities of younger drivers). It should be noted that while younger drivers may have lower

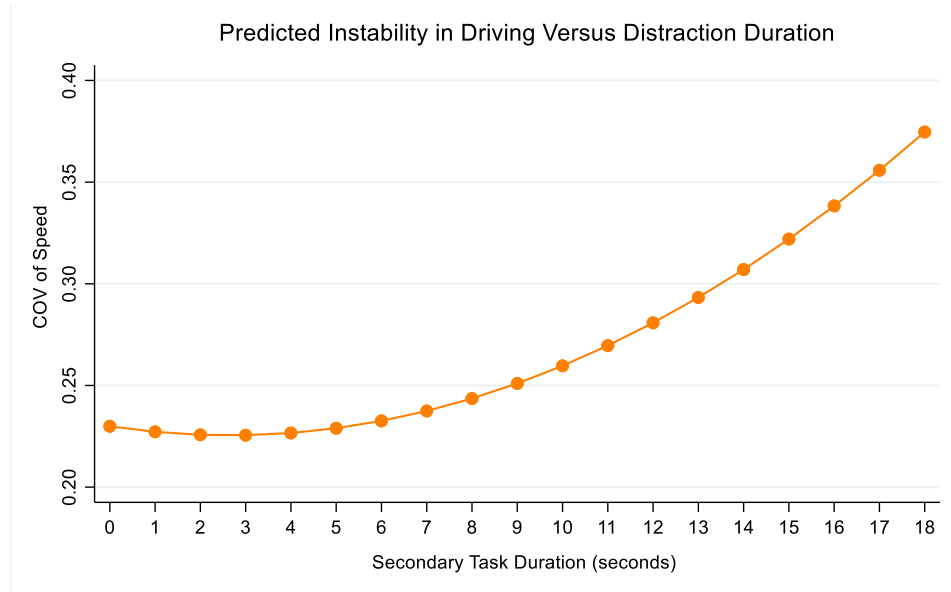


Figure 4. 6 Predicted Instability in Driving versus Duration of Distraction

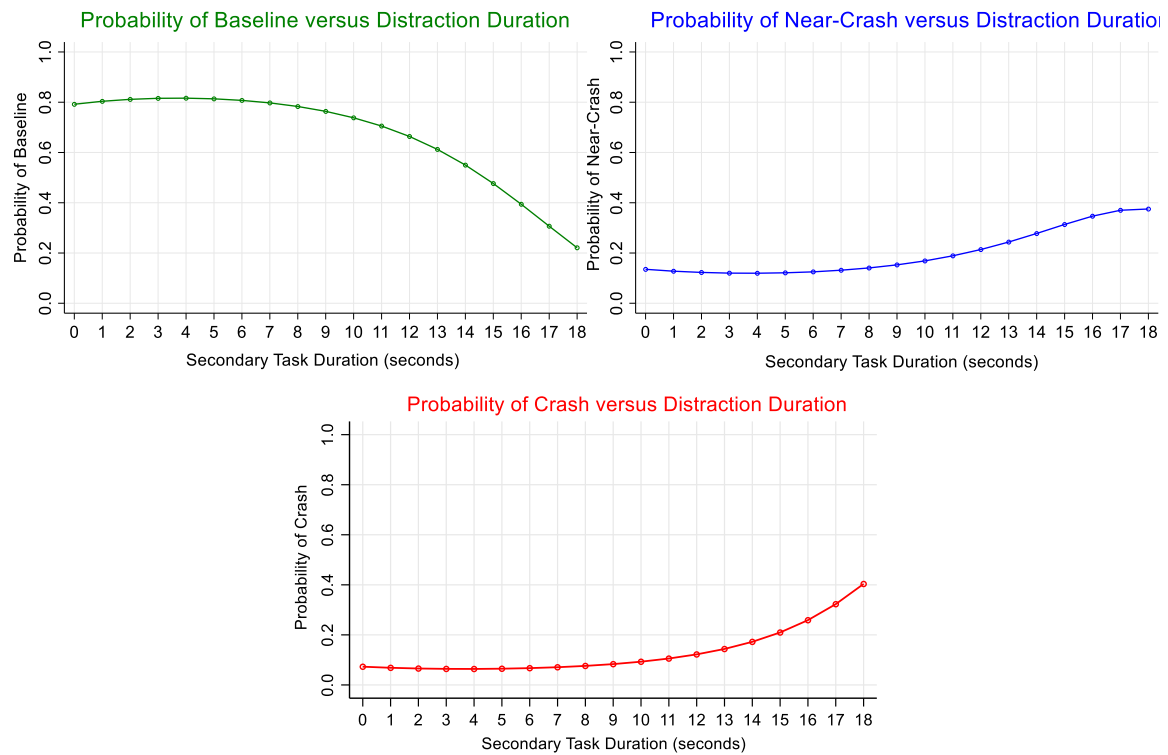


Figure 4. 7 Probabilities of Safety-Critical Events versus Distraction Duration

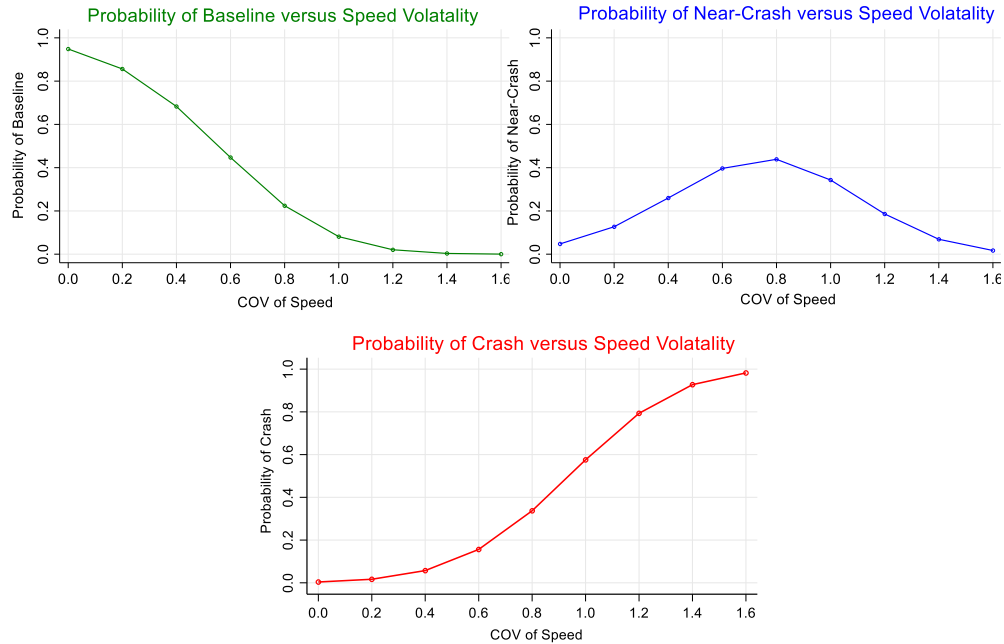


Figure 4. 8 Probabilities of Safety-Critical Events versus Instability in Driving

Note: The above curves are computed till the COV of speed is 1.6, beyond which the probability of baseline, near-crash, and crash follows almost a constant trend. Note that the maximum COV of speed based on the SHRP2-NDS subsample is 3.048 (Table 4.1).

crash risk (instead of higher instability in driving), their chance of crashes resulting in severe injury fatality could be higher.

To summarize the discussion, the key finding is that driving errors, violations, and duration of distraction increases the chance of a crash and a near-crash not only directly, but also indirectly through instability in driving. Furthermore, various roadway environments (urban areas, business/industrial locations, and school zones) and intersection influence (interchange, stop sign/traffic signal, and uncontrolled intersection) increase instability in driving, which in turn increases the chance of SCEs. For brevity and understanding, we illustrate the key results through infographics as shown in Figure 4.9. The images shown above are to illustrate the three roadway environments which are taken from various sources (for picture courtesy, we have provided links to pictures/sources below the references).

Practical Applications

For reducing one or more driving errors, multiple vehicle technologies including the forward collision warning systems, lateral vehicle detection system, adaptive cruise control, and blind-spot detection system can be helpful (Ahmad *et al.* 2021; (Khattak *et al.* 2021a). Similarly, a dilemma zone mitigation system has the potential to reduce violations (Khattak *et al.* 2021a). Potential strategies to reduce the duration of distraction and driving instability are reflected in Figure 4.10. Referring to the practical implications, removing external sources of distraction and educating drivers through seminars and campaigns about the negative consequences of distraction activities can reduce distracted driving. Furthermore, instead of using mobile phones, the use of hands-free technologies including basic phone use, voice-activated controls, and virtual assistants (artificial intelligence) technologies have the potential to reduce distracted driving. Note that recognition errors including distracted driving have contributed to around 38% of crashes. As proactive countermeasures, both fixed and dynamic message signs about distracted driving can help in reducing both distracted driving and driving instability. Using adaptive cruise control, drivers may travel at a safer speed thus reducing driving instability (speed volatility).

Limitations

This study used a subsample of SHRP2-NDS data which is based on specific geographical locations and may not represent other parts of the U.S. Hence, results obtained via this study may not be extended to other locations within the U.S. or abroad as driving behavior, socio-economic, and roadway environment may significantly differ. The information related to roadway environments is coded in a single variable, “locality,” in SHRP2-NDS data, which partially satisfies the research needs (Khattak *et al.* 2021a). In the future, it would be good if SHRP2-NDS provides complete data on the roadway and locality classifications including land use (urban, suburban, and rural), activity centers (playground, school, church, and especially

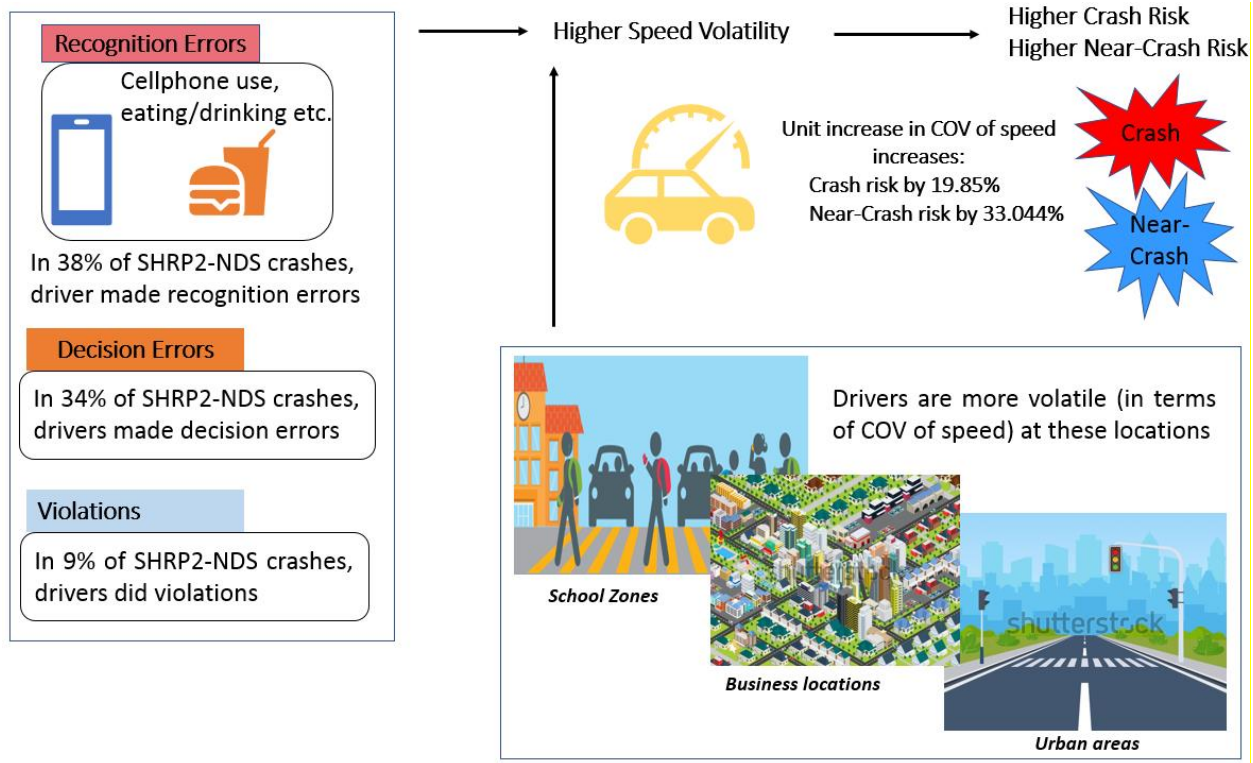


Figure 4. 9 Illustration of Key Findings through Infographics

Note: For pictures used in Figure 4.9, please see the list of the links to the sources below.

- 1). <https://www.safekids.org/infographic/school-zone-infographic>;
- 2). <https://www.shutterstock.com/image-vector/illustration-big-city-isometric-very-many-1995347036>;
- 3). <https://www.shutterstock.com/image-vector/city-road-on-crosswalk-traffic-lights-1947554338>.

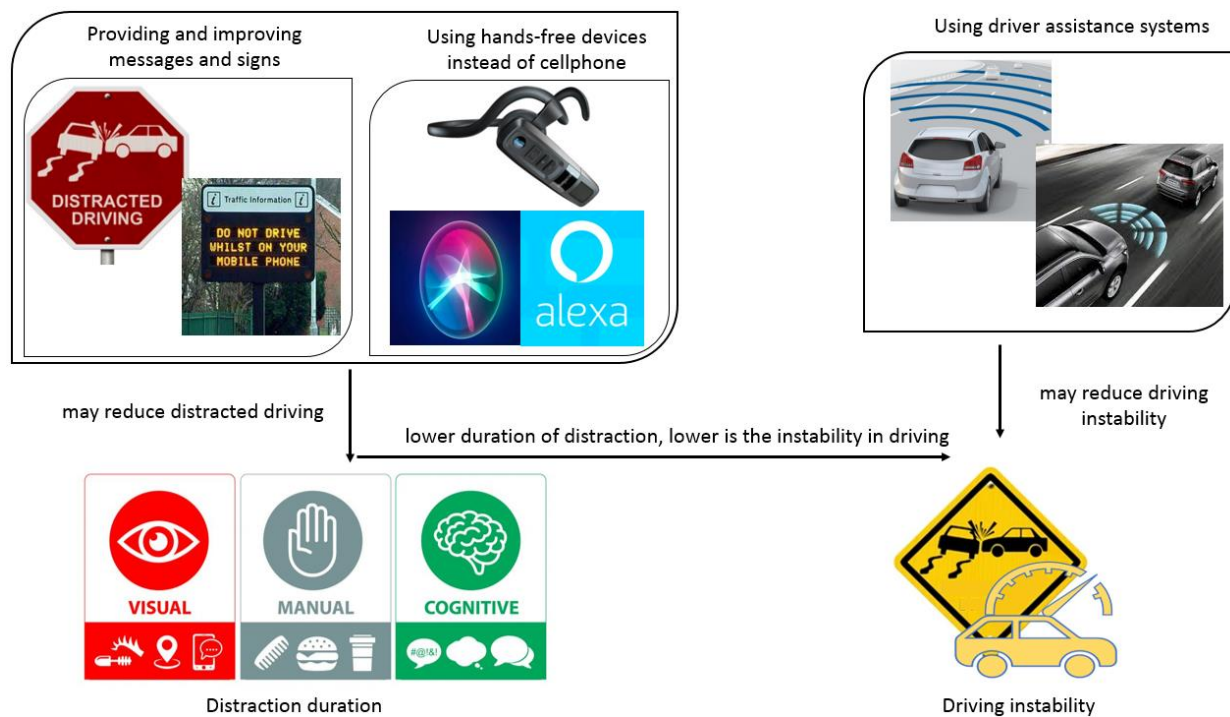


Figure 4. 10 Potential Strategies and technologies for reducing driving distraction and instability

Note: For picture courtesy, please see the links to all pictures shown in Figure 4.10 below.

<https://www.shutterstock.com/image-photo/no-distracted-driving-sign-red-stop-206848018>;

<https://www.pinterest.com/photonplayGroup/vms-variable-message-signs/>

<https://www.bestbuy.com/site/shop/bluetooth-headset-for-phone>

<https://www.cnet.com/tech/mobile/apple-listens-to-some-siri-recordings-to-make-it-better/>

<https://wallpapersafari.com/w/XEGOsF>

<https://www.bosch-mobility-solutions.com/en/solutions/assistance-systems/adaptive-cruise-control/>

<https://pmrpressrelease.com/forward-collision-warning-systems-market-industry-perspective/>

<https://www.supaquick.com/news-article/730/distracted-driving-a-major-cause-of-road-accidents>

<https://www.shutterstock.com/search/distracted+driving+sign>

categorizing industrial and business in separate categories), and roadway functional types (local roads, collectors (major/minor), arterials (major/minor), expressways, and interstate) (Khattak *et al.* 2021a). Furthermore, the SHRP2-NDS subsample used in this research does not include data on important variables like driver demographics (e.g., driver gender, age, and education). In the future, with the availability of such data, it would be good to see how driver age, gender, and education relate to speed volatility and SCEs. Furthermore, this study has used COV of speed as measures of instability in driving as an intermediate variable in the path analysis. Note that other volatility indices (for acceleration, deceleration, jerk) could be important. This study has tried the above indices as an intermediate variable in the joint path analysis along with COV of speed; however, due to collinearity issue of COV of speed with other volatility indices e.g., COV of acceleration and COV of deceleration significantly disturbed the estimation results. Therefore, this study can be replicated by using other volatility measures/indices (instead of COV of speed) as an intermediate variable in the joint path analysis to see how distraction relates to SCEs directly and through other volatility measures.

Conclusions

The paper contributes by jointly estimating models that explore potential pathways leading from errors and violations in diverse roadway environments to safety-critical events through instability in driving. The SHRP2-NDS subsample is harnessed, which includes 80.0% (N = 7,393), 13.4% (N = 1,237), and 6.6% (N = 607) of baselines, near-crashes, and crashes, respectively. Statistics reveal that drivers in crashes and near-crashes were, on average, distracted for longer durations (3.61 and 3.33 seconds, respectively) compared to baselines (1.75 seconds). Furthermore, statistics reveal that driving errors and violations have contributed to 92.42% of crashes where the contribution of recognition errors (37.89%), decision errors (33.77%), and violations (9.39%) was found to be predominant. The joints estimation showed that unobserved factors in the two models estimated for path-analysis are significantly correlated. The COV of speed is associated with increases with all five driving errors and violations. Performance errors showed the strongest positive correlation with crash risk, followed by experience errors, decision errors, and recognition errors. For instance, COV of speed is associated with increases by 0.2524, 0.2429, and 0.1581 units if a subject driver makes a recognition error, performance error, and experience error, respectively. Furthermore, instability in driving is significantly higher in urban areas (by 0.1652 units), business/industrial locations (by 0.0892 units), and school zones (by 0.0820 units) compared to driving on interstates. Importantly, we found that a unit increase in COV of speed increases the crash risk and near-crash risk by 19.85% and 33.04%, respectively. Importantly, our findings suggest significant evidence that both instability in driving and the probability or chance of SCEs increases non-linearly with distraction duration. Path analysis results indicate that all driving errors and violations not only contribute to SCEs

directly but also indirectly through driving instability. For instance, our findings reveal that recognition error (compared to no driving error and violation) increases the crash risk by 6.54% (direct effects); however, it further increases the crash risk (through COV of speed) by 5.01% leading to an overall increase in crash risk by 11.55%. Driving errors and violations can be riskier as they not only increase the chance of SCEs directly but also indirectly through instability in driving. Referring to the total effects of distraction duration on SCEs obtained through path analysis (via predictive margins), the chance of both crash and near-crash significantly increases with an increase in distraction duration beyond 4 seconds. For instance, the overall chance (after including the indirect effects of distraction duration through driving instability) of a crash is 16.46% for a driver distracted for 4 seconds, which increases to 22.03% if a driver is distracted for 10 seconds. From a practical implications standpoint, we consider that multiple vehicle technologies (e.g., forward-collision warning system, adaptive cruise control, lane tracking system, and lateral vehicle detection system) have can reduce one or more driving errors. Similarly, dilemma zone mitigation systems have the potential to reduce a significant percentage of violations. To mitigate distracted driving and duration of distraction, we consider that instead of using mobile phones, the use of hands-free technologies including basic phone use, voice-activated controls, and virtual assistants (artificial intelligence) technologies have the potential to reduce distracted driving. Furthermore, carefully travel in the certain roadway and built environments can potentially reduce various driving errors. For instance, removing sources of external distraction (e.g., billboards) can help to reduce recognition errors and needs to be investigated further. From a policy standpoint, awareness campaigns and mandatory training programs for drivers can reduce performance errors and experience errors. Finally, there is a need to explore the collective impacts of the aforementioned countermeasures like multiple vehicle technologies, roadway (built-environment) changes, and policy interventions.

**CHAPTER 5. PREDICTING SAFETY-CRITICAL EVENTS USING DRIVER BASED
ON DRIVER BEHAVIORS AND PERFORMANCE: APPLICATION OF MACHINE
LEARNING METHODS**

This chapter is in preparation and will be soon submitted to an impact factor journal for publication review.

Abstract

Human factors contribute to almost 93% of road crashes. Previous studies have used traditional statistical models to understand how key human factors including driving errors, violations, distraction duration, and driving instability relate to safety-critical events (SCEs). While previous studies provide useful insights, some of the research gaps still exist which include 1) none of the studies has investigated the effects of all of the key four human factors on SCEs and 2) statistical models could provide useful inferential knowledge but have lower prediction accuracy. Bridging the research gaps, this study harnesses the unique naturalistic driving study (NDS) data obtained through Strategic Highway Research Program 2 (SHRP2) which includes real-world information about driving behavior and performance. After cleaning and preprocessing, a final subsample (N = 9,237) of SHRP2-NDS data was used which was split into training (70%) and test (30%) samples. First, an ordered Probit model was estimated to understand how key human factors relate to the event outcome including baseline, near-crashes, and crashes. Next, three machine learning (ML) methods including Naïve Bayes, K-Nearest Neighbors (KNN), and gradient boosting tree (GBT) classifiers were used to predict the event outcome. Results of the ordered Probit model suggest that chance of crash increases by 5.17%, 5.98%, and 1.16% due to a unit increase in the coefficient of variation of speed, deceleration, and negative jerk respectively. The results also indicate that chance of a crash and near-crash increases with an increase in distraction duration and if a subject driver makes errors and violations. The GBT classifier showed the highest (91.23%) out-of-sample prediction accuracy. For predicting SCEs, the GBT classifier revealed the highest out-of-sample prediction accuracy. The practical applications of the study are discussed in the paper.

Introduction and Literature Review

Road traffic crashes are one of the key causes of fatalities and injuries in the United States. According to the National Highway Safety Administration (NHTSA), a total of 6,756,000 road crashes, including 33,244 and 1,916,000 as fatal and injury crashes respectively, have occurred across the United States in 2019 (NHTSA 2019). Statistics reveal that 36,096 were killed while 2,740,000 people were injured as a result of these crashes across the U.S. only in 2019 (NHTSA 2019). Due to significant socio-economic losses in road crashes, studies were carried out to understand the key causes of road crashes (Sabey and Staughton 1975, Treat *et al.* 1979b, Singh 2015). It is found that human factors contribute to more than 90% of road crashes (Sabey and Staughton 1975, Treat *et al.* 1979b, Singh 2015). To have a deeper understanding of the role of driving behavior in road crashes, studies have developed different classification frameworks to classify

driving errors and violations which could help in identifying appropriate countermeasures (Treat *et al.* 1979b, Reason *et al.* 1990, Wierwille *et al.* 2002, Iden and Shappell 2006, Stanton and Salmon 2009). One of the leading studies used the perception-reaction (PR) process to classify driver behaviors and found that recognition errors and decision errors were the two predominant errors in road crashes (Treat *et al.* 1979b). The above studies provided useful insights about the contribution of driving errors and violations in crashes, one key issue was that these studies were based on police crash reports which could be subjective. Numerous relevant studies have acknowledged that a police officer could not be present at the time of crash occurrence due to which the data about pre-crash driver behavior and performance could either be subjective (Dingus *et al.* 2006a, Hankey *et al.* 2016, Arvin *et al.* 2019a, Ahmad *et al.* 2020). Besides, there could be some issue of underreporting in the police crash report which have been acknowledged (Dingus *et al.* 2006a, Hankey *et al.* 2016, Arvin *et al.* 2019a, Ahmad *et al.* 2020).

On other hand, the unique naturalistic driving study (NDS) data collected through the Second Strategic Highway Research Program (SHRP2) includes real-world microscopic details about pre-crash driver behavior and performance which were obtained through fitted cameras, radars, other sensing equipment fitted in the subject (instrumented) vehicles (Dingus *et al.* 2006a, Hankey *et al.* 2016). The SHRP2-NDS includes data about near-crashes and baselines (where a subject driver was not involved in safety-critical events “SCEs”) which makes it unique from the traditional police crash data (Hankey *et al.* 2016, Arvin *et al.* 2019a, Ahmad *et al.* 2021, Khattak *et al.* 2021a). Note that baselines could be used as a surrogate to exposure for computing the risk of crash and near-crash (Hankey *et al.* 2016, Arvin *et al.* 2019a, Ahmad *et al.* 2021, Khattak *et al.* 2021a). One unique aspect of the SHRP2-NDS data is that it includes data about instantaneous speed, acceleration, and deceleration which could be used to access the driving instability (volatility) by computing various indices of driving volatility (Wali *et al.* 2018c, Arvin *et al.* 2019a, Wali *et al.* 2019, Wali and Khattak 2020). Note that driving volatility is defined as the deviation from the normal value of the specific parameter like speed, acceleration, and deceleration (Wali *et al.* 2018c, Arvin *et al.* 2019a, Wali *et al.* 2019, Wali and Khattak 2020). For driving volatility, studies have used various volatility indices including coefficient of variation (COV), quartile coefficient of variation, standard deviation, time-varying stochastic volatility of speed, acceleration, deceleration, and jerk (Arvin *et al.* 2019a). To capture driving distraction, the SHRP2-NDS include data about types (e.g., cellphone texting, cellphone browsing, cellphone talking, talking to a fellow passenger, and eating/drinking while driving) and duration of secondary tasks due to which driver attention could have deviated from the driving task (Dingus *et al.* 2006a, Hankey *et al.* 2016). We also note that the SHRP2-NDS data also includes information about the roadway and environment which could be used to understand how these factors may relate to SCEs including crashes and near-crash compared to baseline.

Utilizing the SHRP2-NDS subsample, one of the recent studies was conducted to understand the role of human factors in road crashes and it was found that human factors still contribute to almost 93% of road crashes (Khattak *et al.* 2021a). The study also proposed a systematic taxonomy of naturalistic driving errors and violations (TDEV) to understand the nature and role of various driver behaviors in crashes (Khattak *et al.* 2021a). Based on the TDEV, driving behaviors could be classified into 1) Recognition errors, 2) Decision errors, 3) Performance errors, 4) Physical conditions-related errors, 5) Experience or exposure errors, and 6) Violations (Khattak *et al.* 2021a). Among all the driving errors and violations, recognition errors were found to be the predominant errors that contributed to 38.63% of road crashes, followed by decision errors (34.32%) and violations (9.06%) (Khattak *et al.* 2021a). Building upon the TDEV, another study has recently applied path analysis while estimating statistical models to understand how roadway and environmental factors may relate to SCEs both directly and through driving errors and violations (Ahmad *et al.* 2021). This study provided some useful insights. For instance, it was found that specific roadway environments (e.g., urban areas) significantly increases the crash risk not only directly but also through recognition error, decision errors, and performance error as the chance of each of these driving errors was significantly higher in urban areas (Ahmad *et al.* 2021). that The findings suggest that there is a 6.44% higher crash risk in urban areas; where the urban area can also induce an additional crash risk of 2.16% through recognition errors due to which the total increase in crash risk in urban areas could be 8.60% (Ahmad *et al.* 2021). These studies provide useful insights about the effects of the roadway and environmental factors and driving errors and violations on SCEs (Ahmad *et al.* 2021, Khattak *et al.* 2021a). However, besides roadway environment and driving errors and violations, we consider that driving volatility (instability) could significantly relate to SCEs which was not captured in the aforementioned study. To this effect, some studies have investigated the role of driving instability in SCEs (Wali *et al.* 2019, Wali and Khattak 2020); however, these studies have not included the naturalistic driving errors and violations in the analysis which could significantly induce SCEs as discussed earlier. The aforementioned studies reveal that volatile or unstable drivers have a higher chance of crash and near-crash (Wali *et al.* 2019, Wali and Khattak 2020). Referring to the effects of driver distraction (seconds) on SCEs, studies have found that the duration of distraction positively correlate with SCEs (Ahmad *et al.* 2021).

In summary, the above discussion shows that 1) the role of human factors including driving errors and violations in SCEs was investigated, 2) the effects of driving instability on SCEs were quantified, 3) the role of distraction duration in SCEs was investigated, and 4) the effects of the roadway and environment factors on SCEs were explored. However, some of the research gaps still exist. First, the effects of each of the driving errors and violations, driving instability, distraction duration, and roadway environmental factors were investigated in separate studies and none of the studies (to the best of authors' knowledge)

looked at the role of all these factors in SCEs together. Based on this, we consider that the effects of overall human factors including driver behavior and performance on SCEs could have not been unveiled. Second, all of the above studies used standard statistical models to understand how one or two of the key human factors relate to SCEs. We note that while statistical models may provide useful inferential knowledge about the correlation of human factors with SCEs, these models have lower prediction accuracy. On other hand, machine learning (ML) or Artificial intelligence (AI) methods show promising performance in out-of-sample prediction. Bridging the aforementioned research gaps, the key focus of this study is to apply various ML/AI methods to predict SCEs based on all of the available human features including driving errors, violations, driving instability, and distraction duration. Note that this study also considers some of the key control factors including intersection influence and roadway environments in the analysis to investigate the risk of crash and near-crash in diverse roadway environments. First, the study estimates an ordered Probit model to understand how human factors relate to the event outcome (baseline, near-crash, and crash). Next, the study applies three key MI methods including Naïve Bayes, K-Nearest Neighbor (KNN), and gradient boosting tree (GBT) classifiers to predict the event outcome based on the human factors. After training the statistical model and machine learning methods to the training data, the study accesses the out-of-sample prediction accuracy for each model to select the most accurate model/method which could predict the event outcome using the real-world information about human factors in the SHRP2-NDS data.

Methodology

Study Framework

Statistical models can provide useful inferential knowledge about the key relations in the data; however, these models usually show weaker performance in prediction compared to ML methods. While acknowledging the strengths and weaknesses of the two methods, studies show that the decision of preferring one over the other is based on the scope of the study. Note that the key aim of this study is to accurately predict the event outcome (baseline, near-crash, and crash) using human factors including driving errors, violations, driving instability, and distraction duration while controlling for key roadway environmental factors. To achieve the primary objective, this study applies more sophisticated supervised ML methods including NB, KNN, and GBT classifiers to predict the event outcome. In addition, the study also estimates the traditional statistical model like ordered Probit model not only to spot the key relations in the data but also to consider this statistical model as a base to understand the amount (percentage) of change (increase or decrease) in the out-of-sample prediction accuracy of each of the ML methods. The study framework is shown in Figure 5.1. For a brief discussion about each of the methods used in this study including statistical model and ML methods, please refer to the subsections below.

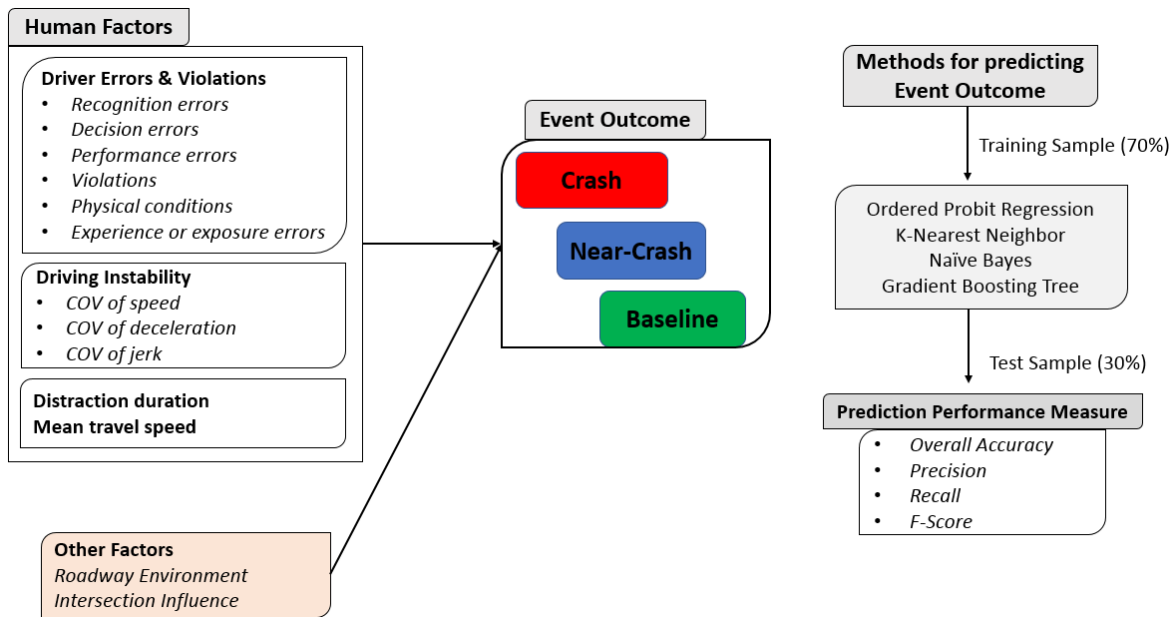


Figure 5. 1 Study Framework

Ordered Probit Regression

Given the ordinal nature of the event outcome (including baseline, near-crash, and crash), this study estimates an ordered Probit model to model the event outcome which can be given below (Kockelman and Kweon 2002, Abdel-Aty 2003, Rifaat and Chin 2007, Arvin *et al.* 2019b).

$$Y = \beta_0 + \beta_i X_i + \varepsilon \quad (1)$$

Where Y indicates the event outcome, β_0 shows the constant term, β_i is the set of parameters to be estimated for the set of associated factors X_i , and ε indicates a set of errors or unobserved factors with normal distribution $\sim N(0, \sigma^2)$. To understand the change (increase/decrease) in the probability of a specific outcome (baseline, near-crash, and crash) with a unit increase (from its mean value) in a specific continuous explanatory variable while keeping all other variables in the model at their mean values, we compute marginal effects (MEs) which can be given as below (Kockelman and Kweon 2002, Arvin *et al.* 2019).

$$\frac{\partial P(Y_2^* = l)}{\partial X} = -[\Phi(\mu_l - \beta_2 X_i) - \Phi(\mu_{l+1} - \beta_2 X_i)]\beta_2 \quad (2)$$

For the marginal effects of an indicator or dichotomous variable i.e., when it switches from 0 to 1, the MEs can be computed as below (Washington *et al.* 2010).

$$\partial P(Y_2^* = l|X) = P(Y_2^* = l|X_i = 1) - P(Y_2^* = l|X_i = 0) \quad (3)$$

The final ordered Probit model is systematically derived while including the most important variables statistical significance, specification parsimony, and intuition. In doing so, the likelihood ratio test, Akaike information criterion “AIC” (Akaike 1974), and Bayesian information criterion “BIC” (Schwarz 1978) statistics were examined. In comparing the competing models, the one with lower AIC and BIC statistics is the best fit model after adjusting for degrees of freedom. Other performance measures can also be used to access the model performance depending on the study scope. For instance, BIC and Schwarz criterion “SBIC” are considered as more parsimonious criteria; however, SBIC outperforms the BIC (Nickelsburg 1985). For the sake of completeness, SBIC (Nickelsburg 1985), Consistent AIC “CAIC” (Bozdogan 1987), and Bozdogan’s Information Complexity “ICOMP” (Bozdogan 1990) is considered (where applicable) to select the final model while comparing it to its counterparts. For details about various criteria used to access the model performance, please refer to the studies in parenthesis (Akaike 1974, Schwarz 1978, Nickelsburg 1985, Bozdogan 1987, 1990).

Naïve Bayes Classifier

The naïve Bayes classifier is a simple and more cost-effective probabilistic algorithm that is based on the Bayes' theorem and conditional independence (naïve) assumption (Dissanayake and Lu 2002, Hlaing *et al.* 2019, Jha *et al.* 2021). The NB classification has shown significant improvement in prediction performance compared to the traditional statistical models (Dissanayake and Lu 2002, Khera and Singh 2015, Hlaing *et al.* 2019, Jha *et al.* 2021). This method is especially found more promising for larger datasets. To predict a specific class of a response variable (in this case, the event outcome including base, near-crash, and crash) using a set of predictors (features), the NB classifier can be expressed as below (Dissanayake and Lu 2002, Hlaing *et al.* 2019, Jha *et al.* 2021).

$$P(Y|X_1, X_2, \dots, X_n) = \frac{P(Y) * P(X_1, X_2, \dots, X_n|Y)}{P(X_1, X_2, \dots, X_n)} \quad (4)$$

Where;

- $P(Y|X_1, X_2, \dots, X_n)$ shows the posterior probability of Y conditioned on predictors $(X_1, X_2, \dots, X_n)$.
- $P(Y)$ indicates the prior probability of Y .
- $P(X_1, X_2, \dots, X_n|Y)$ shows the posterior probability of predictors $(X_1, X_2, \dots, X_n)$ conditioned on Y .
- $P(X)$ indicates the prior probability of predictors $(X_1, X_2, \dots, X_n)$ conditioned on Y .

The prior information in the NB classifier enhances the learning and consequently prediction performance. One basic way is to train the NB classifier with apriori is to use the probabilities of various categories of the response variable in the data. To further enhance the NB classifier, studies suggest using a grid-search optimization procedure to determine the optimal value of the tuning parameter i.e., kernel width.

K Nearest Neighbor Classifier

The KNN classifier is another non-parametric method that can be used to predict the response outcome (in this case, baseline, near-crash, and crash) for a specific observation based on the nearest (closest) k observations (Cover and Hart 1967, Peterson 2009). The KNN assigns a specific class to an observation to which most of the k observations belong (Soucy and Mineau 2001, Peterson 2009, Qin *et al.* 2013, Iranitalab and Khattak 2017, Jha *et al.* 2021). In the KKN algorithm, the appropriate number of k nearest samples or neighbors from the training sample are selected which can be used to predict a specific class (Soucy and Mineau 2001, Qin *et al.* 2013, Iranitalab and Khattak 2017, Jha *et al.* 2021). The two important decisions in implementing the KNN classifier include the distance function and the value of k i.e., number of nearest neighbors. For determining the optimal value of k , a grid-search optimization procedure, to

increase prediction accuracy and reducing the loss function, is generally used (Khera and Singh 2015, Iranitalab and Khattak 2017). For the distance function, this study uses various distance types to select the optimal number of nearest neighbors and specific distance that has the lowest loss (lowest classification error).

Gradient Boosting Trees Classification

In the GBT classifier, a set of predictors (X_i) can be used to predict the multiple categories of the response outcome "Y" via function $\tilde{F}(X)$ (Semanjski and Gautama 2015). The GBT classifier aims to predict the response outcome via function $\tilde{F}(X)$ while minimizing the loss function " $L(Y, F(X))$ " for the joint distribution of all possible combinations of predictors and response outcome (X_i, Y) as given below (Semanjski and Gautama 2015).

$$\tilde{F}(X) = \underset{F}{\operatorname{argmin}} E_X [E_Y(L(Y, F(X)))|X] \quad (6)$$

If the function $F(X)$ is restricted to parameterized class function " $F(X, P)$ ", P indicates a set of finite parameters identifying individual class members, the function optimization problem is changed into parameter problem (Shafique and Hato 2015).

$$\tilde{F}(X) = F(X, \check{P}) \quad (7)$$

Where $\check{P}$ is computed as a summation of the initial guess (P_0) and overall boosts (successive increment). The GBT classifier enhances the stability of the GBT while increasing the weights of the misclassified training events to form a new tree (Roe *et al.* 2005).

Performance Measures

To select the model/method with higher out-of-sample prediction accuracy, this study computes the confusion matrix the output of which is then used to calculate various performance indices based on the test sample. A sample confusion matrix for the three classes can be shown below. Where A, E, and I indicate the number of correctly predicted baselines, near-crashes, and crashes respectively (see Table 5.1).

In general, the out-of-sample prediction performance of a specific model/method is accessed via the overall out-of-sample prediction accuracy which can be computed as below.

$$Accuracy = \frac{\text{Total Observations correctly predicted as Baselines, Near-crashes, and Crashes}}{\text{Total Number of Observations}} * 100\% \quad (8)$$

To understand how accurately a model/method predicts a specific class (baseline, near-crash, and crash), the precision and recall can be computed both of which are context-specific and can be used depending on the scope. Since the precision and recall (formula shown in Table 5.1) are context-specific; therefore, we compute F_1 which is a more comprehensive performance measure of the prediction accuracy for a specific class as it takes into account both precision (False Positive) and recall (False Negative). To calculate the F_1 score for a specific model/model for the various classes, please see the equations below.

$$F_1 \text{ score (Baseline)} = \frac{2*A}{(2*A+B+C+D+G)} \quad (9)$$

$$F_1 \text{ score (Near - Crash)} = \frac{2*E}{(2*E+D+F+B+H)} \quad (10)$$

$$F_1 \text{ score (Crash)} = \frac{2*I}{(2*I+G+H+C+F)} \quad (11)$$

Data

Data Source

This study analyzes a subsample of SHRP2-NDS data which provides unique and real-world data on pre-crash driving behaviors, distraction (secondary task) types and duration, vehicle kinematics including speed, acceleration, decelerations, and other key variables related to roadway and environment (Dingus et al. 2006, Hankey *et al.* 2016). The second-to-second data on driver behavior, driver performance, and eye glance of the subject drivers were recorded through recording equipment like video cameras (Dingus et al. 2006, Hankey *et al.* 2016). Data about the instantaneous speed, deceleration, and acceleration can be used to compute various volatility measures for the subject drivers (Dingus et al. 2006, Hankey *et al.* 2016). These volatility measures include COV, standard deviation, and quantile coefficient of variation of speed, acceleration, deceleration, and jerk (Arvin *et al.* 2019b). In the SHRP2-NDS, data was carefully coded from the recorded videos by the data reductionist in the Virginia Tech Transportation Institute (VTTI) who following a thorough procedure to ensure the quality data extraction (Dingus et al. 2006, Hankey *et al.* 2016). Note that the availability of baselines in the SHRP2-NDS data provides an opportunity to compute crash risk and near-crash risk which is not possible in police crash data. The police crash data can also be subjective especially for pre-crash driving behaviors and performance (Ahmad *et al.* 2021, Khattak *et al.* 2021a). To understand how SHRP-NDS defines baseline, near-crash, and crash, please see below.

- **Baseline:** In the SHRP2-NDS, the baselines were selected for the sake of comparison with SCEs (Dingus *et al.* 2006a, Hankey *et al.* 2016). Note that the typical duration of baseline in the SHRP2-

Table 5. 1 Sample Confusion Matrix for Baselines, Near-Crashes, and Crashes

Confusion Matrix		Predicted			False Negative (FN)	Recall
		Baseline	Near-Crash	Crash		
Observed	Baseline	A	B	C	B+C	$\frac{A}{(A + B + C)}$
	Near-Crash	D	E	F	D+F	$\frac{E}{(D + E + F)}$
	Crash	G	H	I	G+H	$\frac{I}{(G + H + I)}$
False Positive (FP)		D+G	B+H	C+F	<i>Overall Accuracy</i> $(A + E + I)$ $= \frac{\quad}{Total\ Observations}$	
Precision		$\frac{A}{(A + D + G)}$	$\frac{E}{(B + E + H)}$	$\frac{I}{(C + F + I)}$		

NDS was 20-30 seconds and the baselines were randomly selected by the team of data reductionists to select at least one baseline for a subject driver who was involved in an SCE (Dingus *et al.* 2006a, Hankey *et al.* 2016). We note that in the SHRP2-NDS data, the number of baselines for a specific subject driver was proportional to the overall driving time of that driver in the experiment of SHRP2-NDS (Dingus *et al.* 2006a, Hankey *et al.* 2016). One of the key criteria for selecting the baseline was that the vehicle speed (miles per hour) of the subject vehicle should not drop below 5 miles per hour for more than 2 seconds (Dingus *et al.* 2006a, Hankey *et al.* 2016). These lower speed baseline observations were excluded in the SHRP2-NDS to exclude the long stopping effects and to include only those baseline drivers who could have a potential risk of an at-fault crash (Dingus *et al.* 2006a, Hankey *et al.* 2016).

- **Near-Crash:** The near-crash refers to a situation where the subject driver, other drivers, cyclist, animal, or pedestrian was required to take a rapid evasive maneuver to avoid a crash (Dingus *et al.* 2006a, Hankey *et al.* 2016). Based on the SHRP2-NDS data dictionary, a near-crash should fulfill the four criteria which include 1) Not a crash 2) Not pre-meditated 3) Evasion required, and 4) Rapidity required (Dingus *et al.* 2006a, Hankey *et al.* 2016). For further details about the four criteria, please refer to (Dingus *et al.* 2006a, Hankey *et al.* 2016, Ahmad *et al.* 2021, Khattak *et al.* 2021a).
- **Crash:** The NDS crash refers to a situation where the subject vehicle had any contact with a fixed or moving object which could have led to the transfer/dissipation of kinetic energy. According to the NDS dictionary, situations including non-premeditated departures from the roadway i.e., if at least one tire leaves the paved/intended travel surface are also considered crashes

Human Factors

The key human factors considered in this study include driving errors, violations, driving instability, and distraction duration. Using information coded in the two variables including driver behavior and secondary task in the SHRP2-NDS data, this study uses the recently developed TDEV which is based on the theoretical perception-reaction (PR) process (Khattak *et al.* 2021a). In each phase (perception, recognition, decision, and performance) of the PR process, a subject driver could have made a specific error (Khattak *et al.* 2021a). For instance, in the recognition and decision phase, a subject driver could have made recognition errors and decision errors, respectively. Similarly, in the performance phase, a subject driver could have made performance errors while executing his/her intended action or could have violated the traffic regulations which include both intentional and unintentional violations (Khattak *et al.* 2021a). According to the TDEV, some of the subject drivers could have made physical condition-related errors or experience/exposure errors which might not directly relate to the PR process but could influence driving capabilities in all four key phases of the PR process (Khattak *et al.* 2021a). For further understanding about the TDEV and the key

types of naturalistic driving errors and violations, please refer to (Khattak *et al.* 2021a). Using the TDEV, this study categorizes the driving behaviors into 1) Recognition error, 2) Decision error, 3) Performance error, 4) Physical condition-related error, 5) Experience or exposure error, and 6) violation. For computing various indices of the volatility measures such as COV of speed, deceleration, acceleration, and jerk, please refer to (Arvin *et al.* 2019a). Note the acceleration/deceleration refers to the first derivate whereas jerk refers to the second derivate of the travel speed. The negative jerk refers to the derivative of deceleration i.e., the second derivate of the travel when the vehicle is decelerating and vice versa for the positive jerk.

Results

Descriptives

Before discussing the descriptive statistics of the key variables, we note that after cleaning and preprocessing the data, a total of 9,237 observations with complete information were considered for the analysis. To access the prediction accuracy of statistical model and machine learning methods, the data (N = 9,237) is randomly split into training (N = 6,466) and test samples (N = 2,771) with a ratio of 70:30 respectively. Descriptive statistics of the key variables in the training sample are provided in Table 5.2. The mean COV of speed, mean COV of deceleration, and mean COV of negative jerk is significantly lower in the baselines compared to their corresponding values in the SCEs. For instance, the mean COV of speed for baselines is found to be 0.142 which is significantly lower than its value in near-crashes (0.556) and crashes (0.642). Similarly, the mean duration of distraction in crashes and near-crashes is found to be 3.814 seconds and 3.333 seconds which are significantly higher than that in the baselines (1.745 seconds). For driving errors and violations, the statistics reveal that while baseline drivers were not involved in any SCEs, these drivers have still made almost 10% of the errors and violations which can lead to crashes and near-crashes in the future. On other hand, subject drivers have made errors and violations in 61.26% and 92.74% of the near-crashes and crashes, respectively. For details about the descriptive statistics of key variables, please refer to Table 5.2.

Empirical Results

Ordered Probit Model

Results of the ordered Probit model with the lowest performance measures (AIC, BIC, CAIC, SBIC, and ICOMP) among all the competing models is selected which includes the key variables and other important

Table 5. 2 Descriptive Statistics of Key Variables (N = 6,466)

Explanatory Variables	Baselines (N = 5,195)				Near-Crashes (N = 844)				Crashes (N = 427)			
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
Measures of Volatility												
COV of Speed	0.142	0.161	0.001	1.920	0.556	0.421	0.019	3.048	0.642	0.385	0.014	2.425
COV of Deceleration	0.752	0.232	0.000	2.110	1.245	0.401	0.171	2.661	1.049	0.359	0.000	2.584
COV of Negative Jerk	0.790	0.234	0.000	2.780	1.291	0.453	0.489	3.005	1.075	0.413	0.412	2.770
Secondary Task Duration (seconds)	1.745	2.142	0.000	8.960	3.333	3.889	0.000	15.683	3.814	4.184	0.000	18.263
Mean Travel Speed (miles per hour)	17.754	8.154	0.002	40.935	11.551	7.447	0.162	42.753	8.208	5.998	0.159	34.401
	Frequency		%		Frequency		%		Frequency		%	
Driving Errors												
No error	4,676		90.01		327		38.74		31		7.26	
Recognition Error	13		0.25		279		33.06		171		40.05	
Decision Error	138		2.66		120		14.22		144		33.72	
Performance Error	4		0.08		7		0.83		29		6.79	
Violation	297		5.72		89		10.55		39		9.13	
Physical Condition-related Error	63		1.21		17		2.01		7		1.64	
Experience or Exposure Error	4		0.08		5		0.59		6		1.41	
Intersection Influence												
No Intersection Influence	4,229		81.41		267		31.64		176		41.22	
Interchange Influence	112		2.16		79		9.36		15		3.51	
Stop Sign or Traffic Signal Influence	626		12.05		275		32.58		105		24.59	
Uncontrolled Intersection Influence	104		2		91		10.78		34		7.96	
Parking Lot or Driveway Entrance/Exit	99		1.91		98		11.61		87		20.37	
Other Intersection Influence	25		0.48		34		4.03		10		2.34	
Roadway Environment												
Urban area indicator (1/0)	80		1.54		69		8.18		32		7.49	
Moderate Residential (1/0)	1,165		22.43		124		14.69		89		20.84	

variables are shown in Table 5.3. To decide about including a specific explanatory variable in the final model, we have considered the statistical significance, intuition, and model parsimony. For the statistical significance of a specific variable, we have used 95% confidence level criteria. To understand the effect of a unit increase in a specific explanatory variable from its mean value (or if an indicator variable switches from 0 to 1) while keeping all other variables in the model at their mean values on the probability of a specific category of the response variable (baseline, near-crash, or crash), we have computed the MEs (when multiplied with 100 can be interpreted as chance). The results indicate that all human factors including driving instability measures (COV of speed, deceleration, and negative jerk), driving errors and violation (compared to no driving error and violation), and secondary task duration showed a statistically significant and positive association with both crash and near-crash. For a detailed discussion about the findings, please refer to the discussion section in this paper.

Naïve Bayes Classifier

First, this study applies a Naïve Bayes (NB) classifier while assuming equal prior probabilities of baseline, near-crash, and crash (0.333, 0.333, and 0.333 respectively) to the training sample. Next, the NB classifier is trained using informative prior probabilities for the baseline, near-crash, and crash (0.8034, 0.1305, and 0.0660 respectively) which were determined based on the distribution of the event outcome in the training sample (please see Table 5.4). Finally, a grid search optimization procedure was used to select the optimal kernel width for the NB classifier which is found to be 0.11215 at which the minimum objective loss (classification error) is obtained (Figure 5.2 and Figure 5.3). Comparing the resubstituting and cross-validation errors for the three models, it is found that the NB classifier with the optimal kernel width (0.11215) shows the minimum loss (misclassification error) compared to its counterparts (Table 5.4). To understand how driver instability and distraction duration relate to posterior probabilities of baseline and SCEs obtained through the optimal NB classifier in the training sample please see Figure 5.4. As COV of speed, deceleration, and negative jerk increases, the likelihood of both crash and a near-crash increase compared to baseline. The distraction duration also shows a positive association with the posterior probabilities of SCEs (compared to baseline) which suggests those subject drivers who are distracted for a longer duration have a higher chance of involving in a crash and near-crash.

KNN Classifier

For fitting the KNN classifier to the training sample, this study first uses a grid-search optimization procedure to select the optimal number of nearest neighbors based on a specific type of distance. To do this, various types of distances including spearman, seuclidean, Mahalanobis, Jaccard, hamming, Euclidean, cosine, correlation, chebychev, and cityblock were considered. Based on the gird-search, the Mahalanobis

Table 5. 3 Estimation Results of the Ordered Probit Model

Key Explanatory Variables (Predictors)	Coeff.	t-stats	Chance in % (= Marginal Effects*100)		
Measures of Volatility			Baseline	Near-Crash	Crash
COV of Speed	0.956	9.97	-11.74	6.57	5.17
COV of Deceleration	1.107	13.25	-13.59	7.61	5.98
COV of Negative Jerk	0.214	2.76	-2.63	1.47	1.16
Secondary Task Duration (seconds)	0.032	4.18	-0.39	0.22	0.17
Mean Travel Speed (miles per hour)	-0.022	-4.96	0.26	-0.15	-0.12
Driving Errors (Base = no driving error)					
Recognition Error	1.681	22.49	-20.64	11.55	9.08
Decision Error	1.867	26.13	-22.93	12.84	10.09
Performance Error	2.544	11.91	-31.24	17.49	13.75
Violation	1.009	13.29	-12.39	6.94	5.45
Physical Condition-related Error	0.978	6.19	-12.01	6.72	5.29
Experience or Exposure Error	1.781	5.87	-21.86	12.24	9.62
Intersection Influence (Base = no intersection influence)					
Interchange Influence	0.785	7.76	-9.64	5.39	4.24
Stop Sign or Traffic Signal Influence	0.247	4.03	-3.03	1.70	1.33
Uncontrolled Intersection Influence	0.593	6.38	-7.28	4.07	3.20
Parking Lot or Driveway Entrance/Exit	0.755	8.80	-9.27	5.19	4.08
Other Intersection Influence	0.560	3.59	-6.87	3.85	3.02
Roadway Environment					
Urban area indicator (1/0)	0.319	3.08	-3.92	2.20	1.73
Moderate Residential (1/0)	-0.203	-3.32	2.50	-1.40	-1.10
Threshold parameters					
μ ₁	2.801	27.92	---	---	---
μ ₂	4.292	37.42	---	---	---
Summary Statistics					
N	6,464				
Pseudo R ²	0.4925				
Loglikelihood (null)	-4013.632				
Loglikelihood (convergence)	-2036.74				
AIC	4113.479				
BIC	4248.959				
SBIC	4231.411				
CAIC	4249.411				
ICOMP	4099.032				

Table 5. 4 Comparing NB Classifiers: Default Prior, Informative Prior, and Optimal Kernel Width

Overall Classification Error	NB (Default Prior)	NB (Sample Prior)	Optimized NB
Resubstituting Error	0.3123	0.1563	0.0871
Cross validation Error	0.3223	0.1596	0.1002

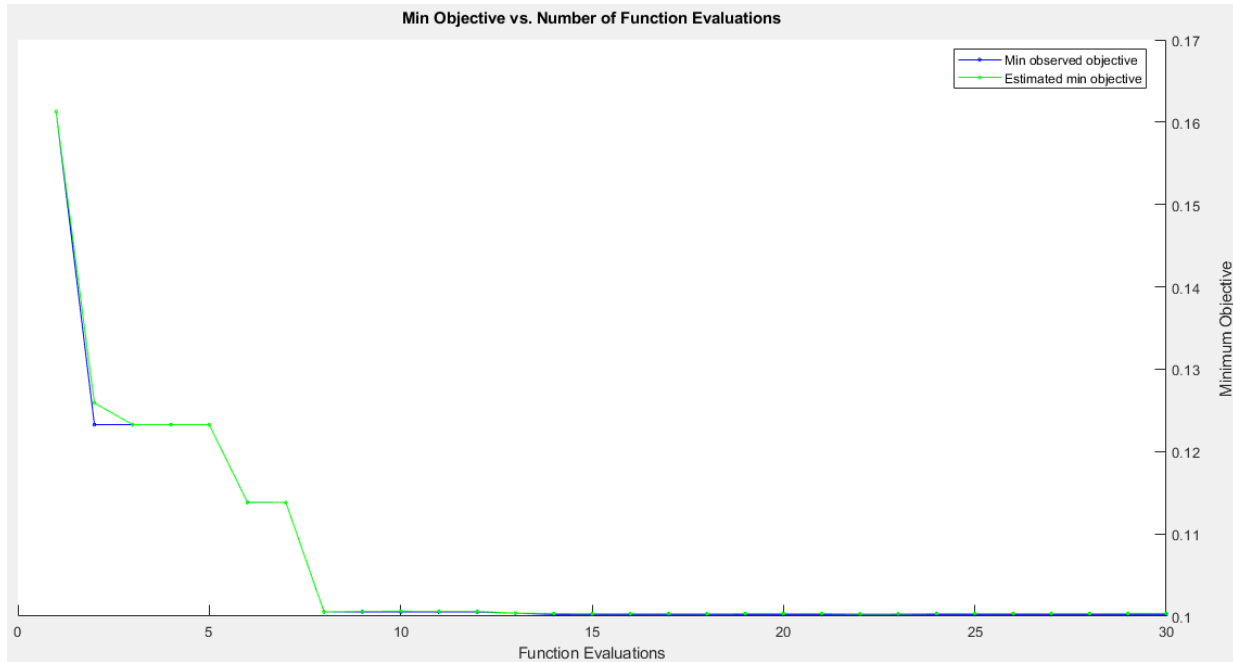


Figure 5. 2 Search for Minimum Objective Function (Optimization for NB)

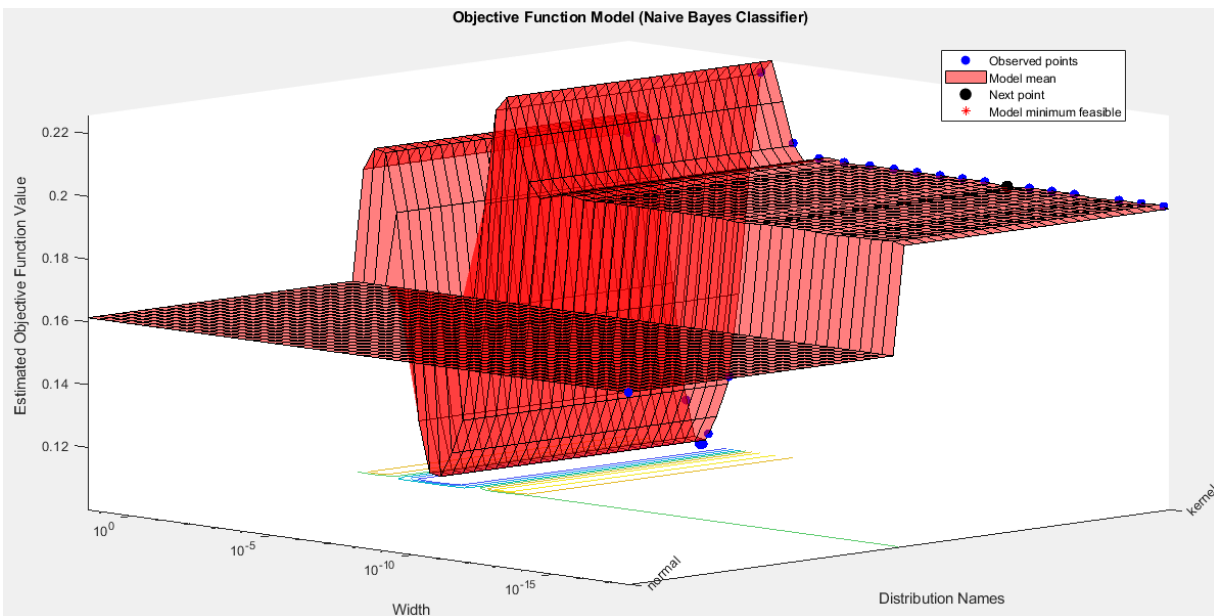


Figure 5. 3 Minimum Objective Function vs. Kernel Width (Optimization for NB Classifier)

Posterior Probabilities versus Driving Instability and Distraction Duration

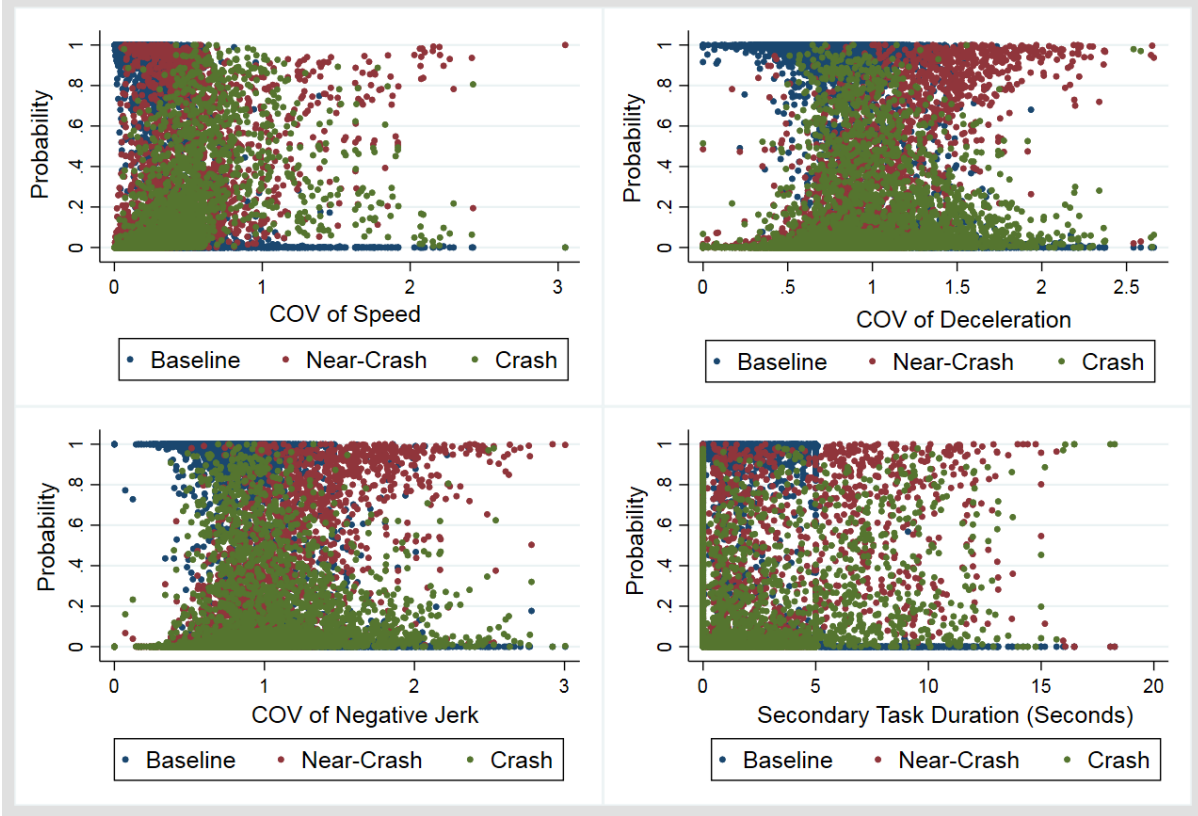


Figure 5. 4 Posterior Predicted Probabilities versus Driving Instability and Distraction Duration

distance with 3 nearest neighbors is found to have the lowest loss (minimum misclassification error). Therefore, the final KNN classifier with 3 nearest neighbors based on the mahalanobis distance was applied to the training data. Note that the key output of the KNN classifier is the distance matrix and not parameter estimates. Note that the minimum objective function (misclassification error) value and estimated function evaluation time are found to be 0.1125 and 4.4061 respectively which refer to the 3 nearest neighbors with Mahalanobis distance (Figure 5.5 and Figure 5.6). For the minimum objective function for the KNN classifier through grid-search optimization procedure, please see Figure 5.5 and Figure 5.6.

GBT Classifier

Before fitting a GBT classifier, a grid-search 10-fold cross-validation procedure was applied to select the optimal values of the tuning parameters including the number of trees, shrinkage, and interaction depth. For regularization, we considered a range of values for the number of trees (50, 200, 300, 400, and 1,000), shrinkage (0.01, 0.05, and 0.1), and interaction depth (1, 2, and 3). Through grid-search, the highest classification accuracy for the event outcome (baseline, near-crash, and crash) was obtained when the number of trees, interaction depth, and shrinkage parameters were 200, 2, and 0.1 respectively (Figure 5.7). Using the optimal values of the three parameters, the final GBT classifier was fitted to the training data.

Comparing Models: Out-of-Sample Prediction Accuracy

After estimating the ordered probit model and fitting the optimal ML methods to the training data, this study computes the out-of-sample prediction accuracy for the statistical model and ML methods to select the best model/method which could accurately predict the event outcome in the test (hold-out) sample. The confusion matrices for all the ordered Probit model and NB, KNN, and GBT classifiers for the test sample are presented in Table 5.5 the output of which is used to compute various performance measures of the out-of-sample prediction accuracy (Table 5.6). Based on the overall out-of-sample prediction accuracy, the GBT classifier shows the highest accuracy correctly predicting 91.23% of the observations in the test sample followed by NB (89.75%) and KNN (88.70%) classifier. Note that the ordered Probit model shows the lowest out-of-sample prediction accuracy (85.75%). Referring to the out-of-sample prediction performance accuracy for individual classes, the results indicate that the GBT classifier shows promising performance, compared to its counterparts, in predicting SCEs (crashes and near-crashes). For the baselines in the test sample, the recall (%) for the GBT is slightly lower than that of the ordered Probit model; however, the difference is minimal. However, the precision and F1 score of the GBT classifier for the baselines are also significantly higher than all its counterparts. Since the study focuses on predicting the SCEs for which the GBT classifier shows the highest out-of-sample prediction accuracy compared to its

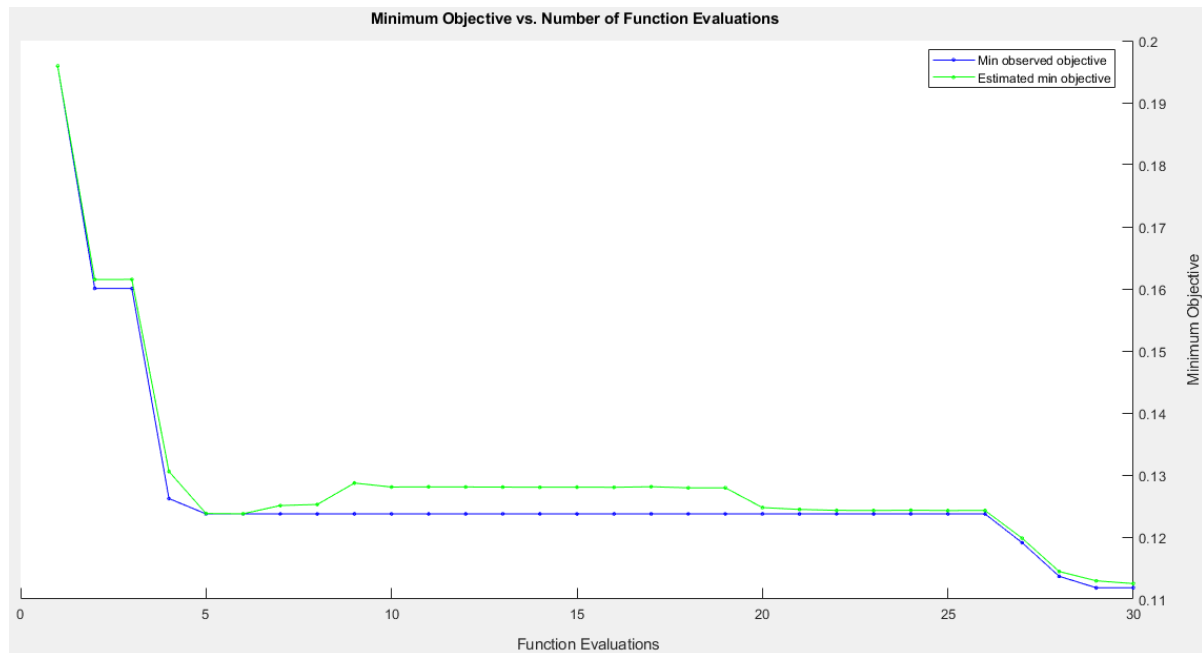


Figure 5. 5 Search for Minimum Objective Function (Optimization for KNN)

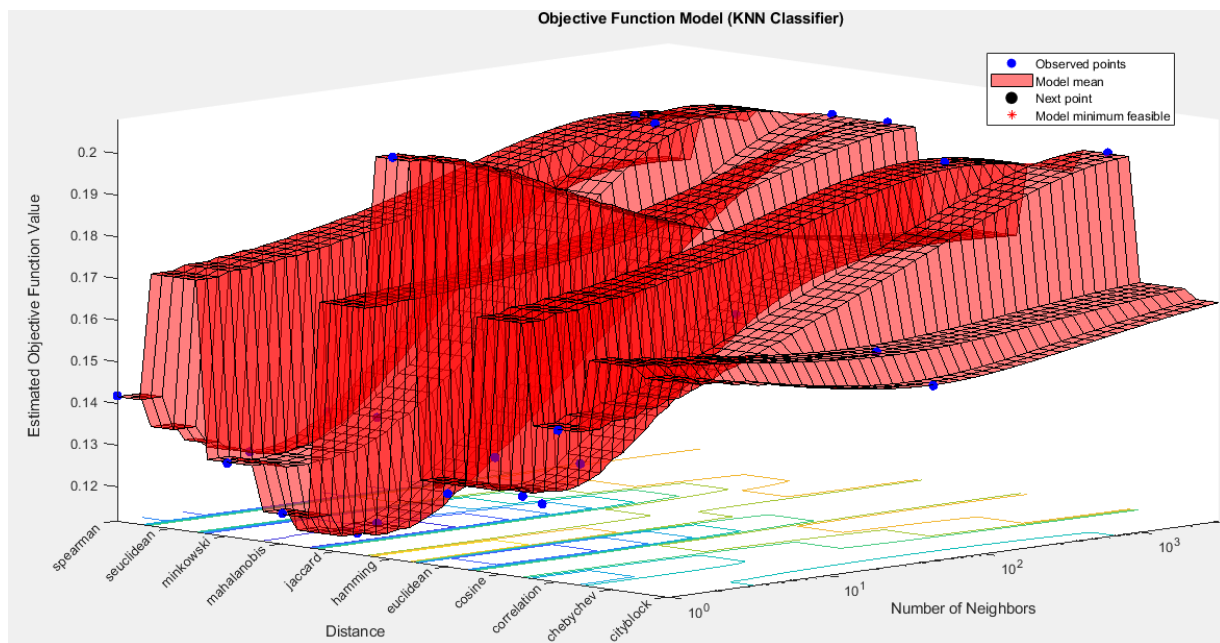


Figure 5. 6 Minimum Objective Function vs. Number of Nearest Neighbors (Optimization of KNN)

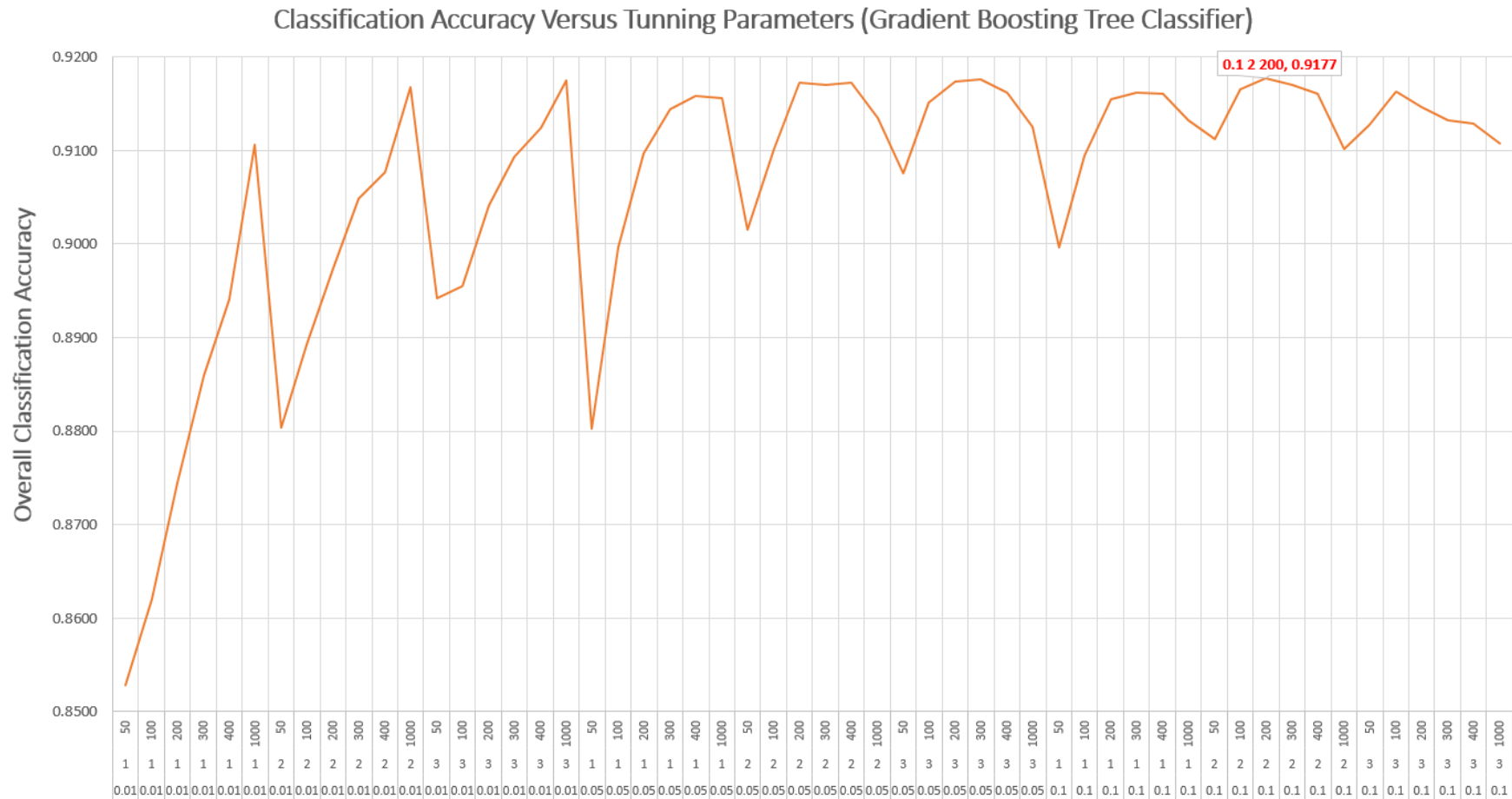


Figure 5. 7 Overall Classification Accuracy versus Tuning Parameters for the GBT Classifier

Note: The top (first), second, and third-row along the x-axis indicates different values of the number of trees, interaction depth, and shrinkage parameters respectively. The grid-search optimization reveals that the highest (91.77%) overall classification accuracy for the GBT classifier is obtained when the shrinkage, interaction depth, and the number of trees are 0.1, 2, and 200, respectively.

Table 5. 5 Confusion Matrices for Ordered Probit Model and NB, KNN, and GBT Classifiers

Observed (Event Outcome)	Predicted (Event Outcome)			Total	False Negative (FN)
	Baseline	Near-Crash	Crash		
Ordered Probit Model					
Baseline	2,161	33	4	2,198	37
Near-Crash	155	153	85	393	240
Crash	31	87	62	180	118
Total	2,347	273	151	2,771	---
False Positive (FP)	186	120	89	---	---
Optimal Naïve Bayes Classifier					
Baseline	2131	42	25	2198	67
Near-Crash	69	270	54	393	123
Crash	28	66	86	180	94
Total	2228	378	165	2771	---
False Positive (FP)	97	108	79	---	---
Optimal KNN Classifier					
Baseline	2135	52	11	2198	63
Near-Crash	103	243	47	393	150
Crash	24	76	80	180	100
Total	2262	371	138	2771	---
False Positive (FP)	127	128	58	---	---
Gradient Boosting Tree Classifier					
Baseline	2160	26	12	2198	38
Near-Crash	62	281	50	393	112
Crash	23	70	87	180	93
Total	2245	377	149	2771	---
False Positive (FP)	85	96	62	---	---

Table 5. 6 Comparing Overall and Class-level Out-of-Sample Prediction Accuracy

Performance Measure	Ordered Probit	NB	KNN	GBT
Overall Accuracy (%)	85.75	89.75	88.70	91.23
Baseline				
Recall (%)	98.32	96.95	97.13	98.27
Precision (%)	92.07	95.65	94.39	96.21
F1 Score (%)	95.09	96.29	95.74	97.23
Near-Crash				
Recall (%)	38.93	68.70	61.83	71.50
Precision (%)	56.04	71.43	65.50	74.54
F1 Score (%)	45.95	70.04	63.61	72.99
Crash				
Recall (%)	34.44	47.78	44.44	48.33
Precision (%)	41.06	52.12	57.97	58.39
F1 Score (%)	37.46	49.86	50.31	52.89

three counterparts. To conclude based on overall out-of-sample prediction accuracy and individual class-level out-of-sample prediction accuracy for SCEs (crashes and near-crashes), the GBT classifier is selected as the best model.

Discussion

The key focus of this study is to select the model/method which could accurately predict the event outcome especially SCEs. While the ML methods generally show promising performance in out-of-sample prediction accuracy compared to statistical models, we discuss the results of the ordered Probit model to 1) to understand how human factors relate to SCEs due to the interpretability of the statistical model, and 2) to use the ordered Probit model as a base for comparing the change (increase or decrease) in out-of-sample prediction accuracy with that of the ML methods. Referring to the results of the ordered Probit model, it is found that there is a 5.17%, 5.98%, and 1.16% higher chance of the subject driver to involve in a crash due to a unit increase in COV of speed, COV of deceleration, and COV of a negative jerk, respectively. These findings were expected as more volatile (unstable) drivers have a higher risk of involving in a crash (Arvin *et al.* 2019a). Similar findings were obtained for the near-crash.

Moreover, the results of the ordered Probit model suggest that driving errors and violations are critical as each of these errors and violations significantly increase the chance of both crash and near-crash compared to baseline. In particular, compared to no driving error and violations, performance error lead to the highest increase (by 13.75%) in crash risk followed by decision error and recognition error which increases the crash risk by 10.09% and 9.08%, respectively. Previous studies show similar findings which suggest that chance of both crash and near-crash risk is higher if a subject driver makes a driving error or violation (Ahmad *et al.* 2021, Khattak *et al.* 2021a). For distraction duration (seconds), we found that crash risk increases by 0.17% due to a unit increase in distraction duration while keeping other variables at their mean values. Also, the results suggest that compared to intersection influence, the crash risk is significantly higher by 4.24%, 4.08%, 2.20% at the interchange, parking lot or driveway entrance/exit, and uncontrolled intersection, respectively. Referring to the ML methods, this study first uses a grid-search optimization procedure to select the optimal values of the tuning parameters. For the NB classifier, the one with optimal kernel width showed the lowest classification error in the training sample and was selected as the final NB classifier compared to the NB classifiers which were trained with default prior probabilities (0.333, 0.333, and 0.333 respectively) and information prior probabilities (0.8034, 0.1305, and 0.0660, respectively). For the KNN classifier, the grid-search optimization showed that the minimum objective value (misclassification error) was found to be the lowest (0.1125) for the “Mahalanobis” distance with 3 nearest neighbors which are selected as the optimal KNN classifier. Referring to the GBT classifier, the analysis indicates that the optimal values for the tuning parameters including the number of trees, shrinkage, and

interaction depth are found to be 200, 0.1, and 2 respectively. Comparing the out-of-sample prediction accuracy, the results indicate that the GBT classifier accurately predicts 91.23% of observations in the test sample which is significantly higher than the corresponding out-of-sample prediction accuracy of its all counterparts. Besides, the GBT classifier also shows the highest precision, recall, and F1 score for crashes and near-crashes in the test sample. To conclude, the GBT classifier is selected as the most accurate method to predict crashes and near-crashes based on the key human factors and roadway and environment factors. Since the GBT classifier is selected as the most accurate method; therefore, we discuss the feature importance of the GBT classifier. Our findings indicate that among 18 predictors, the top five most important features used by the GBT classifier to predict SCEs include COV of speed, COV of deceleration, secondary task duration (seconds), recognition errors, and COV of the negative jerk (Figure 5.8). Note that all of these five features relate to human behavior and performance which signifies the negative role of human behavior and performance in driving. In general, the roadway and environment factors including intersection influence, and urban or moderate land use had lower importance in the GBT classifier to predict the event outcome (Figure 5.8).

Referring to the practical implications, we consider that multiple vehicle technologies including adaptive cruise control, forward-collision warning system, lateral vehicle detection system, dilemma zone mitigation systems could mitigate the driving errors/violations and/or driving instability. Furthermore, distracted driving and recognition errors could be reduced through vehicle technologies like warning systems and by educating the general public (in particular drivers) through seminars and campaigns about the negative consequences of distracted driving. As a part of future research, it would be good to develop an AI algorithm based on the GBT classifier in the instrumented vehicle which can be collected and process the information from the fitted cameras and equipment about driver behavior and performance, and roadway environment to predict the risk of both crash and near-crash in realtime. In particular, the automotive engineers may consider the research findings helpful; however, we consider that there is a need of implementing the TDEV in the algorithm to first classify the driving errors, violations, and compute driving instability (volatility) measures in real-time which could be then used in predicting (and revealing it in advance to the subject driver) the real-time crash and near-crash risks.

Limitations

This study has mainly used human factors (driving errors, violations, measures of driving instability, and distraction duration) along with key roadway and environment factors as the key features to predict the event outcome including baseline, near-crash, and crash. While the authors consider that driver demographics including age and gender could significantly relate to SCEs, SHRP2-NDS subsample, to which the authors have access, does not include these details. Furthermore, SHRP2-NDS data was collected

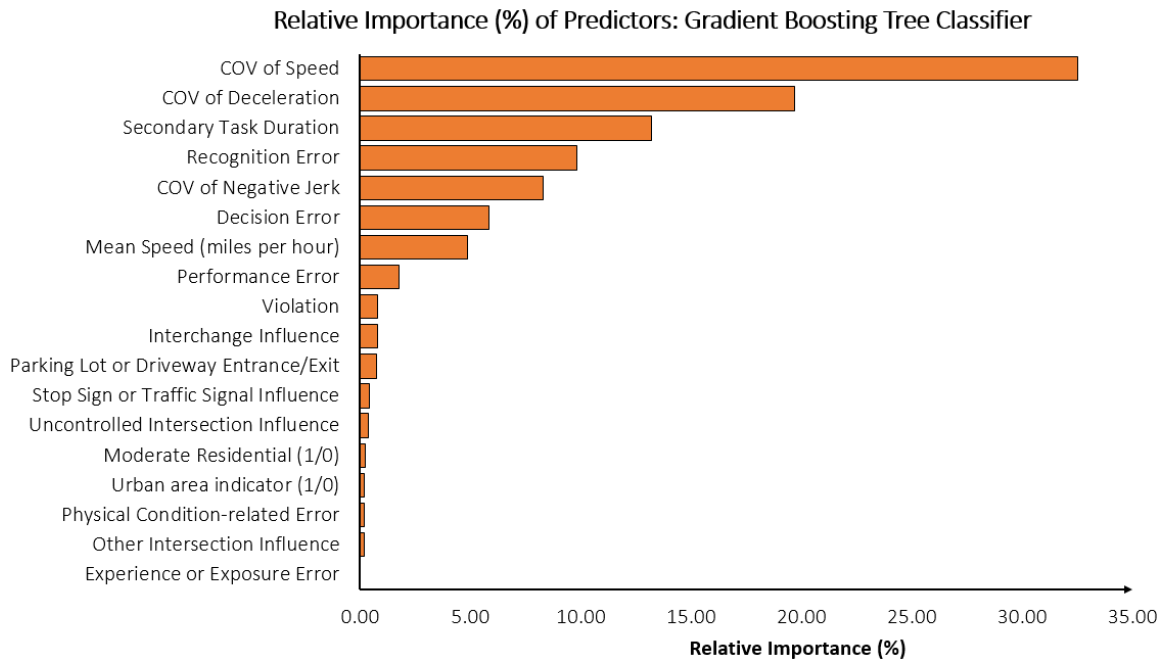


Figure 5. 8 Feature Importance Plot for the GBT Classifier

for volunteer drivers in six states including Indiana, Florida New York, Pennsylvania, North Carolina, and Washington in the U.S. Based on this, we consider the findings in this study may not be generalizable to other states across the U.S. or other countries due to potential variations in driver demographics, roadway conditions, and other important attributes.

Conclusions

Driver behavior and performance are the key contributing factors in SCEs including crashes and near-crashes. Studies have used traditional statistical models to understand how human factors including driving errors, violations, driving instability, and distraction duration relate to SCEs while analyzing real-world microscopic SHRP2-NDS data. While previous studies provide useful insights about the relation of human factors with SCEs, some research gaps still exist. First, none of the studies has looked into the effects of all of the four key human factors on SCEs. Second, while statistical models may provide useful inferential insights about the key relations in the data, these models have lower out-of-sample prediction accuracy which could be enhanced through the application of AI/ML methods. This study contributes by applying statistical model and ML methods to predict SCEs based on the four human factors while controlling for key road and environment factors. First, the statistical (ordered Probit) model and ML methods (NB, KNN, and GBT classifier) were fitted to the training data which provides useful insights. For instance, results of the ordered Probit model suggest that there is a 5.17%, 5.98%, and 1.16% significantly higher chance of subject driver to involve in a crash due to a unit increase in COV of speed, COV of deceleration, and COV of a negative jerk. According to the ordered Probit model, driving errors and violations showed a statistically significant and positive correlation with SCEs. For instance, compared to no driving error and violation, the chance of a crash increases by 9.08%, 10.09%, and 13.75% if a subject driver makes recognition error, decision error, and performance, respectively. Furthermore, results also indicate that the chance of SCEs is significantly higher for a subject driver who is distracted for a longer duration. Next, the three ML methods including Naïve Bayes, K-Nearest Neighbor, and gradient-boosting classifiers were applied to the training sample after selecting the optimal values of the corresponding tuning parameters obtained through the grid search optimization procedure. Finally, the study applies the ordered Probit model and ML methods to the new data (test sample) to compare the out-of-sample prediction accuracy. Based on the overall out-of-sample prediction accuracy, the GBT classifier is found to be the most accurate predictive method which accurately predicts the event outcome for 91.23% of observations in the test data. Besides the overall accuracy, the GBT classifier also showed the highest out-of-sample prediction accuracy for the crashes and near-crashes among all the four methods used in this study. Referring to the feature importance plot of the GBT classifier, COV of speed is found to be the most important feature followed by COV of deceleration, distraction duration, recognition errors, and COV of negative jerk used in predicting the event

outcome. In summary, the findings in this study suggest that human factors are critical and significantly increase the chance of SCEs which if mitigated could enhance safety. From the practical implication standpoint, vehicle technologies including forward-collision warning systems, adaptive cruise control, and lateral vehicle detection systems could help in reducing driving errors, violations, and driving instability which could consequently minimize the risk of SCEs. We also consider that the built environment (roadway) changes like removing sources of external distraction (e.g., billboards) could be helpful. Besides, educational seminars and awareness campaigns for drivers about the negative consequences of distraction could help in reducing driving distraction. As a part of future research, AI algorithms in instrumented vehicles could be developed to compute and reveal the risk of crash and near-crash to subject drivers based on their driving behaviors and performance in real-time.

**CHAPTER 6. THE ROLE OF HUMAN FACTORS IN PEDESTRIAN INJURIES GIVEN
A CRASH: APPLICATION OF FINITE MIXTURE MODEL**

A portion of this chapter by Numan Ahmad and Asad J. Khattak is under review for presentation and publication in Transportation Research Board 101st Annual Meeting and Journal of Transportation Research Record respectively.

Ahmad, N. and Khattak, A. J. Improving Safety by Investigating the Role of Human Errors and Violations in Pedestrian-Vehicle Crashes. In a review for presentation and publication in TRB2022 and Journal of Transportation Research Record respectively.

Abstract

The key objective of this study is to see how human factors including errors, violations, and impairment by both pedestrians and drivers related to pedestrian injury given a crash. Data on pedestrian-vehicle crashes (N = 24,886) were obtained from North Carolina (NC) Pedestrian and Bicycle Crash Data Tool (PBCAT) database. First, the pre-crash behaviors by both pedestrians and drivers are classified into various errors and violations. Next, the study considers that there could be some unobserved factors (not used in analysis or inherently unavailable) based on which pedestrians (or crashes) could be categorized into latent classes. These unobserved factors may include impact speed, pedestrian conspicuity, and riskiness due to which the effects of observed variables on pedestrian injury may significantly vary across various unobserved classes. To achieve the study objective, both simple ordered Probit (baseline) and finite mixture ordered Probit models were developed. Statistics reveal that 9%, 9.26%, 39.23%, 37.25%, and 5.26% of the crashes have resulted in a fatality, suspected serious injury, suspected minor injury, possible injury, and no injury to pedestrians respectively. Our results indicate that 58.50% of crashes have occurred due to errors and violations by pedestrians and drivers. Results indicate that chance of pedestrian fatality is significantly higher if pedestrian makes recognition errors and violations. Based on the finite mixture model, significant evidence that pedestrians could belong to two latent groups across which there is significant heterogeneity in pedestrian injuries is observed. If finite mixture models are not used in such cases, the estimates could be biased and inaccurate. Practical implications are discussed in the paper.

Introduction

Walking is considered as one of the sustainable modes which could bring both environmental and public health benefits (Murtagh *et al.* 2014, Peng *et al.* 2020, Risser and Sucha 2020). Promoting active transport modes including walking, cycling, and transit could reduce greenhouse gas (GHG) emissions and improve public health (Murtagh *et al.* 2014, Peng *et al.* 2020). Instead of the potential social, economic, and environmental benefits that walking could bring, pedestrians are found to be the most vulnerable road users with the least protection (Organization 2013, Peng *et al.* 2020). It is imperative to enhance pedestrian safety

while promoting this sustainable mode of transportation. Studies reveal a higher chance of severe pedestrian injuries and fatalities in a road crash. In the U.S., pedestrian fatalities have remarkably increased with a percentage of 53% from 2009 to 2018. In 2018, out of the total road crash fatalities in the U.S., 17% were pedestrians who lost their lives in road crashes. Having said this, there is a need to identify factors that contribute to pedestrian crashes and especially pedestrian injuries which in turn could help to identify more effective countermeasures to reduce pedestrian crashes and their injury given a crash. In this regard, studies have explored how various socio-demographics factors of pedestrians and drivers, built-environment (land use), roadway environment, and traffic conditions correlate with pedestrian injury. For instance, it was found that the built environment significantly affects pedestrian injury given a crash (Clifton *et al.* 2009). It is found that higher connectivity and transit access are negatively correlated with pedestrian injury (Clifton *et al.* 2009). Other factors including socio-demographics of both pedestrians and drivers, roadway geometry, and roadway environmental factors showed a significant correlation with pedestrian injury given a crash (Lee and Abdel-Aty 2005). Referring to socio-demographics factors, studies suggest that odds of pedestrian injury increases for older pedestrians (Kemnitzer *et al.* 2019). Studies also reveal that dark light conditions positively correlate with pedestrian injury (Kemnitzer *et al.* 2019). Besides, human factors are found to be the key contributing factors that could significantly influence both pedestrian crashes and pedestrian injury given a crash. Previous studies provide useful insights into how pre-crash behaviors of both pedestrians and drivers affect pedestrian injury given a crash, few research gaps still exist. For instance, most of the studies have looked at how individual pre-crash behaviors of drivers or pedestrians affect pedestrian injury. We consider that there is a need first to bring all pre-crash behaviors of both drivers and pedestrians into a systematic framework which could be then used to see their correlation with pedestrian injury. This could provide a more holistic picture of how a set of similar pre-crash behaviors (for both pedestrian and drivers) relate to pedestrian injury which could help in identifying more effective countermeasure to mitigate pedestrian injury given a crash. To bridge the research gaps, this study first classifies pre-crash behaviors of both drivers and pedestrians involved in pedestrian-vehicle crashes. Next, the study explores the effects of key errors, violations, and impairment of both pedestrian and driver on pedestrian injury given a crash. Importantly, the study tries to compute and use accurate estimates and prediction via finite mixture (latent class) models which could capture heterogeneity in pedestrian injuries across various latent classes due to some unobserved variable or set of unobserved variables (impact speed, conspicuity of pedestrians, familiarity with roadway environment, aggressiveness of both pedestrians and drivers).

Methodology

First, this study aims to bring all risky behaviors of both drivers and pedestrians involved in pedestrian crashes on roadway segments across North Carolina (2007-2019). For drivers, the recently developed taxonomy of driving errors and violations (TDEV) would be applied which is based on a theoretical perception-reaction process (Khattak *et al.* 2021a). For details, please refer to (Khattak *et al.* 2021a). For the pedestrian, we aim to develop a separate classification system through which the risky pedestrian behaviors would be brought under a systematic framework. The key focus of this study is to see how errors and violations made by pedestrians and drivers, in a pedestrian-involved crash, lead to pedestrian injury given a crash who is the most vulnerable road user. Besides quantifying the effects of pedestrian and driver behaviors on pedestrian injury severity, this study also includes other key control variables including light conditions, traffic control, and pedestrian age to see how these variables relate with pedestrian injury given a crash. For details about data, study design, and methodology, please refer to the subsections below.

Data

This study aims to analyze data on crashes that involved a pedestrian and occurred on roadway segments (2007-2019). The North Carolina Department of Transportation (NCDOT) used a special tool called NC Pedestrian and Bicycle Crash Data Tool (PBCAT) which includes separate four separate crash data sets which include: 1). NC Pedestrian Crashes on roadway segments; 2). NC Pedestrian Crashes on non-roadway; 3). NC Bicyclist Crashes on roadway segment; 4). NC Bicyclist Crashes on non-roadway (Harkey *et al.* 2006). The PBCAT data were derived from the NC Division of Motor Vehicles (DMV) Crash Report Form (DMV-349) which is used by the reporting officers to report a motor vehicle crash in NC. In NC, a reportable crash should meet one of the criteria which include: 1). A crash that involves a fatality or non-fatal injury to a person involved in a crash, or 2). A crash that causes overall property damage of \$ 1,000 or more, or 3). A crash causes property damage of any value to a vehicle that is seized (Harkey *et al.* 2006). In this study, we analyze a comprehensive database that includes all the necessary crash-related factors (indicating when, where, who, why, and in which conditions) the pedestrian-vehicle crashes have occurred across North Carolina State. The North Carolina Department of Transportation (NCDOT) used a special tool called NC Pedestrian and Bicycle Crash Data Tool (PBCAT) which includes separate four separate crash data sets which include: 1). NC Pedestrian Crashes on roadway segments; 2). NC Pedestrian Crashes on non-roadway; 3). NC Bicyclist Crashes on roadway segment; 4). NC Bicyclist Crashes on non-roadway (Harkey *et al.* 2006). In NC, a reportable crash should meet one of the criteria which include: 1). A crash that involves a fatality or non-fatal injury to a person involved in a crash, or 2). A crash that causes overall property damage of \$ 1,000 or more, or 3). A crash causes property damage of any value to a vehicle that

is seized (Harkey *et al.* 2006). The data was derived from the police crash reports. However, crash type and some additional unique attributes were specially coded by professional data experts in the PBCAT tool using the crash narrative in the police crash reports (Levitt). While police-reported pedestrian-involved crash data can have gaps in crash details, the PBCAT provides a more comprehensive high-quality database of pedestrian-vehicle crashes. The PBCAT is a crash typing software that is intended to help local and state bicycle/pedestrian coordinators, engineers, and planners in improving the safety of pedestrians and bicyclists by providing a comprehensive database of pedestrian-vehicle and bicyclist-vehicle crashes occurring across North Carolina (Levitt). The “crash type” variable in the PBCAT data includes details about the risky behaviors of drivers and pedestrians who were involved in pedestrian-vehicle crashes (Levitt). Other interesting variables in the PBCAT data includes impairment (due to alcohol or drugs) status of pedestrians and drivers, location or position of a pedestrian at the time of crashes, roadway factors (functional classes of roadways, roadway configuration), lighting conditions, speed limit, land use and types of surrounding development, weather conditions, geocoded location of the crash, and demographics details (e.g., age, gender, and race) of pedestrians and drivers who were involved in pedestrian-vehicle crashes (Levitt). We have selected pedestrian-vehicle crashes (N= 24,886) for the analysis which have occurred between 2009-2019 on roadway segments across North Carolina. The injury severity of both pedestrians and drivers is reported on a five-level KABCO scale (Levitt). It is important to mention that NC-PBCAT data also provide comprehensive data about bicycle-vehicle crashes; however, given the study scope, this study analyzes only the pedestrian-vehicle crashes which have occurred on roadways. We consider that this comprehensive data on pedestrian-vehicle crashes can be used to investigate several important safety issues related to pedestrians. Given the study objectives, we only focus on understanding the effects of human behaviors on pedestrian injury severity. Several other control variables related to the presence/types of traffic control devices, lighting conditions, pedestrian demographics (age and gender), and impairment due to alcohol by pedestrian and driver are also considered in the analysis.

Conceptual and Methodological Framework

The key focus of this study is to understand how pre-crash risky behaviors (errors and violations) of both drivers and pedestrians affect pedestrian injury severity given a pedestrian-motor vehicle crash. The conceptual framework for this study is shown in Figure 6.1.

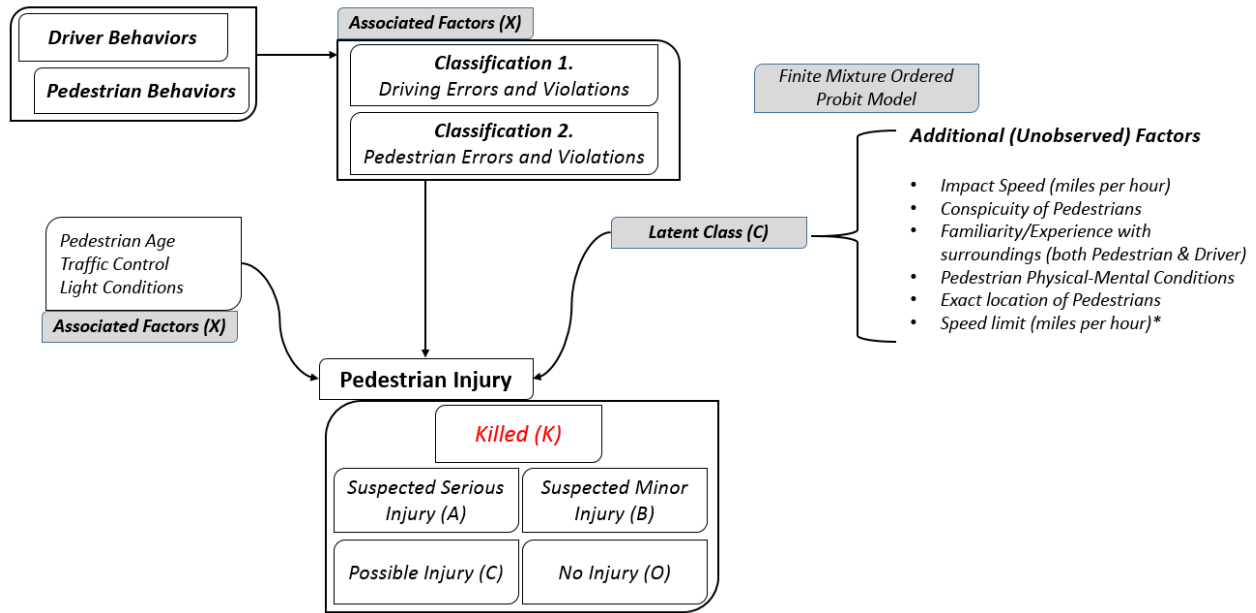


Figure 6. 1 Conceptual Framework

Note: Additional unobserved factors may refer to the factors which are available in the data but not used in the analysis (like speed limit) or indicate the ones which could be inherently unavailable (e.g., impact speed, conspicuity of pedestrian, the unfamiliarity of pedestrians or drivers with roadway). Also, please note that some specific explanatory observed variable (e.g., speed limit) or inherently unobserved factors could be considered as a base for classification which depends on the scope and objective of the study. In this study, we go with the later approach where we hypothesize the latent class split of pedestrians (or crashes in which pedestrians were involved) based on important unobserved factors like pedestrian conspicuity, impact speed, riskiness.

The motivation behind this research is based on some of our key hypotheses. For instance, one of our key hypotheses is that errors and/or violations by drivers may contribute more to pedestrians due to their vulnerability. Similarly, we also consider that some of the key pre-crash behaviors of pedestrians themselves could lead to serious injuries or fatalities by pedestrians. The interesting aspect of research is that we want to see the effects of various combinations of impairments by drivers and pedestrians. For instance, compared to crashes where neither driver nor pedestrian is impaired, how would the chance of a pedestrian being killed or get seriously injured change if 1) The driver was impaired but the pedestrian was not 2) The driver was not impaired, but the pedestrian was impaired, and 3) Both driver and pedestrian were impaired. First, pre-crash behaviors of drivers and pedestrians will be classified.

Importantly, we consider that the effects of errors, violations, and impairment by drivers and pedestrians, and other control variables including light conditions and age may not be the same as we consider that there could be two or more latent classes based on some unobserved or set of unobserved variables. For instance, studies may suggest that if a pedestrian-vehicle crash occurs during nighttime without streetlights, the chance of most severe pedestrian injury would be higher which could be inappropriate if the model does not include information about pedestrian conspicuity. We consider that based on outfits, pedestrians could be classified into two or more groups. Having said this, we consider that the effect of lighting conditions and other factors may vary across the two groups. Hence, it would be more appropriate to account for such unobserved heterogeneity (across two or more classes) within pedestrian injury based on a set of unobserved factors which we consider important and could have a significant effect on pedestrian injury. Some of such unobserved variables are shown in Figure 6.1.

Methods

Given the ordinal nature of the pedestrian injury, this study uses the ordered probit model which is briefly discussed in the subsection below. Furthermore, the study aims to account for unobserved heterogeneity in the effects of several variables on pedestrian injury across two or more latent classes which could exist based on some unobserved or set of unobserved variables. To account for such heterogeneity between classes, finite mixture or latent class models could be used which are also discussed after the ordered probit model.

Ordered Probit Model

Given the ordinal nature of the pedestrian injury, we use the ordered probit model. The unobserved pedestrian injury severity (Y_i^*) could be a linear function of specific explanatory variables (X_i) and can be given below (Wali *et al.* 2018a, Ahmad *et al.* 2019, Washington *et al.* 2020).

$$Y_i^* = \beta X_i + \varepsilon_i \quad (1)$$

where Y_i^* indicate an unobserved measure of pedestrian injury in a crash i , β refers to a set of parameters to be estimated for a set of explanatory variables X_i , and ε_i indicates unobserved factors which are assumed to have normal distribution $\sim N(0,1)$. Given a crash, the maximum pedestrian injury could fall within any of the five categories (KABCO). The latent Y_i^* and observed (Y_i) pedestrian injury are related as below (Wali *et al.* 2018a, Ahmad *et al.* 2019c, Washington *et al.* 2020).

$$Y_i = \begin{cases} 0 & \text{if } -\infty \leq Y_i^* \leq \mu_0 & (\text{No Injury}) \\ 1 & \text{if } \mu_0 < Y_i^* \leq \mu_1 & (\text{Possible Injury}) \\ 2 & \text{if } \mu_1 < Y_i^* \leq \mu_2 & (\text{Suspected Minor Injury}) \\ 3 & \text{if } \mu_2 < Y_i^* \leq \mu_3 & (\text{Suspected Serious Injury}) \\ 4 & \text{if } \mu_3 < Y_i^* \leq +\infty & (\text{Fataality}) \end{cases} \quad (2)$$

Where " μ " refers to the thresholds parameters to be estimated. The probability of each of the five categories of pedestrian injury severity can be computed as below (Kockelman and Kweon 2002, Ahmad *et al.* 2019, Menon *et al.* 2019).

$$\begin{aligned} \text{Prob}(Y_i = 0|X_i) &= \Phi(-\beta X_i) \\ \text{Prob}(Y_i = 1|X_i) &= \Phi(\mu_1 - \beta X_i) - \Phi(-\beta X_i) \\ \text{Prob}(Y_i = 2|X_i) &= \Phi(\mu_2 - \beta X_i) - \Phi(\mu_1 - \beta X_i) \\ \text{Prob}(Y_i = 3|X_i) &= \Phi(\mu_3 - \beta X_i) - \Phi(\mu_2 - \beta X_i) \\ \text{Prob}(Y_i = 4|X_i) &= 1 - \Phi(\mu_3 - \beta X_i) \end{aligned} \quad (3)$$

Referring to the interpretation of the results obtained through the ordered Probit model, positive β would suggest an increase in the probability of the highest injury (pedestrian being killed) and decrease in the probability of the lowest injury (pedestrian with no injury) (Ahmad *et al.* 2019). While parameters obtained through the ordered Probit model are useful, they just show the direction of association of specific explanatory variables with the lowest and highest categories of the response variable (Wali *et al.* 2018a). To understand the change in probability of each category of the pedestrian injury with a unit increase in a specific continuous explanatory variable (or if indicator switches from 0 to 1), while keeping all other variables in the model at their mean values, marginal effects (MEs) can be computed. For continuous explanatory variable, MEs can be computed as below (Wali *et al.* 2018a, Washington *et al.* 2020).

$$\frac{\partial \text{Prob}(Y_i=n)}{\partial x} = -[\Phi(\mu_{n-1} - \beta X_i) - \Phi(\mu_n - \beta X_i)]\beta \quad n = 0,1,2,3,4 \quad (4)$$

While μ_{n-1} and μ_n are the threshold parameters and n represent various levels of the response variable. For MEs of indicator variables, the following equations can be used.

$$\partial P(Y_i = n|X_i) = P(Y_i = n|X_i = 1) - P(Y_i = n|X_i = 0) \quad (5)$$

The final ordered Probit model is systematically derived while including the most important variables statistical significance, specification parsimony, and intuition. In doing so, the likelihood ratio test, Akaike information criterion “AIC” (Akaike 1974), and Bayesian information criterion “BIC” (Schwarz 1978) statistics were examined. In comparing the competing models, the one with lower AIC and BIC statistics is the best fit model after adjusting for degrees of freedom. Other performance measures can also be used to access the model performance depending on the study scope. For instance, BIC and Schwarz criterion “SBIC” are considered as more parsimonious criteria; however, SBIC outperforms the BIC (Nickelsburg 1985). For the sake of completeness, SBIC (Nickelsburg 1985), Consistent AIC “CAIC” (Bozdogan 1987), and Bozdogan’s Information Complexity “ICOMP” (Bozdogan 1990) are considered (where applicable) to select the final model while comparing it to its counterparts. For details about various criteria used to access the model performance, please refer to the studies in parenthesis (Akaike 1974, Schwarz 1978, Nickelsburg 1985, Bozdogan 1987, 1990).

Finite Mixture Ordered Probit Model

The simple ordered probit model (as discussed above) quantifies the effects of known (observed) factors, included in the model, and pedestrian injury, and assumes that all pedestrians (or crashes) are the same based on unobserved factors. However, there could be significant heterogeneity in pedestrian injuries given a crash based on some unobserved class variables or a set of class variables. For instance, studies may quantify the effects of lighting conditions on pedestrian injury through a simple ordered probit/logit model and suggest that chance of most severe injury to a pedestrian is higher if the crash occurs during the nighttime with lighting. However, we consider that such findings and estimates which are just based on lighting conditions without including information about pedestrian conspicuity could be less accurate. Furthermore, we consider that the effects of lighting conditions on pedestrian injury may vary across latent (unobserved) classes based on pedestrian conspicuity (pedestrian with bright clothes versus dark clothes). Having said this, the effects of some or all variables included in the model could be different among various latent “unobserved” classes considering an unobserved factor or a set of unobserved factors. If latent classes, based on some unobserved variables exist, but are not considered, this could lead to biased estimates and incorrect predictions. In Stata, the finite mixture model provides an opportunity to take such unobserved factors (e.g., pedestrian conspicuity, impact speed, and speed limit) using which the pedestrian or crashes could be categorized into two or more classes. These models provide two or more set of estimates which

could be used for inference or predictions after accessing the latent class probabilities which show the appropriate latent class of an observation.

Classification of Human Factors and Descriptive Statistics

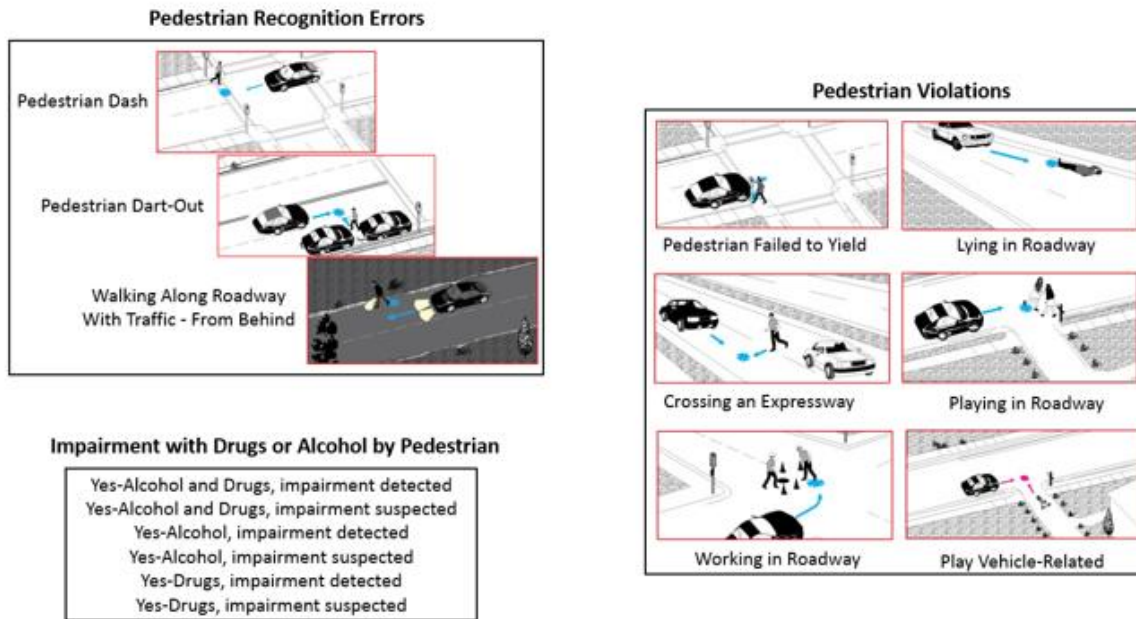
Classification of Human Behaviors

In this section, we discuss the classification of behaviors of both pedestrians and drivers in pedestrian-vehicle crashes. For driving behaviors, we use the recently developed TDEV which has classified driving behaviors into five key driving errors (recognition, decision, performance, physical-condition related, and experience errors) and violations (Khattak *et al.* 2021a). Based on the available details in the NC-PBCAT data, we have classified driving behaviors into 1) recognition errors (backing vehicle along the roadway, in parking lot, or intersection), 2) Performance errors (motor vehicle loss of control), and 3) Violation (motorist failed to yield). Note that impairment of drivers due to drug or alcohol is placed in a separate category which in fact could negatively influence driving behaviors during one or more of the driving phases (perception, recognition, decision, and reaction); however, it does not directly relate to the perception-reaction process – for details, please refer to (Khattak *et al.* 2021a). For pedestrians, we have classified pedestrian behaviors into possible recognition errors (pedestrian dash and pedestrian dart out) and violations (crossing an expressway, lying in a roadway, walking in a roadway, playing in the roadway, working in the roadway, play vehicle-related, and pedestrian failed to yield). Similar to the driver, we have placed the pedestrian impairment as a separate pedestrian human-related factor that could have contributed to pedestrian-vehicle crashes. For some pedestrian behaviors like dart or dash-out, we consider that these behaviors could have led to recognition errors by pedestrians themselves where they could have not appropriately recognized the incoming vehicle due to rushing. For a summary of the human factors related to both drivers and pedestrians, please refer to Figure 6.2.

Descriptive Statistics

According to NC PBCAT (2007-2019), a total of 27,300 pedestrian-motor vehicle crashes were reported on roadways. For N = 2,247 of the total crashes, either pedestrian injury was either missing or reported as unknown. Similarly, pedestrian age was missing in another N = 167 crashes. To get a cleaner dataset for analysis, such observations were excluded from the final dataset considered for analysis. After cleaning data, a total of 24,886 pedestrian-motor vehicle crashes which have occurred on roadways were considered for analysis (Table 6.1). Statistics reveal that out of total pedestrian-vehicle crashes (N = 24,886) in NC, in 9% (N = 2,240), 9.26% (N = 2,305), 39.23% (N = 9,763), 37.25% (N = 9,269) and 5.26% (N = 1,309) of the crashes, pedestrian was killed (K), seriously injured (A), received minor injury (B), received possible

a. Pedestrian Behaviors (Human Factors) in Pedestrian-vehicle Crashes



b. Driver Behaviors (Human Factors) in Pedestrian-vehicle Crashes

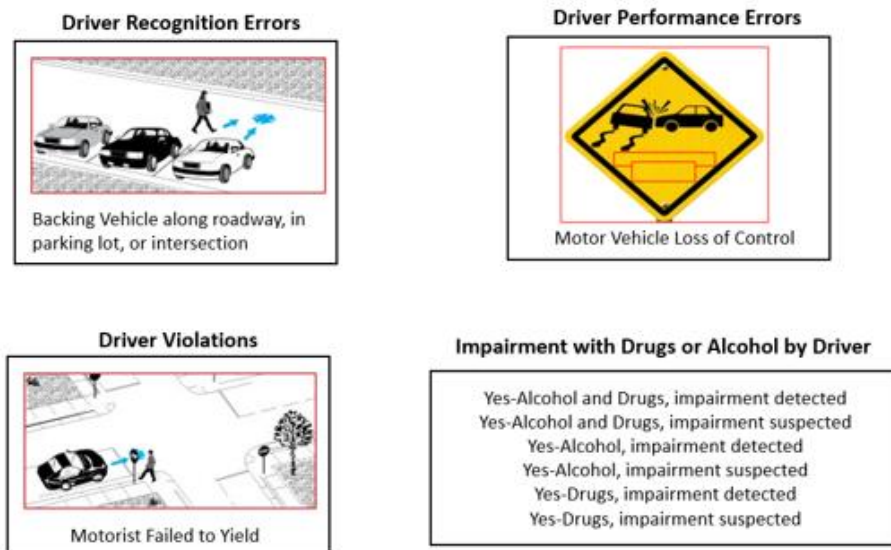


Figure 6. 2 Classifying Pre-Crash Behaviors of Drivers and Pedestrians in Crashes

For sources to pictures, please refer to https://www.pedbikeinfo.org/pbcat_us/ped_images.cfm

Table 6. 1 Descriptive Statistics of Key Variables

Key Variables	Total Crashes	“O” Injury	“C” Injury	“B” Injury	“A” Injury	“K” Injury
	(N = 24,886)	(N = 1,309)	(N = 9,269)	(N = 9,763)	(N = 2,305)	(N = 2,240)
	Freq (%)	Freq (%)	Freq (%)	Freq (%)	Freq (%)	Freq (%)
Pedestrian Behaviors						
Recognition Error (Yes/No)	5,481 (22.01%)	220 (16.81%)	1,813 (19.56%)	2,450 (25.09%)	565 (24.51%)	433 (19.33%)
Violation (Yes/No)	7113 (28.58%)	321 (24.52%)	1,922 (20.74%)	2,876 (29.46%)	864 (37.48%)	1,130 (50.45%)
Driver Behaviors						
Recognition Error (Yes/No)	527 (2.12%)	47 (3.59%)	296 (3.19%)	150 (1.54%)	21 (0.91%)	13 (0.58%)
Performance Error (Yes/No)	667 (2.68%)	29 (2.22%)	206 (2.22%)	247 (2.53%)	106 (4.60%)	79 (3.53%)
Violation (Yes/No)	775 (3.11%)	51 (3.90%)	395 (4.26%)	267 (2.73%)	48 (2.08%)	14 (0.63%)
Impairment with Alcohol or Drugs						
Both Driver and Pedestrian were Sober	21,366 (85.86%)	1,178 (89.99%)	8,461 (91.28%)	8,377 (85.80%)	1,800 (78.09%)	1,550 (69.20%)
Driver Impaired, but Pedestrian was Sober	569 (2.29%)	16 (1.22%)	134 (1.45%)	295 (2.00%)	101 (4.38%)	123 (5.49%)
Driver was Sober, but Pedestrian was Impaired	2,770 (11.13%)	109 (8.33%)	654 (7.06%)	1,123 (11.50%)	372 (16.14%)	512 (22.86%)
Both Driver and Pedestrian were Impaired	181 (0.73%)	6 (0.46%)	20 (0.22%)	68 (0.70%)	32 (1.39%)	55 (2.46%)
Light Condition						
Unknown or other	58 (0.23%)	4 (0.31%)	19 (0.2%)	22 (0.23%)	5 (0.22%)	8 (0.36%)
Daylight	12,026 (48.32%)	772 (58.98%)	5,285 (57.02%)	4,749 (48.64%)	787 (34.14%)	433 (19.33%)
Dawn	411 (1.65%)	16 (1.22%)	168 (1.81%)	165 (1.69%)	38 (1.65%)	24 (1.07%)
Dusk	702 (2.82%)	39 (2.98%)	283 (3.05%)	286 (2.93%)	58 (2.52%)	36 (1.61%)
Dark but lighted	5,833 (23.44%)	257 (19.63%)	1,953 (21.07%)	2,459 (25.19%)	615 (26.68%)	549 (24.51%)
Dark unknown lighting	133 (0.53%)	10 (0.76%)	60 (0.65%)	43 (0.44%)	13 (0.56%)	7 (0.31%)
Dark not lighted	5,723 (23.00%)	211 (16.12%)	1,501 (16.19%)	2,039 (20.88%)	789 (34.23%)	1,183 (52.81%)
Traffic Control						
Traffic Control: No Control (1/0)	14,107 (56.69%)	733 (56.00%)	4,873 (52.57%)	5,656 (57.93%)	1,401 (60.78%)	1,444 (64.46%)
Traffic Control: Human Control (1/0)	266 (1.07%)	24 (1.83%)	131 (1.41%)	85 (0.87%)	18 (0.78%)	8 (0.36%)
Socio-demographics						
Pedestrian is female (1/0)	8,831 (35.49%)	440 (33.61%)	3,693 (39.84%)	3,399 (34.82%)	693 (30.07%)	606 (27.05%)
	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)
Pedestrian age (Years) * (Min = 2; Max = 98)	35.60 (17.92)	34.35 (17.49)	35.02 (17.29)	34.39 (18.12)	37.08 (17.81)	42.52 (18.33)

Note: * indicates that for pedestrian age, the values in the columns indicate the mean (standard deviation) based on total crashes and each pedestrian injury severity category. For all the remaining variables, we have shown frequency (percentage). S.D., Min, and Max indicate standard deviation, minimum, and maximum respectively. Where “K”, “A”, “B”, “C”, and “O” injury indicates fatal, suspected serious injury, suspected minor injury, possible injury, and no injury to the pedestrian, respectively.

injury (C), and no injury (O) respectively. As expected, in almost 95% of the crashes, pedestrian was reported to have some injury ranging from possible injury to fatal injury. Referring to driver errors and violations, in 2.12%, 2.68%, and 3.11% of the total pedestrian-vehicle crashes, drivers have made recognition errors, performance errors, and violations respectively. Based on errors and violations by pedestrians, in 22.01% and 28.58% of the total crashes, pedestrians were involved in recognition errors and violations respectively. The percentages of recognition errors and violations by pedestrians are significantly higher compared to the ones made by drivers. For instance, recognition errors by pedestrians have resulted contributed into 22.01% of the overall crashes which is significantly higher than the percentage (2.12%) of these errors by drivers. Whereas, violations by pedestrians were almost nine times of the ones made by drivers. Referring to the impairment with alcohol or drugs, in 85.86% of the total crashes, neither driver nor pedestrian was impaired. However, statistics show that in 2.29%, 11.13%, and 0.73% of the total pedestrian-vehicle crashes, only driver, only pedestrian, and both driver and pedestrian were impaired respectively. Similar to recognition errors and violations by pedestrian, this also seems concerning that pedestrian was impaired in 11.13% of the crashes which is significantly higher than the percentage of crashes (2.29%) were only driver was impaired. To conclude, we found that errors and violations by pedestrians and drivers have contributed to 58.50% of the total pedestrian-vehicle crashes. Statistics also reveal that in 14.15% of the crashes, either pedestrian, driver, or both pedestrian and driver were impaired with alcohol and drugs. Furthermore, statistics show that the mean age of pedestrians involved in a crash was 35 years with minimum and maximum values of 2 and 98 years respectively. For pedestrian age (years), the first and second columns in the table show mean and standard deviations (S.D) respectively. To understand the distribution of key human factors and other attributes in various categories of pedestrian injuries, please refer to Table 6.1.

Results

Ordered Probit Model

To achieve the study objectives, this study estimates a traditional ordered Probit model which we term as the “baseline” model. Results of the final baseline model (selected with the lowest AIC, BIC, CAIC, SBIC, and ICOMP compared to its competing models) including the most important variables are shown in Table 6.2. Note that the decision of including a variable was based on statistical significance, intuition, and parsimony. For variable significance, 95% confidence criteria were used. However, for the sake of completeness, some of the variables or categories of the variables, were included in the final model even if they showed marginal significance or no significance such as an indicator of violation by a driver, indicator

Table 6. 2 Results of Simple Ordered Probit (Baseline) Model

Key Variables	Coeff.	t-stats	Chance (=Marginal Effects*100)				
			“O” Injury	“C” Injury	“B” Injury	“A” Injury	“K” Injury
Pedestrian Behaviors							
Recognition Error (Yes/No)	0.2379	12.78	-2.49	-6.22	2.84	2.34	3.53
Violation (Yes/No)	0.4295	24.94	-4.50	-11.23	5.14	4.23	6.36
Driver Behaviors							
Recognition Error (Yes/No)	-0.2865	-5.89	3.00	7.49	-3.43	-2.82	-4.25
Performance Error (Yes/No)	0.3610	8.41	-3.78	-9.44	4.32	3.55	5.35
Violation (Yes/No)	-0.0704	-1.74	0.74	1.84	-0.84	-0.69	-1.04
Impairment with Alcohol or Drugs (Base = Both Driver and Pedestrian were Sober)							
Driver Impaired, but Pedestrian was Sober	0.5286	11.48	-5.54	-13.82	6.32	5.20	7.83
Driver was Sober, but Pedestrian was Impaired	0.2049	9.00	-2.15	-5.36	2.45	2.02	3.04
Both Driver and Pedestrian were Impaired	0.7387	9.09	-7.74	-19.31	8.83	7.27	10.95
Light Condition (Base = Daylight)							
Unknown or other	0.2339	1.66	-2.45	-6.12	2.80	2.30	3.47
Dawn	0.1545	2.88	-1.62	-4.04	1.85	1.52	2.29
Dusk	0.0691	1.66	-2.03	-5.07	2.32	1.91	2.87
Dark but lighted	0.1940	10.94	-2.03	-5.07	2.32	1.91	2.87
Dark unknown lighting	-0.0727	-0.78	0.76	1.90	-0.87	-0.72	-1.08
Dark not lighted	0.5293	29.06	-5.55	-13.84	6.33	5.21	7.84
Traffic Control							
Traffic Control: No Control (1/0)	0.0664	4.67	-0.70	-1.74	0.79	0.65	0.98
Traffic Control: Human Control (1/0)	-0.3972	-5.89	4.16	10.39	-4.75	-3.91	-5.89
Socio-demographics							
Pedestrian is female (1/0)	-0.0451	-3.12	0.47	1.18	-0.54	-0.44	-0.67
Pedestrian age (Years)	0.0058	14.80	-0.06	-0.15	0.07	0.06	0.09
Threshold parameters							
μ1	-1.0948	-46.22	---	---	---	---	---
μ2	0.3960	18.15	---	---	---	---	---
μ3	1.5762	67.62	---	---	---	---	---
μ4	2.0610	83.04	---	---	---	---	---
Model Summary (SBIC = 63315.64; CAIC = 63333.64; ICOMP = 63200.01; Log-likelihood at Convergence = -31566.72)							
AIC	63177.44						
BIC	63356.13						

of unknown or another lighting, and indicator dark unknown lighting. The remaining variables included in the model showed an intuitive and significant correlation with pedestrian injury severity. Marginal effects (MEs) for the explanatory variables are also provided which show the change in the probability of a specific severity level of pedestrian injury with a unit increase in a continuous variable (pedestrian age) or if an indicator switches from 0 to 1 while keeping all other variables at their mean values. We note that MEs when multiplied by 100 can be interpreted as chance. The results suggest that the chance of a pedestrian being killed is higher by 3.53% and 6.36% if a pedestrian is involved in recognition error and violation respectively. Results also indicate that chance of fatal injury and suspected serious injury to a pedestrian is higher by 5.35% if a driver makes a performance error. On other hand, the chance of fatal injury and suspected serious injury is significantly lower if a driver makes a recognition error or violation. For a detailed discussion about the interesting findings, please refer to the discussion section.

Finite Mixture Ordered Probit Model

We estimate the finite mixture ordered Probit model assuming two latent classes. Note that latent classes of more than two were also applied; however, the model with more than two latent classes did not converge. Referring to the results of FMOPM with two latent classes, we noticed a significant improvement in AIC, BIC, and log-likelihood at convergence compared to the simple “baseline” ordered Probit model (please see Table 6.2 and Table 6.3). Also, significant variation in parameter estimates of observed variables on pedestrian injury severity across the two latent classes can be noticed. Based on the AIC, BIC, and loglikelihood values, the FMOPM can be selected as the best model. Please note that the class variable can be simply considered as an unobserved variable or indicator of a set of unobserved variables which is used by FMOPM for splitting the observations into two unobserved classes. For a detailed discussion about the results obtained through the FMOPM, please refer to the discussion section.

Discussion

Statistics reveal that that out of total pedestrian-vehicle crashes ($N = 24,886$), a pedestrian was killed (K), seriously injured (A), received minor injury (B), received possible injury (C), and no injury (O) in NC, in 9%, 9.26%, 39.23%, 37.25% and 5.26% crashes, respectively (Table 6.1). After classifying the errors and violations by pedestrians and drivers in pedestrian-vehicle crashes, errors and violations by drivers and pedestrians have contributed to 58.50% of the overall crashes. It is important to mention that errors and violations by pedestrians contributed to around 50% of the overall crashes which is significantly higher than the percentage of errors and violations by drivers which have contributed to around 8% of the overall pedestrian-vehicle crashes only. Statistics also reveal that in 14.15% of the crashes, either pedestrian or

Table 6. 3 Results of Finite Mixture Ordered Probit Model

Key Variables	Class 1		Class 2	
	Coeff.	t-stats	Coeff.	t-stats
Pedestrian Behaviors				
Recognition Error (Yes/No)	0.3502	6.28	0.2357	5.97
Violation (Yes/No)	0.1350	2.97	0.6551	18.09
Driver Behaviors				
Recognition Error (Yes/No)	-0.0649	-0.32	-0.4554	-3.19
Performance Error (Yes/No)	0.0803	0.74	0.6083	7.89
Violation (Yes/No)	0.0766	0.55	-0.1877	-1.94
Impairment with Alcohol or Drugs (Base = Both Driver and Pedestrian were Sober)				
Driver Impaired, but Pedestrian was Sober	0.1322	1.15	0.7972	10.37
Driver was Sober, but Pedestrian was Impaired	0.1345	2.50	0.3150	7.89
Both Driver and Pedestrian were Impaired	0.4631	2.38	1.0609	7.50
Light Condition (Base = Daylight)				
Unknown or other	-0.3095	-1.11	0.6705	2.90
Dawn	0.0625	0.38	0.2438	2.33
Dusk	-0.0660	-0.54	0.1702	1.99
Dark but lighted	0.0408	0.79	0.3107	8.17
Dark unknown lighting	-0.4524	-2.11	0.1755	1.12
Dark not lighted	-0.1942	-3.40	0.9977	19.13
Traffic Control				
Traffic Control: No Control (1/0)	-0.0611	-1.58	0.1555	5.26
Traffic Control: Human Control (1/0)	-0.3722	-2.24	-0.4140	-2.93
Socio-demographics				
Pedestrian is female (1/0)	0.1080	2.61	-0.1555	-5.09
Pedestrian age (Years)	-0.0069	-5.60	0.0139	12.47
Constant	Base	Base	0.3121	2.87
Threshold parameters				
μ_1	-1.5579	-17.96	-1.0475	-13.57
μ_2	-0.3846	-3.37	0.9428	8.94
μ_3	1.7275	12.55	1.8065	26.81
μ_4	2.5183	10.15	2.4038	34.59
Model Summary				
AIC	62572.39			
BIC	62937.88			
Log-likelihood at Convergence	-31241.2			
N	24,886			

driver or both pedestrian and driver were impaired with alcohol and drug. Referring to the results, several useful insights were obtained from the simple ordered Probit “baseline” model. For instance, the results of the baseline model suggest that chance of fatal injury and suspected serious injury to a pedestrian is significantly higher by 3.53% and 6.36% if a pedestrian makes recognition error and violation. These findings were expected as the recognition error by pedestrians include pedestrian dash and dart-out without realizing the oncoming vehicle due to which the driver could have no or very little chance to control his/her vehicle thus leading to serious pedestrian injuries. The results also suggest that performance error by drivers significantly increases the chance of a pedestrian being killed and seriously injured. This finding was expected as performance error by driver refers to the situation where the driver loses control of his/her vehicle. On other hand, recognition error by the driver is found to reduce the chance of a pedestrian being killed or seriously injured. While one would expect that recognition error by the driver could be associated with a higher chance of a pedestrian being killed or seriously injured, we note that recognition error by driver refers to a situation where the driver was backing his/her vehicle which usually has a lower speed. Referring to the effect of impairment with alcohol or drug on pedestrian injury, the results of the baseline model suggest that chance of pedestrian fatality increases by 7.83%, 3.04%, and 10.95% if only the driver was impaired and the pedestrian was sober, a pedestrian was impaired and the driver was sober, and both driver and pedestrian were impaired respectively. It can be noticed that the chance of pedestrian fatality is the highest if both pedestrian and driver are impaired with alcohol or drug followed by one only pedestrian is impaired with alcohol or drug. Studies suggest similar findings indicating that if a pedestrian is intoxicated, the chances of pedestrian fatality or severe injury significantly increase (Holubowycz 1995, Miles-Doan 1996, Öström and Eriksson 2001, Pour-Rouholamin and Zhou 2016, Kemnitzer *et al.* 2019). Referring to pedestrian demographics, the results suggest that chance of most severe pedestrian injury (fatality) increases with an increase in pedestrian age. For the lighting condition, the results reveal that compared to day light, the chance of pedestrian fatality increases by 2.87, and 7.84% of crashes occur in the night (dark) time with lighting and nighttime without lighting respectively. These findings are intuitive and refer to the visibility of pedestrians.

Referring to the results obtained from the FMOPM, there is significant statistical evidence of two unobserved (latent) classes across which the effects of the observed variables on pedestrian injury severity vary (Table 6.3). According to the post estimation results of the FMOPM, there is a 42.29% and 57.74% chance of pedestrians falling into Class 1 and Class 2 respectively (Table 6.4). The plots based on posterior probabilities of the two latent classes are symmetrical which indicates that the probability of pedestrian (or crash) belong to the two classes should add to 1 (Figure 6.3). Similar to Table 6.4, Figure 6.3 shows that

there are significant percentages of observations (pedestrians or crashes) that belong to Class 2 (red bars mostly exceed 0.5) compared to those which are predicted to be in class 1.

To compare the two latent classes based on the posterior probabilities of various severity levels of pedestrian injuries, the estimates from the FMOPM for class 1 predict only 0.47% ($= 0.0047 \times 100$) of the fatality (K) (Table 6.5). Similarly, the percentage of crashes with suspected serious injury (A) in class 1 is significantly lower (1.57%) than those obtained in class 2 (14.82%). The FMOPM estimates for Class 1 mostly predict suspected minor injuries (B). In summary, it can be said that Class 1 does not include (or predict) fatality (K) which could be regarded as relatively safer compared to Class 2. As discussed earlier, this could be due to some unobserved factors (based on which the pedestrians could be categorized). The estimates for Class 2 predicted 14.82% and 13.87% as fatal injury and suspected serious injury respectively which could be therefore regarded as “Fatality and Serious Injuries Group”. On other hand, in Class 1, only 3.39% of the fatality and suspected injuries are predicted using the estimates from the FMOPM for Class 1. Therefore, Class 1 could be regarded as minor and possible injuries group.

The probability distribution of various severity levels predicted based on the estimates of the base model and those of FMOPM models for the two different classes is illustrated in Figure 6.4. Referring to the estimation results of the FMOPM model, it can be seen that there is a significant variation in the effects of the observed factors. For instance, the results indicate that the probability of pedestrian being killed is significantly higher if a pedestrian-vehicle crash occur in dark not lighted conditions for pedestrians belong to Class 2. We note that in the same light condition (dark not lighted), the probability of pedestrian being killed significantly reduces in Class 1. Similar findings can be seen for dark lighted and dark unknown lighting conditions where the probability of pedestrian being killed is significantly higher in Class 2 compared to Class 1 (Table 6.3). These findings are in line with our hypothesis that Pedestrian in Class 1 may good visibility (with bright color clothes) compared to those in Class 2. These findings can also be due to the fact that pedestrian in Class 1 can be less riskier or hit with lower impact speed for which the information is not available in the dataset used for analysis; however, the FMOPM takes into account these factors thorough latent classes. Building upon these findings, our results indicate that if a pedestrian in Class 2 (for instance with lower conspicuity) makes violation, the probability of pedestrian being killed in Class 2 is significantly higher than those in Class 1 (refer to coefficients in Class 1 and Class2 in Table 6.3). Similarly, the coefficients of driver being impaired (but pedestrian is not), pedestrian being impaired (but driver is not), and both pedestrian and driver being impaired in Class 2 are significantly higher than their corresponding values in Class 1 indicating the while the probability of pedestrian being killed, due to impairment by either pedestrian or driver, could be higher for all pedestrians (both in Class 1 and Class 2),

Table 6. 4 Probabilities of Two Classes: Postestimation Results FMOPM

Class	Probability
1	0.4229
2	0.5774

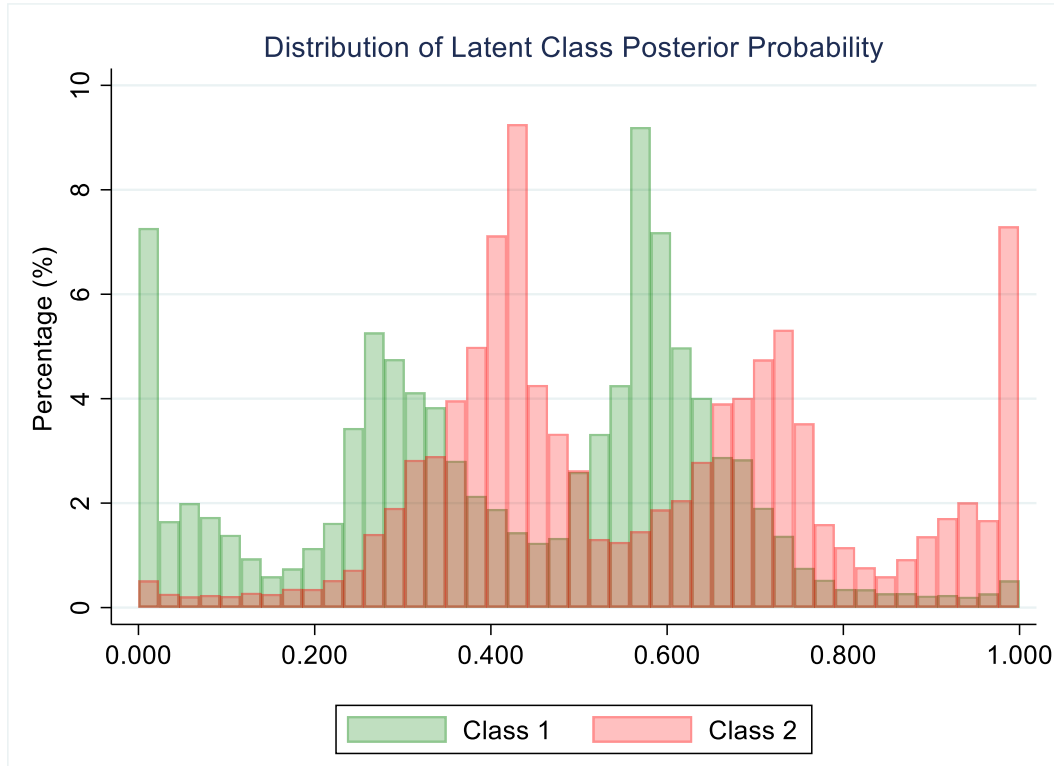


Figure 6. 3 Probability Distribution of Pedestrians (or Crashes): Two Latent Classes

Table 6. 5 Mean Posterior Probabilities of Pedestrian Injuries: Base versus FMOPM Models

Pedestrian Injury	Baseline Model	FMOPM	
		Class 1	Class 2
Fatality (K)	0.0881	0.0047	0.1482
Suspected Serious Injury (A)	0.0952	0.0292	0.1387
Suspected Minor Injury (B)	0.3950	0.5587	0.2709
Possible Injury (C)	0.3669	0.3234	0.4103
No Injury (O)	0.0546	0.0838	0.0317

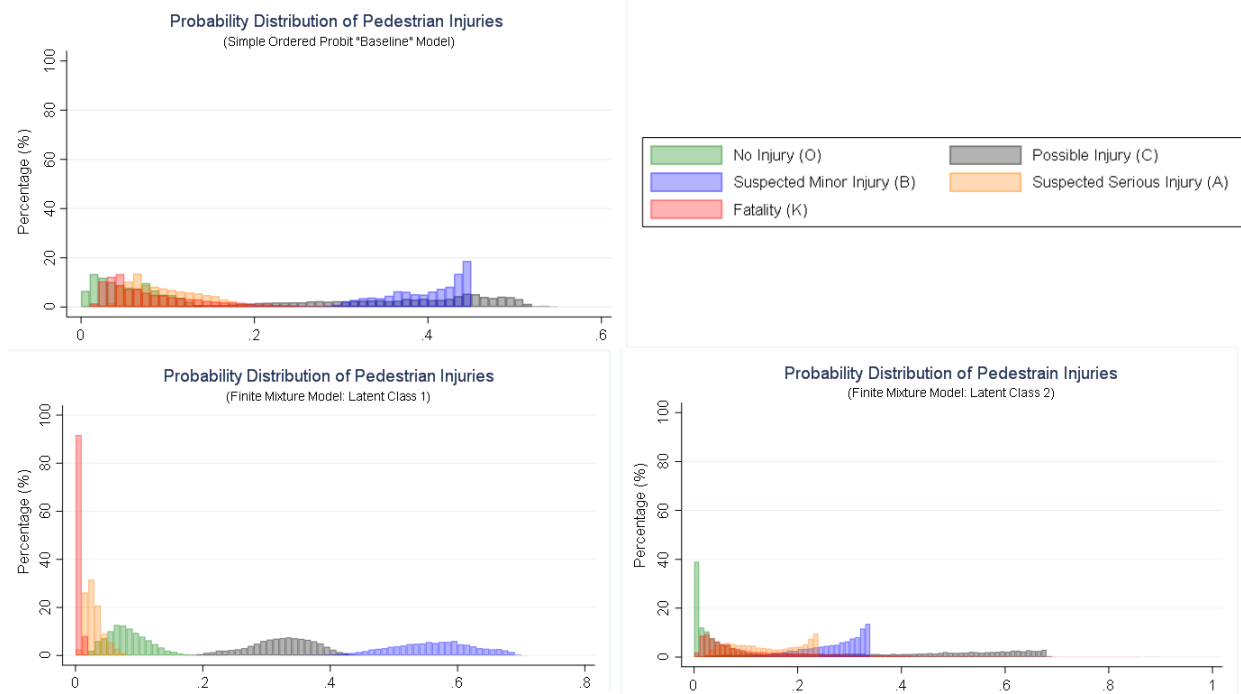


Figure 6. 4 Posterior Probabilities Distribution of Various Pedestrian Injury Severity Levels

it further increases for pedestrians belonging to Class 2 which in fact seems due to unobserved factor. For instance, if a pedestrian has lower conspicuity then impairment by driver or pedestrian could lead to more drastic outcomes for pedestrian in terms of fatality and serious injuries in pedestrian-vehicle crashes. To sum it up, our findings indicates that the effects of various key variables (light conditions, impairment, pedestrian behaviors) are different across different latent classes which if not taken into account, the estimates could be less accurate and unreliable.

Practical Implications

Referring to the practical implications of this research, we discuss certain countermeasure strategies which have the potential to reduce pedestrian-vehicle crashes and pedestrian injury severity as illustrated in Figure 6.5. This study has not investigated the countermeasures directly; however, based on the study findings, existing literature, and engineering judgment, we discuss some of the key countermeasures which could be considered for improving pedestrian safety especially in North Carolina. For instance, pedestrian recognition errors (during dash and dart-out) could be minimized by avoiding mobile phone use especially texting or browsing before crossing roadways as these may cause visual distraction leading to crashes. This could be achieved by educating the general public through awareness campaigns, by providing information through message signs, and use of hands-free (only for unavoidable and necessary calls) by pedestrians. For the vehicle, enhancing the pedestrian identification system and automatic braking system could compensate for both recognition errors by pedestrians and drivers. Note that our findings indicate that recognition errors by drivers are less likely to result in pedestrian fatalities and serious injuries; however, these errors in our data sample were made when the driver was backing where the speed could be lower. In general, we consider that using hands-free devices (including Siri and Alexa) by drivers could help reducing recognition errors which could lead to more severe pedestrian injuries at places where travel speed could be higher. Providing and improving message signs about distracted driving and pedestrian presence (in pedestrian-prone locations) could be beneficial in this regard.

Violations by both pedestrians and drivers could be potentially reduced through awareness seminars and campaigns. Such programs could be more beneficial for pedestrians who intentionally cross high-speed roadways like expressways, walk in the roadway, play in the roadway, and lay in the roadway. These programs could be focused to highlight the negative consequences of these activities by pedestrians which are found to be dangerous for pedestrians (given a crash) as per our findings. Note that improving the pedestrian identification system and automatic braking system could also be supportive in compensating for the violations by pedestrians. Referring to the light conditions, our results indicate that chance of fatal and serious injuries to pedestrians increases if crashes occur during nighttime (darkness) for which we



Figure 6. 5 Potential Countermeasures for Pedestrian-vehicle Crashes

Note: For picture courtesy, we provide links to the sources of the images below.

<https://www.bestbuy.com/site/shop/bluetooth-headset-for-phone>

<https://www.bosch-mobility-solutions.com/en/solutions/assistance-systems/adaptive-cruise-control/>

<https://pmrpressrelease.com/forward-collision-warning-systems-market-industry-perspective/>

<https://www.supaquick.com/news-article/730/distracted-driving-a-major-cause-of-road-accidents>

https://www.caricos.com/cars/t/toyota/2017_toyota_c-hr/images/137.html

<https://www.latimes.com/business/autos/la-fi-hy-auto-safety-20160317-story.html>

<https://www.safetyandhealthmagazine.com/articles/17348-impairment-is-impairment-study-finds-drugs-about-on-par-with-alcohol-as-cause-of-fatal-crashes>

<https://www.shutterstock.com/image-vector/no-texting-while-walking-use-smart-482023981>

consider that pedestrians may use bright color clothes especially at nighttime. Furthermore, the provision of street lighting in dense pedestrian areas could make it easier for drivers to perceive and recognize pedestrians thus may reduce driver recognition errors. Furthermore, based on the findings, we consider that provision of control devices especially stop-signs and stop and go signals in pedestrian areas could be beneficial. Finally, based on our findings, considering the reduction in speed limits and illumination solutions in highly populated pedestrian areas across the state could also be helpful (Prato *et al.* 2018).

Limitations

The findings in this study are based on the analysis of pedestrian-vehicle crashes in North Carolina. The key focus of this study is to see how errors and violations by both pedestrians and drivers relate to pedestrian injuries given a crash. Since the NC pedestrians and drivers could have their unique behaviors and socio-demographics; therefore, the findings from this study could not be extended to other parts of the U.S. However, we consider that similar regional or state-based findings could be obtained by replicating the methodology used in this study. Moreover, there could be other interesting factors in NC PBCAT data that could be explored to see their association with the severity of a pedestrian injury or driver injury given a crash. The focus of this study is more on human behaviors; therefore, we have included few other key factors in the analysis which showed a significant and intuitive correlation with pedestrian injury along with the human. In the future, it would be good to harness the unique NCPBCAT data for other important factors related especially related to roadway environment, roadway classes, land use, and driver demographics to see how these factors relate to pedestrian injury given a pedestrian-vehicle crash.

Conclusions

This study explores the effects of pre-crash errors and violations by both pedestrians and drivers on pedestrian injury severity given a crash. The unique aspect of this research is to test the hypothesis which assumes that there could be some unobserved factors (e.g., impact speed, pedestrian conspicuity, and aggressive/riskiness of drivers and/or pedestrians) based on which pedestrians could be split into unobserved (latent) groups. Both the simple ordered Probit “base” model and finite mixture ordered Probit model are used to model pedestrian injury severity given a crash. The findings indicate that errors and violations by pedestrians have contributed to around 50% of the crashes which is significantly higher than the errors and violations by drivers which have contributed to only 8% of these crashes. To sum it up, errors and violations by drivers and pedestrians have collectively contributed to 58% of pedestrian-vehicle crashes. It is important to mention that in 14.15% of the pedestrian-vehicle crashes, either pedestrian, driver, or both pedestrian and driver were impaired with alcohol or drugs. Referring to the estimation results, the

simple ordered Probit model reveals that there is a 6.36%, 5.35%, and 3.53% higher chance of pedestrian fatality if a pedestrian makes a violation, a driver makes performance error, and a pedestrian makes violation respectively. Results also indicate that chance of a pedestrian being killed or seriously injured is significantly higher by 10.95%, 7.83%, and 3.04% if both pedestrian and driver are impaired with alcohol or drugs if a pedestrian is impaired but the driver is sober, and if the driver is impaired and the driver is not respectively. These findings indicate that impairment by a pedestrian with alcohol or drugs could be more fatal. Based on the results, it was also found that compared to daylight, the chance of a pedestrian being killed and seriously injured is significantly higher if a pedestrian-vehicle crash occurs at night time, dusk, and dawn times; however, the chance of pedestrian is the highest if a crash occurs at night with no light. Importantly, the results of the finite mixture model with two latent classes show significant evidence of unobserved factors which could have significant effects on the pedestrian injury and could be used to categorize the pedestrians into two latent classes across which the magnitude (or even directions) of the parameters of the observed factors significantly vary. For instance, a crash occurring during darkness without lighting increases and decreases the chance of a pedestrian being killed in class 2 and class 1 respectively. This could be attributed to better conspicuity of pedestrians (who are more visible) due to brighter clothes in Class 1 than those in Class 2 which makes the individuals in Class 1 more visible and safer. Furthermore, this could also be due to safer unobserved behaviors of pedestrians in Class 1 compared to Class 2 as revealed by the posterior mean probabilities of fatal injury and serious injuries being significantly lower in Class 1 than Class 2. From a practical implication standpoint, we consider pedestrians with more bright outfits and safer behaviors could reduce their chance of fatality and/or serious injury in pedestrian-vehicle crashes. Furthermore, effective enforcement of drunk driving and drunk walking could help reduce serious and fatal injuries to pedestrians. The use of hands-free devices by pedestrians (for unavoidable calls only) during walking could be beneficial. To mitigate performance errors, multiple vehicle technologies including collision warning systems, automatic braking systems, and pedestrian detection systems could be helpful. From a policy standpoint, conducting awareness campaigns and seminars about the potential consequences of drug and alcohol use by both pedestrians and drivers could potentially reduce serious pedestrian injuries.

CHAPTER 7. OVERALL RESEARCH CONTRIBUTION AND PRACTICAL IMPLICATIONS

This dissertation generates new knowledge and foundation, mainly through real-world SHRP2-NDS data, which provide a deeper understanding of the role of driver behavior and performance in safety-critical events. The study findings provide useful insights which could help in selecting more effective countermeasures that could contribute to the Zero Vision via a safe system. The dissertation also unveils the role of human factors (errors, violations, and impairments by pedestrians and drivers) in pedestrian-vehicle crashes to see how these factors relate to pedestrian injury. Overall, the research findings aid to the existing state-of-the-art through various conceptual and methodological rigors as below.

- The dissertation reinforces the need to address human behavior and performance as the key drivers in road safety, the novel idea is to quantify pre-crash driving behaviors and secondary tasks (given that a particular secondary task has contributed to crashes) and bring them into a systematic behavioral framework. While previous studies have developed taxonomies for driving errors and violations (Treat *et al.* 1979, Reason *et al.* 1990, Petridou and Moustaki 2000, Wierwille *et al.* 2002, Stanton and Salmon 2009), these studies were based on police crash reports which could be subjective. On other hand, SHRP2-NDS was a unique data collection effort that provide real-world data on pre-crash driving behaviors and performance. Using SHRP2-NDS data, a different set of driving errors are identified as contributing factors in crashes compared to the past. One of the unique aspects of this dissertation is that it utilizes the baselines (non-event driving) in the SHRP2-NDS data to compute the crash risk associated with key driving errors, violations, and diverse roadway environments which could not be possible using police crash data.
- While previous studies have investigated the effects of roadway environment on crash occurrence (Dumbaugh and Li 2010), none have taken into account the indirect effects of diverse roadway environments on SCEs through one or more driving errors and violations. The key rationale behind this research was that if significant indirect effects of a specific roadway environment on SCEs through one or more driving errors and violations exist but are not considered, the overall effects of the roadway environment on SCEs could not be captured. Since such complex relationships could not be unveiled through simple statistical models, we have applied rigorous path analysis to quantify both direct and indirect pathways that lead from diverse roadway environments to SCEs. From the methodological standpoint, besides the rigorous path analytic framework, this dissertation relaxes the parallel assumption (for those explanatory variables which violate the assumption) through a more sophisticated partially proportional odds model in the path analysis which links the diverse roadway environments, and driving errors and violations to SCEs.
- Furthermore, one of the unique aspects of SHRP2-NDS is that it includes data about distraction duration, instantaneous speed, instantaneous acceleration, and instantaneous deceleration where the

later three could be used to derive their respective volatility measures. Building upon the newly developed taxonomy of naturalistic driving errors and violations (Khattak et al. 2021a) along with the volatility measures, this dissertation contributes by investigating how driving errors, violations, and duration of distraction relate to SCEs directly and through instability in driving which was unexplored to date. From the methodological standpoint, we have used the joint estimation procedure which accounts for the potential correlation between the unobserved factors associated with both instability in driving and the event outcome (baseline, near-crash, and crash). We note that most of the relevant studies have assumed that the unobserved factors in the two models (in path analysis) were uncorrelated (Liu and Khattak 2017, 2018, Arvin et al. 2019). We note that if a significant correlation between the unobserved factors in the two models exist but is not considered, this could lead to inefficient estimates (Ahmad et al. 2021). Furthermore, studies have quantified only the direct effects of duration of distraction on crashes (Simons-Morton et al. 2014, Arvin and Khattak 2020), we consider that there could be significant indirect effects of distraction duration on crashes (and near-crashes) through driving instability which should be accounted to understand the overall effects of distraction duration on SCEs. In this regard, one of the unique aspects of this research is to account for the non-linear association of duration of distraction with SCEs both directly and through driving instability in a joint path analytic framework which none of the studies have accounted for to date.

- After understanding the role of driver behavior and performance in SCEs via sophisticated statistical models, the dissertation aims to assess the significance of driver behavior and performance in predicting SCEs using machine learning or artificial intelligence methods. The novel aspect is that this research not only predicts SCEs based on human factors via artificial intelligence methods but also develops a model which could link all the available human factors (including driving errors, violations, duration of distraction, and instability in driving) to SCEs in diverse roadway environments – something which was overlooked till date. Using both frequentist and machine learning methods (Naïve Bayes, KNN, and gradient boosting tree classifications), the key aims were 1) to select a more accurate model that could be used to predict SCEs using human factors, and 2) to assess the importance of driver behavior and performance-related factors used in the prediction of SCEs compared to roadway environmental factors.
- Finally, this dissertation investigates how human factors (errors, violations, and impairment by pedestrians and drivers) relate to pedestrian injuries in pedestrian-vehicle crashes. The research contributes to the state of art in several ways. First, one of the unique aspects is the use of NCPBCAT data which includes comprehensive information about pre-crash behaviors by pedestrians and drivers extracted by professional experts from crash narratives in police crash reports. Second, while previous studies have acknowledged the effects of individual pedestrian behaviors (Stavrinos et al. 2011,

Byington and Schwebel 2013, Nasar and Troyer 2013, Kemnitzer et al. 2019) and pedestrian conditions (Holubowycz 1995, Miles-Doan 1996, Öström and Eriksson 2001, Pour-Rouholamin and Zhou 2016, Kemnitzer et al. 2019) on pedestrian injuries, this dissertation has looked at the pre-crash behaviors of both pedestrians and drivers in a comprehensive manner by bringing them into systematic frameworks. Finally, the unique methodological aspect of this research is the use of a finite mixture model to capture heterogeneity in pedestrian injuries across different latent (unobserved) classes based on some unobserved variables (e.g., aggressiveness by pedestrians and/or driver, conspicuity of pedestrians, and impact speed) – something which is usually overlooked.

Based on the thorough empirical analyses in this dissertation, several useful insights were obtained which highlight the role of human factors in road crashes. These findings provide useful knowledge and foundation that could help in selecting appropriate countermeasures to enhance various elements of the safe system. To spot the key findings based on SHRP2-NDS data, please refer to Table 7.1. For instance, based on our findings, in a school zone, business or industrial area, urban location, there is a higher chance of SCEs. However, the three roadway types lead to higher instability in driving and a higher chance of recognition errors and decision errors where the two errors and instability in driving themselves are associated with a higher chance of SCEs. The results also indicate that drivers distracted for a longer duration have a higher chance of SCEs (crash and near-crash) where the chance of SCEs further increase if the effects of duration of distraction on SCEs through recognition error, decision error, performance error, and instability in driving are considered. These findings provide a foundation that could be useful in selecting effective countermeasures which could potentially improve safety through multiple protective layers of the safe system (Table 7.2). For instance, the ratio of safer road users can be increased through campaigns and seminars while educating the general public (especially drivers) about the negative consequences of driving errors, intentional violations, duration of distraction, and instability in driving. Furthermore, driver assistance systems have the potential to reduce driving errors, intentional violations, and driving instability thus improving road safety through safer road users, safer vehicles, and safer speed (Table 7.2). For instance, studies reveal that the forward collision warning systems (Ervin 2005, Rakha 2011), adaptive cruise control (Ervin 2005, Rakha 2011), blind-spot assistance systems (Liu et al. 2017a, Sossa and Zamora 2020), and lane tracking systems (Ervin 2005, Hanowski et al. 2008, Lerner et al. 2010, Rakha 2011) have the potential to mitigate both recognition and decision errors by drivers. Similarly, forward collision warning systems (Rakha 2011), adaptive cruise control (McDonald et al. 2001, Dang et al. 2015), and lane tracking systems (Ervin 2005, Hanowski et al. 2008, Lerner et al. 2010, Rakha 2011) have the potential to reduce performance errors by drivers. Referring to violations, studies suggest that dilemma zone mitigation systems have the potential to reduce violations by subject drivers (U.S. DOT

2009, Rakha 2011). For a detailed discussion about the effectiveness of various driver assistance systems in mitigating one or more driving errors and violations, please refer to Chapter 2 (Khattak et al. 2021a) and Chapter 3 (Ahmad et al. 2021). Referring to the safer speed, driver assistance systems especially adaptive cruise control can help in maintaining constant speed thus reducing speed instability of the subject driver. We also consider that monitoring speed variation in real-time and generating warning to drivers who deviates from his/her normal speed range could help to ensure safer speed by the drivers. Furthermore, we consider that the objective of safer road users can be achieved through use of hands-free devices (instead of texting or holding cellphone) by drivers for urgent calls only during driving. Referring to the safer roadway environments, crashes (and near-crashes) can be reduced or eliminated by removing external sources of distraction (e.g., billboards) and by improving signs (both fixed and variable messages signs to warn drivers about the negative effects of risky behaviors like distracted driving).

Based on NCPBCAT data, our findings indicate that chance of a pedestrian being killed and seriously injured is significantly higher if a pedestrian makes a recognition error or violation, a driver makes a performance error, and either pedestrian or driver is impaired due to alcohol or drugs. While the results of the traditional ordered Probit model indicate that chance of a pedestrian being killed and seriously injured is significantly higher if a crash occurs in dark light conditions (compared to daytime) for all the observations, the finite mixture model shows that the effects of darkness on pedestrian injury are different across the observations based on some unobserved factors. Using the finite mixture model, we have found sufficient statistical evidence that pedestrians could be split into two latent classes across which the effects of observed factors on pedestrian injury severity could not be the same. For instance, in Class 2, the dark light conditions have even a higher chance of fatality and serious injuries compared to Class 1. This could be due to the better conspicuity, less risky behaviors, or lower impact speed (all the three factors are unobserved in the data) for pedestrians who belong to Class 1 compared to Class 2. Furthermore, violations by pedestrians and impairment by either pedestrian or driver are associated with even a higher chance of a pedestrian being killed which further increases if pedestrians belong to Class 2. We note that our findings indicate that the posterior probabilities of fatal injury and suspected serious injury to a pedestrian in Class 2 are significantly higher than Class 1 suggesting that Class 2 is more unsafe. Based on the class probabilities obtained via the finite mixture model, 57% of the pedestrians belong to Class 2 which is alarming. Referring to the practical implications, we consider that educating pedestrians about the potential benefits of using brighter clothes (at nighttime) and negative consequences of risky or aggressive behaviors by pedestrians could improve pedestrian safety by increasing the ratio of safer pedestrians. Similarly, effective enforcement of laws related to drunk driving and drunk walking could potentially increase the ratio of sober and safer pedestrians thus mitigating pedestrians being killed or seriously injured. From the technological standpoint, the use of hands-free devices (only for necessary calls) by pedestrians could help

in avoiding distracted walking due to cellphone texting or browsing. To mitigate the performance errors by drivers which lead to fatality and serious injuries to pedestrians, we consider that driver assistance systems including automatic brake systems, collision warning systems, and pedestrian detection systems can potentially enhance pedestrian safety. As discussed, all of the aforementioned countermeasures have the potential to improve safety; however, we consider that there is a need for further research prior to firm recommendation of implementing them.

Table 7. 1 Key Relationships of Human Factors and Roadway Environments with SCEs: Insights from SHRP2-NDS data

Explanatory Variables	Outcomes							
	Driving Errors and Violations						Driving Instability	Crash Risk
Roadway Environment	Recognition Errors	Decision Errors	Performance Errors	Physical Condition Errors	Experience Errors	Violations		
School zones	+	+	---	---	---	---	+	+
Urban areas	+	+	+	---	---	---	+	+
Business or Industrial locations	+	+	+	---	+	---	+	+
Driving Errors and Violations								
Recognition Errors	---	---	---	---	---	---	+	+
Decision Errors	---	---	---	---	---	---	+	+
Performance Errors	---	---	---	---	---	---	+	+
Physical Condition Errors	---	---	---	---	---	---	+	+
Experience Errors	---	---	---	---	---	---	+	+
Violations	---	---	---	---	---	---	+	+
Driving Instability	---	---	---	---	---	---	---	+
Duration of Distraction	+	+	+	---	+	---	+	+

Table 7. 2 Potential Countermeasures: Improving Elements of Safe Systems

Elements of Safe Systems	Potential Countermeasures				
	Educational Campaigns	Driver Assistance Systems	Hands-free devices	Improving Signs	Roadway Environment (Removing billboards)
Safer Roadway Environment	---	---	---	✓	✓
Safer Road Users	✓	✓	✓	---	✓
Safer Vehicles	---	✓	---	---	---
Safer Speed	---	✓	---	---	---

CHAPTER 8. CONCLUSIONS

Human factors including driving errors and violations are highly relevant to the safe system approach as these factors tend to contribute to a majority of road crashes. The safe system approach to roadway safety is a more holistic procedure that requires safety experts to access the roadway system more comprehensively. For instance, roadway safety could be substantially improved through multiple layers of protection for roadway users which include safe road users, safe vehicles, safe speed, safe roads, safe roadway environment, and an appropriate post-crash care system. Road safety can be improved by applying the safe systems approach to harness new forms of real-world data. Contrary to crash data obtained through police crash reports, emerging data sources including SHRP2-NDS data provide opportunities to explore pre-crash behaviors and performance from unique perspectives which could not be possible through conventional police crash data. The SHRP2-NDS data include real-world microscopic details about pre-crash driving behaviors, performance, distractions (and their durations), and vehicle kinematics (speed, acceleration, and deceleration) of the subject (in instrumented vehicle) drivers. In SHRP2-NDS data, all of the aforementioned real-world data is continuously captured through advanced equipment, including video cameras and radars, in the subject vehicles. Moreover, one of the unique features of SHRP2-NDS data is including information about baseline (non-event driving) which could be used to quantify crash risk and near-crash risk. All of the aforementioned features make SHRP2-NDS data more unique and realistic which could be analyzed to generate knowledge related to driving behaviors in SCEs (crashes and near-crashes). Taking advantage of the unique SHRP2-NDS data, this study explores the role of driving behaviors and performance in SCEs at various levels.

Through a safe system approach, this dissertation aims to investigate the role of human factors (driving errors and violations), driving instability, and roadway “built” environment in SCEs. Using real-world data, the key objective of this dissertation is to generate new knowledge which could help in selecting more effective countermeasures which would aid to Vision Zero through safe road users, safe speeds, and safe roadway environments. Harnessing SHRP2-NDS data, this dissertation will make a timely contribution by developing a systematic behavioral framework for all pre-crash driving behaviors which contribute to crashes. Continuing the analysis, this dissertation will apply the taxonomy to empirical data to understand how diverse roadway environments relate to different driving errors and violations. Our findings obtained through rigorous path analysis suggest that school zones, business/industrial locations, and urban areas induce crash risk not only directly but also through driving errors and violations making the overall effect on crash occurrence even higher. For instance, our findings suggest that chance of subject drivers to make recognition errors increases by 4.78%, 2.51%, and 8.00% on roadways passing through school zones, urban areas, and business/industrial locations respectively. The three roadway environments are also found to increase the chance of decision error by the subject drivers. The findings are relevant to safe road users and

safe roadway environments as they provide new knowledge and foundation which could be used to select appropriate countermeasures for reducing the crash risk associated with driving errors, violations, and hotspot locations. Referring to safe speed which could have a significant role in a safe roadway system, this dissertation uses new measures of driving instability (speed volatility) which is then used to discover multidimensional connections from driving errors, violations, and roadway environment to SCEs through driving instability. For instance, our findings suggest that school zones, business/industrial locations, and urban areas increase speed volatility which in turn is positively associated with SCEs. Furthermore, the results suggest that all driving errors and violations lead to higher speed volatility thus inducing additional indirect crash risk. For instance, our findings reveal that recognition error (compared to no driving error and violation) increases the crash risk by 6.54% (direct effects); however, it further increases the crash risk (through COV of speed) by 5.01% leading to an overall increase in crash risk by 11.55%. In the context of a safe system, these findings related to the effects of roadway environment and human behavior on SCEs, through driving instability, are useful and could be helpful to achieve goals relevant to safe speed on roadways. Realizing the significant negative role of distraction duration in SCEs, this dissertation looked at the effects of distraction duration (seconds) on SCEs both directly and through driving instability. In this regard, our findings reveal that if a subject driver is distracted for 10 seconds, his/her chance of involving in a crash is 16.88% which is based on the direct effect of distraction duration on a crash. However, if its indirect effects on crash risk through speed volatility are considered, the overall chance of the driver to involve in a crash is significantly higher (22.03%). Such findings could only be explored through rigorous path analysis. The findings also show significant statistical evidence about non-linearity in the effects of duration of distraction on both instability in driving and the chance of SCEs. After investigating the nature and magnitude of human factors (driving errors, violations, duration of distraction, and instability in driving) on SCEs, an additional research effort was made to predict SCEs based on the human factors in diverse roadway and environments. Since conventional statistical models are found weaker in out-of-sample prediction, machine learning methods including Naïve Bayes, K-Nearest Neighbor, and Gradient Boosting Tree classification algorithms were applied. The findings indicate that the gradient boosting tree classifier has the highest out-of-sample prediction accuracy of 91.23%. Furthermore, the gradient boosting tree classifier also showed the highest out-of-sample prediction accuracy for SCEs including crash and near-crash. Referring to the feature importance plot of the gradient boosting classifier, the COV of speed is found to be the most important feature followed by COV of deceleration, distraction duration, recognition errors, and COV of negative jerk used in predicting the event outcome. In summary, the findings in this study suggest that human factors are critical and significantly increase the chance of SCEs which if mitigated could enhance safety. After understanding the effects of human behaviors and performance on crash risk and near-crash risk, this dissertation aims to explore the links of human factors to injury severity

of vulnerable road users. Due to data limitation of SHRP2-NDS regarding injury severity (given a crash) and pedestrian-vehicle crashes, this dissertation will harness conventional NC-PBCAT data to see how human factors (errors, violations, and impairments by both pedestrians and drivers) relate to pedestrian injury severity given a crash. Referring to the overall analysis conducted in this dissertation, various approaches including standard path analysis, joint estimation procedure, path analysis through predictive margins, and finite mixture models will be used to spot the multidimensional pathways which lead from human factors (along with other key factors) to safety-critical outcomes.

From the prediction standpoint, we consider that the findings in this dissertation especially those obtained through the machine learning models could be used to develop the AI algorithms which can compute and reveal the risk of SCEs in real-time using data on pre-crash behaviors and performance of the driver in diverse roadway environments. However, prior to developing the AI algorithms which were trained in this study via driving errors and violations obtained through the TDEV, the TDEV should be programmed in the algorithm which could continuously classify the pre-crash driver behavior into various types of errors and violations using real-time data. Furthermore, based on the findings in Chapter 3, various roadway environments (e.g., school zone, urban areas, and business and industrial locations) could lead to recognition and decision errors. These findings could be used in the development of AI algorithms that could predict the real-time chance of specific driving errors and violations when the driver is traveling in a specific roadway environment. However, the key limitation in this regard is that the SHRP2-NDS data which is analyzed in this dissertation refers to a few states in the U.S. and may not be applicable in other areas due to potential variations in driver demographics, driver behaviors, and roadway and environment conditions. Using the real-time driver behavior and performance data, the ML algorithm could be trained in a manner that could generate timely warnings to the subject drivers which could help them in taking timely actions to overcome their risky behaviors which otherwise could lead to SCEs. While automated vehicles (AVs) and/or Connected Vehicles (CVs) are widely considered as the most effective solution that could significantly reduce or eliminate driving errors and violations, such vehicles could still lead to hazardous scenarios which need further research. We consider that the key driver errors (recognition errors, decision errors, and performance errors) and violations discussed in this dissertation could be still expected in AVs and even CVs where in the latter two cases such errors and violations could be more drastic. For instance, the AV may fail to recognize or unintentionally misclassify an object (or hazard) in the roadway resulting in recognition error due to the system (NTSB, 2017, McDonald, 2019, Louw, 2019). Such recognition errors can also occur due to the malfunction of AVs or their limitations. Similar to recognition errors, AVs can be prone to decision errors, performance errors, and violations. As mentioned, errors due to the malfunction or limitation of AVs/CVs could be more problematic as in such scenarios these vehicles

may not be able to generate warnings to the driver. If such “silent failures” due to machine failure occur, the responsibility may fall on the subject driver who must detect that the system has failed. The important question that needs further research is to investigate driver behaviors (errors and violations) when some control is given to the vehicles. For instance, if the subject driver is also distracted and is unable to detect the silent failure when no warning is issued by AVs/CVs, it would be a riskier situation compared to when the system fails but the driver is active enough to detect the problem and take control. To sum it up, while the current study provides useful knowledge of naturalistic driving errors and violations which can provide useful knowledge for developing AVs/CVs, still further research is needed to see the behavior and performance of drivers in partially automated vehicles. This would not only help in understanding the predominant malfunctions and silent failures of AVs/CVs but also provide useful insights about the corresponding errors and violations by drivers in such scenarios. Investigating errors and violations by drivers and partially automated vehicles could provide insights which could be more beneficial in improving the safety performance of more advanced AVs and even CVs in future.

REFERENCES

- Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research* 34 (5), 597-603.
- Ahmad, N., Ahmed, A., Shah, A.A., 2019a. Is effective enforcement of seatbelt law helpful in reduction of road crash fatalities: An exploratory empirical analysis using aggregate data. *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences* 56 (4), 31–41-31–41.
- Ahmad, N., Ahmed, A., Wali, B., Saeed, T.U., 2019b. Exploring factors associated with crash severity on motorways in Pakistan. *Proceedings of the Institution of Civil Engineers-Transport*, 1-10.
- Ahmad, N., Wali, B., Khattak, A.J., Dumbaugh, E., 2021. Built environment, driving errors and violations, and crashes in naturalistic driving environment. *Accident Analysis & Prevention* 157, 106158.
- Ahmad, N., Wali, B., Khattak, A. J., & Dumbaugh, E., 2021. Built environment, driving errors and violations, and crashes in naturalistic driving environment. *Accident analysis & prevention*. *Accident Analysis & Prevention* (157), 106158.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6), 716-723.
- Alioua, N., Amine, A., Rziza, M., Aboutajdine, D., 2011. Driver's fatigue and drowsiness detection to reduce traffic accidents on road. In: *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pp. 397-404.
- Anastasopoulos, P.C., 2016. Random parameters multivariate Tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. *Analytic Methods in Accident Research* 11, 17-32.
- Anastasopoulos, P.C., Mannering, F.L., Shankar, V.N., Haddock, J.E., 2012. A study of factors affecting highway accident rates using the random-parameters Tobit model. *Accident Analysis & Prevention* 45, 628-633.
- Anastasopoulos, P.C., Tarko, A.P., Mannering, F.L., 2008. Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis & Prevention* 40 (2), 768-775.
- Arvin, R., Kamrani, M., Khattak, A.J., 2019a. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. *Accident Analysis & Prevention* 132, 105226.
- Arvin, R., Khattak, A.J., 2020. Driving impairments and duration of distractions: Assessing crash risk by harnessing microscopic naturalistic driving data. *Accident Analysis & Prevention* 146, 105733.
- Bao, F., Huang, B., Year. The design of vehicle low-speed intelligent braking system. In: *Proceedings of the 3rd International Conference on Mechatronics, Robotics and Automation*.
- Bhat, C.R., Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological* 43 (7), 749-765.

- Blanco, M., Hickman, J.S., Olson, R.L., Bocanegra, J.L., Greening, M., Madison, P., Holbrook, G.T., Hanowski, R.J., Nakata, A., Bowman, D.S., 2016. Investigating critical incidents, driver restart period, sleep quantity, and crash countermeasures in commercial vehicle operations using naturalistic data collection. United States. Department of Transportation. Federal Motor Carrier Safety Administration.
- Blatt, A., Pierowicz, J., Flanigan, M., Lin, P.-S., Kourtellis, A., Lee, C., Jovanis, P., Jenness, J., Wilaby, M., Campbell, J., 2015. Naturalistic driving study: Field data collection.
- Boggs, A.M., Wali, B., Khattak, A.J., 2020. Exploratory analysis of automated vehicle crashes in California: A text analytics & hierarchical bayesian heterogeneity-based approach. *Accident Analysis & Prevention* 135, 105354.
- Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52 (3), 345-370.
- Bozdogan, H., 1990. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics-Theory and Methods* 19 (1), 221-278.
- Byington, K.W., Schwebel, D.C., 2013. Effects of mobile internet use on college student pedestrian injury risk. *Accident Analysis & Prevention* 51, 78-83.
- Campbell, K.L., 2012. The SHRP 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. *TRB News* (282).
- Carney, C., Mcgehee, D., Harland, K., Weiss, M., Raby, M., 2015a. Using naturalistic driving data to assess the prevalence of environmental factors and driver behaviors in teen driver crashes.
- Casado-Sanz, N., Guirao, B., Lara Galera, A., Attard, M., 2019. Investigating the risk factors associated with the severity of the pedestrians injured on Spanish crosstown roads. *Sustainability* 11 (19), 5194.
- Celik, A.K., Oktay, E., 2014. A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars provinces of turkey. *Accident Analysis & Prevention* 72, 66-77.
- Chen, Z., Wu, C., Huang, Z., Lyu, N., Hu, Z., Zhong, M., Cheng, Y., Ran, B., 2017. Dangerous driving behavior detection using video-extracted vehicle trajectory histograms. *Journal of Intelligent Transportation Systems* 21 (5), 409-421.
- Choudhary, P., Velaga, N.R., 2017a. Analysis of vehicle-based lateral performance measures during distracted driving due to phone use. *Transportation Research Part F: Traffic Psychology and Behaviour* 44, 120-133.
- Choudhary, P., Velaga, N.R., 2017b. Modeling driver distraction effects due to mobile phone use on reaction time. *Transportation Research Part C: Emerging Technologies* 77, 351-365.

- Chovan, J., Tijerina, L., Alexander, G., Hendricks, D., 1994. Examination of lane change crashes and potential countermeasures. Final report.
- Cicchino, J.B., 2018. Effects of blind spot monitoring systems on police-reported lane-change crashes. *Traffic Injury Prevention* 19 (6), 615-622.
- Cicchino, J.B., 2019. Real-world effects of rear automatic braking and other backing assistance systems. *Journal of Safety Research* 68, 41-47.
- Clifton, K.J., Burnier, C.V., Akar, G., 2009. Severity of injury resulting from pedestrian–vehicle crashes: What can we learn from examining the built environment? *Transportation Research Part D: Transport and Environment* 14 (6), 425-436.
- Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory* 13 (1), 21-27.
- Dang, R., Wang, J., Li, S.E., Li, K., 2015. Coordinated adaptive cruise control system with lane-change assistance. *IEEE Transactions on Intelligent Transportation Systems* 16 (5), 2373-2383.
- Debnath, A.K., Blackman, R., Haworth, N., 2014. A Tobit model for analyzing speed limit compliance in work zones. *Safety Science* 70, 367-377.
- 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* 113 (10), 2636-2641.
- Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., 2006a. The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment. United States. Department of Transportation. National Highway Traffic Safety Administration.
- Dingus, T.A., Neale, V.L., Klauer, S.G., Petersen, A.D., Carroll, R.J., 2006c. The development of a naturalistic data collection system to perform critical incident analysis: An investigation of safety and fatigue issues in long-haul trucking. *Accident Analysis & Prevention* 38 (6), 1127-1136.
- Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object–passenger car crashes. *Accident Analysis & Prevention* 34 (5), 609-618.
- Distner, M., Bengtsson, M., Broberg, T., Jakobsson, L., Year. City safety—a system addressing rear-end collisions at low speeds. In: *Proceedings of the Proc. 21st International Technical Conference on the Enhanced Safety of Vehicles*.
- Dumbaugh, E., Li, W., 2010. Designing for the safety of pedestrians, cyclists, and motorists in urban environments. *Journal of the American Planning Association* 77 (1), 69-88.
- Eluru, N., Paleti, R., Pendyala, R.M., Bhat, C.R., 2010. Modeling multiple vehicle occupant injury severity: A copula-based multivariate approach. *Transportation Research Record* 2165, 1-11.

- Ervin, R., 2005. Automotive collision avoidance system field operational test methodology and results, volume 1: Technical report. University of Michigan, Ann Arbor, Transportation Research Institute.
- Farmer, C.M., 2008. Crash avoidance potential of five vehicle technologies. Insurance Institute for Highway Safety.
- FHWA, Facts and statistics work zone injuries and fatalities
- FHWA, 1989. FHWA functional classification guidelines. US Federal Highway Administration, Washington, DC.
- Frawley, W.E., Eisele, W.L., 2004. Investigation of access point density and raised medians: Crash analysis and micro-simulation.
- Fu, X., Nie, Q., Liu, J., Khattak, A., Hainen, A., Nambisan, S., 2021. Constructing spatiotemporal driving volatility profiles for connected and automated vehicles in existing highway networks. *Journal of Intelligent Transportation Systems*, 1-14.
- Gaspar, J.G., Schwarz, C.W., Brown, T.L., Kang, J., 2019. Gaze position modulates the effectiveness of forward collision warnings for drowsy drivers. *Accident Analysis & Prevention* 126, 25-30.
- Gershon, P., Sita, K.R., Zhu, C., Ehsani, J.P., Klauer, S.G., Dingus, T.A., Simons-Morton, B.G., 2019. Distracted driving, visual inattention, and crash risk among teenage drivers. *American Journal of Preventive Medicine* 56 (4), 494-500.
- Gkritza, K., Kinzenbaw, C.R., Hallmark, S., Hawkins, N., 2010. An empirical analysis of farm vehicle crash injury severities on Iowa's public road system. *Accident Analysis & Prevention* 42 (4), 1392-1397.
- Greene, W.H., 2003. *Econometric analysis*, prentice hall. New Jersey, 16.
- Gregory, B., Irwin, J.D., Faulks, I.J., Chekaluk, E., 2014. Speeding in school zones: Violation or lapse in prospective memory? *Journal of Experimental Psychology: Applied* 20 (3), 191.
- Guo, F., Klauer, S.G., Fang, Y., Hankey, J.M., Antin, J.F., Perez, M.A., Lee, S.E., Dingus, T.A., 2017. The effects of age on crash risk associated with driver distraction. *International Journal of Epidemiology* 46 (1), 258-265.
- 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. Virginia Tech Transportation Institute.
- Hanowski, R.J., Blanco, M., Nakata, A., Hickman, J.S., Schaudt, W.A., Fumero, M., Olson, R.L., Jermeland, J., Greening, M., Holbrook, G., 2008. The drowsy driver warning system field operational test: Data collection methods.
- Harbluk, J.L., Noy, Y.I., Trbovich, P.L., Eizenman, M., 2007. An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis & Prevention* 39 (2), 372-379.

- Harkey, D., 2020. Accelerating automated vehicle acceptance. TRB Webinar, National Academy of Sciences Washington D.C.
- Harkey, D.L., Tsai, S., Thomas, L., Hunter, W.W., 2006. Pbcats-pedestrian and bicycle crash analysis tool version 2.0. United States. Federal Highway Administration.
- Hlaing, K.P., Aung, N.T.T., Hlaing, S.Z., Ochimizu, K., 2019. Analysis of accident severity factor in road accident of Yangon using FRAM and classification technique. In: Proceedings of the 2019 International Conference on Advanced Information Technologies (ICAIT), pp. 256-261.
- Holubowycz, O.T., 1995. Age, sex, and blood alcohol concentration of killed and injured pedestrians. *Accident Analysis & Prevention* 27 (3), 417-422.
- Huang, H., Abdel-Aty, M.A., Darwiche, A.L., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. *Transportation Research Record* 2148 (1), 27-37.
- Iden, R., Shappell, S.A., Year. A human error analysis of us fatal highway crashes 1990–2004. In: Proceedings of the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 2000-2003.
- Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention* 108, 27-36.
- Isaksson-Hellman, I., Lindman, M., Year. The effect of a low-speed automatic brake system estimated from real life data. In: Proceedings of the Annals of Advances in Automotive Medicine/Annual Scientific Conference, pp. 231.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning Springer.
- Jermakian, J.S., 2011. Crash avoidance potential of four passenger vehicle technologies. *Accident Analysis & Prevention* 43 (3), 732-740.
- Jha, A.N., Chatterjee, N., Tiwari, G., 2021. A performance analysis of prediction techniques for impacting vehicles in hit-and-run road accidents. *Accident Analysis & Prevention* 157, 106164.
- Jing-Liang, L., Zhao-Du, L., Year. Design of acc controller based on double level Mamdani reasoning. In: Proceedings of the 2009 WRI Global Congress on Intelligent Systems, pp. 233-237.
- Juvale, H.B., Mahajan, A.S., Bhagwat, A.A., Badiger, V.T., Bhutkar, G.D., Dhabe, P.S., Dhore, M.L., 2009. Design of drodeasys (drowsy detection and alarming system). *Advances in computational algorithms and data analysis*. Springer, pp. 75-79.
- 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.
- Kemnitzer, C.R., Pope, C.N., Nwosu, A., Zhao, S., Wei, L., Zhu, M., 2019. An investigation of driver, pedestrian, and environmental characteristics and resulting pedestrian injury. *Traffic Injury Prevention* 20 (5), 510-514.

- Khattak, A.J., Ahmad, N., Wali, B., Dumbaugh, E., 2021a. A taxonomy of driving errors and violations: Evidence from the naturalistic driving study. *Accident Analysis & Prevention* 151, 105873.
- Khattak, A.J., And, B.W., Ahmad, N., 2019. A taxonomy of naturalistic driving errors and violations and its variations across different land-use contexts – a path analysis approach. *Transportation Research Board 98th Annual Meeting Transportation Research Board. National Academy of Sciences, Washington D.C.*
- Khattak, A.J., Khattak, A.J., Council, F.M., 2002. Effects of work zone presence on injury and non-injury crashes. *Accident Analysis & Prevention* 34 (1), 19-29.
- Khattak, Z.H., Fontaine, M.D., Li, W., Khattak, A.J., Karnowski, T., 2021b. Investigating the relation between instantaneous driving decisions and safety critical events in naturalistic driving environment. *Accident Analysis & Prevention* 156, 106086.
- Khera, D., Singh, W., 2015. Prediction and analysis of injury severity in traffic system using data mining techniques. *Int J Comput Applic* 7, 1-7.
- 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.
- Klauer, S.G., Guo, F., Simons-Morton, B.G., Ouimet, M.C., Lee, S.E., Dingus, T.A., 2014. Distracted driving and risk of road crashes among novice and experienced drivers. *New England Journal of Medicine* 370 (1), 54-59.
- Klauer, S.G., Guo, F., Sudweeks, J., Dingus, T.A., 2010. An analysis of driver inattention using a case-crossover approach on 100-car data.
- Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: An application of ordered probit models. *Accident Analysis & Prevention* 34 (3), 313-321.
- Kummetha, V.C., Kondyli, A., Schrock, S.D., 2020. Analysis of the effects of adaptive cruise control on driver behavior and awareness using a driving simulator. *Journal of Transportation Safety & Security* 12 (5), 587-610.
- Lagro Jr, J., 2005. Land-use classification.
- Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. *Accident Analysis & Prevention* 37 (4), 775-786.
- Lee, D., Oh, S., Heo, S., Hahn, M., Year. Drowsy driving detection based on the driver's head movement using infrared sensors. In: *Proceedings of the 2008 Second International Symposium on Universal Communication*, pp. 231-236.
- Lee, J.D., Caven, B., Haake, S., Brown, T.L., 2001. Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors* 43 (4), 631-640.

- Lee, S.E., Olsen, E.C., Wierwille, W.W., 2004. A comprehensive examination of naturalistic lane-changes. United States. National Highway Traffic Safety Administration.
- Lerner, N., Jenness, J., Singer, J., Klauer, S., Lee, S., Donath, M., Manser, M., Ward, N., 2010. An exploration of vehicle-based monitoring of novice teen drivers: Final report. Report number DOT HS 811, 333.
- Levitt, D., Pedestrian crashtypes.
- Liang, Y., Lee, J.D., Yekhshatyan, L., 2012. How dangerous is looking away from the road? Algorithms predict crash risk from glance patterns in naturalistic driving. *Human Factors* 54 (6), 1104-1116.
- Liu, G., Wang, L., Zou, S., Year. A radar-based blind spot detection and warning system for driver assistance. In: *Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 2204-2208.
- Liu, J., 2015. Driving volatility in instantaneous driving behaviors: Studies using large-scale trajectory data.
- Liu, J., Khattak, A., Wang, X., 2017b. A comparative study of driving performance in metropolitan regions using large-scale vehicle trajectory data: Implications for sustainable cities. *International Journal of Sustainable Transportation* 11 (3), 170-185.
- Liu, J., Khattak, A.J., 2017. Gate-violation behavior at highway-rail grade crossings and the consequences: Using geo-spatial modeling integrated with path analysis. *Accident Analysis & Prevention* 109, 99-112.
- Liu, J., Khattak, A.J., 2018. Are gates at rail grade crossings always safe? Examining motorist gate-violation behaviors using path analysis. *Transportation Research Part F: Traffic Psychology and Behaviour* 55, 314-324.
- Liu, J., Khattak, A.J., Richards, S.H., Nambisan, S., 2015. What are the differences in driver injury outcomes at highway-rail grade crossings? Untangling the role of pre-crash behaviors. *Accident Analysis & Prevention* 85, 157-169.
- Louw, T., Kuo, J., Romano, R., Radhakrishnan, V., Lenné, M. G., Merat, N., 2019. Engaging in NDRTs affects drivers' responses and glance patterns after silent automation failures. *Transportation Research Part F: Traffic Psychology and Behaviour*, 62, 870-882.
- Ma, X., Chen, F., Chen, S., 2015. Modeling crash rates for a mountainous highway by using refined-scale panel data. *Transportation Research Record* 2515 (1), 10-16.
- Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research* 1, 1-22.
- Maret, I., Dakan, B., 2003. Sig et gestion de l'étalement urbain (gis and sprawl management). *Bulletin de l'Association de Géographes Français* 80 (2), 220-234.

- McDonald, A. D., Alambeigi, H., Engström, J., Markkula, G., Vogelpohl, T., Dunne, J., Yuma, N., 2019. Toward computational simulations of behavior during automated driving takeovers: a review of the empirical and modeling literatures. *Human factors*, 61(4), 642-688.
- Mcdonald, M., Marsden, G., Brackstone, M., 2001. Deployment of interurban att test scenarios (diats): Implications for the European road network. *Transport Reviews* 21 (3), 303-335.
- McFadden, D., 1981. Econometric model of probabilistic choice. *Structural analysis of discrete data*, 198-272. Cambridge, MA: MIT Press.
- Mckeever, B.B., 1998. Estimating the potential safety benefits of intelligent transportation systems. Mitretek Systems.
- Menard, S., 2002. *Applied logistic regression analysis* SAGE.
- Menon, N., Barbour, N., Zhang, Y., Pinjari, A.R., Mannering, F., 2019. Shared autonomous vehicles and their potential impacts on household vehicle ownership: An exploratory empirical assessment. *International Journal of Sustainable Transportation* 13 (2), 111-122.
- Miles-Doan, R., 1996. Alcohol use among pedestrians and the odds of surviving an injury: Evidence from florida law enforcement data. *Accident Analysis & Prevention* 28 (1), 23-31.
- Mohanty, M., Mishra, A., Routray, A., 2009. A non-rigid motion estimation algorithm for yawn detection in human drivers. *International Journal of Computational Vision and Robotics* 1 (1), 89-109.
- Murtagh, E.M., Nichols, L., Mohammed, M.A., Holder, R., Nevill, A.M., Murphy, M.H., 2014. Walking to improve cardiovascular health: A meta-analysis of randomised control trials. *The Lancet* 384, S54.
- Nasar, J.L., Troyer, D., 2013. Pedestrian injuries due to mobile phone use in public places. *Accident Analysis & Prevention* 57, 91-95.
- Ng, M., Sathasivan, K., 2014. Probabilistic modeling of erroneous human response to in-vehicle route guidance systems: A first look. *Journal of Intelligent Transportation Systems* 18 (2), 131-137.
- NHTSA, 2018. Fact sheet auto industry commitment to IIHS and NHTSA on automatic emergency braking.
- NHTSA, 2019. Traffic safety facts research note. U.S. Department of Transportation.
- Nickelsburg, G., 1985. Small-sample properties of dimensionality statistics for fitting var models to aggregate economic data: A monte carlo study. *Journal of Econometrics* 28 (2), 183-192.
- Norman, D.A., Shallice, T., 1980. Attention to action: Willed and automatic control of behavior technical report no. 8006.
- NTSB., 2017. Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016.
- O'donnell, C., Connor, D., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention* 28 (6), 739-753.

- O'higgins, N., 2009. 'It's not that i'm a racist', it's that they are Roma: Roma discrimination and returns to education in South Eastern Europe.
- WHO, 2013. Strengthening road safety legislation: A practice and resource manual for countries World Health Organization.
- Öström, M., Eriksson, A., 2001. Pedestrian fatalities and alcohol. *Accident Analysis & Prevention* 33 (2), 173-180.
- Oviedo-Trespalacios, O., Haque, M.M., King, M., Washington, S., 2017. Effects of road infrastructure and traffic complexity in speed adaptation behaviour of distracted drivers. *Accident Analysis & Prevention* 101, 67-77.
- Papazikou, E., 2019. Investigating the transition from normal driving to safety-critical scenarios. Loughborough University.
- Papazikou, E., Quddus, M., Thomas, P., 2017. Detecting deviation from normal driving using SHRP2 NDS data.
- Papazikou, E., Quddus, M., Thomas, P., Kidd, D., 2019. What came before the crash? An investigation through SHRP2 NDS data. *Safety Science* 119, 150-161.
- Peng, H., Ma, X., Chen, F., 2020. Examining injury severity of pedestrians in vehicle–pedestrian crashes at mid-blocks using path analysis. *International Journal of Environmental Research and Public Health* 17 (17), 6170.
- Peterson, L.E., 2009. K-nearest neighbor. *Scholarpedia* 4 (2), 1883.
- Petridou, E., Moustaki, M., 2000. Human factors in the causation of road traffic crashes. *European Journal of Epidemiology* 16 (9), 819-826.
- Poudel-Tandukar, K., Nakahara, S., Ichikawa, M., Poudel, K.C., Jimba, M., 2007. Risk perception, road behavior, and pedestrian injury among adolescent students in Kathmandu, Nepal. *Injury Prevention* 13 (4), 258-263.
- Pour-Rouholamin, M., Zhou, H., 2016. Investigating the risk factors associated with pedestrian injury severity in Illinois. *Journal of Safety Research* 57, 9-17.
- Prato, C.G., Kaplan, S., Patrier, A., Rasmussen, T.K., 2018. Considering built environment and spatial correlation in modeling pedestrian injury severity. *Traffic injury prevention* 19 (1), 88-93.
- Qin, L., Li, Z.R., Chen, Z., Bill, M.A., Noyce, D.A., 2019. Understanding driver distractions in fatal crashes: An exploratory empirical analysis. *Journal of Safety Research* 69, 23-31.
- Qin, Z., Wang, A.T., Zhang, C., Zhang, S., Year. Cost-sensitive classification with k-nearest neighbors. In: *Proceedings of the International Conference on Knowledge Science, Engineering and Management*, pp. 112-131.

- Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. *Journal of Safety Research* 33 (4), 445-462.
- Rakha, H., 2011. Feasibility of using in-vehicle video data to explore how to modify driver behavior that causes nonrecurring congestion Transportation Research Board.
- Reason, J., 1990. *Human error* Cambridge university press.
- Reason, J., 2000. Human error: Models and management. *BMJ* 320 (7237), 768-770.
- Reason, J., Manstead, A., Stradling, S., Baxter, J., Campbell, K., 1990. Errors and violations on the roads: A real distinction? *Ergonomics* 33 (10-11), 1315-1332.
- Reason, J.T., Mycielska, K., 1982. *Absent-minded?: The psychology of mental lapses and everyday errors* Prentice Hall.
- Recarte, M.A., Nunes, L.M., 2003. Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied* 9 (2), 119.
- Regan, M.A., Mitsopoulos, E., Haworth, N., Young, K., 2002. Acceptability of in-vehicle intelligent transport systems to victorian car drivers. Monash University Accident Research Centre.
- Richard, C.M., Lee, J., Atkins, R., Brown, J.L., 2020. Using shrp2 naturalistic driving data to examine driver speeding behavior. *Journal of safety research* 73, 271-281.
- Rifaat, S.M., Chin, H.C., 2007. Accident severity analysis using ordered probit model. *Journal of Advanced Transportation* 41 (1), 91-114.
- Rimini-Doering, M., Altmueller, T., Ladstaetter, U., Rossmeier, M., 2005. Effects of lane departure warning on drowsy drivers' performance and state in a simulator.
- Risser, R., Sucha, M., 2020. Start walking! How to boost sustainable mode choice—psychological measures to support a shift from individual car use to more sustainable traffic modes. *Sustainability* 12 (2), 554.
- Roe, B.P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., McGregor, G., 2005. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 543 (2-3), 577-584.
- Roodman, D., Year. Mixed-process models with CMP. In: *Proceedings of the DC09 Stata Conference*.
- Sabey, B.E., Staughton, G., 1975. Interacting roles of road environment vehicle and road user in accidents. *Ceste i mostovi*.
- Salmon, P.M., Lenné, M.G., Stanton, N.A., Jenkins, D.P., Walker, G.H., 2010. Managing error on the open road: The contribution of human error models and methods. *Safety Science* 48 (10), 1225-1235.

- Schorr, J., Hamdar, S.H., Silverstein, C., 2017. Measuring the safety impact of road infrastructure systems on driver behavior: Vehicle instrumentation and real world driving experiment. *Journal of Intelligent Transportation Systems* 21 (5), 364-374.
- Schwarz, C., Gaspar, J., Miller, T., Yousefian, R., 2019. The detection of drowsiness using a driver monitoring system. *Traffic Injury Prevention* 20 (sup1), S157-S161.
- Schwarz, G., 1978. Estimating the dimension of a model. *The Annals of Statistics*, 461-464.
- Semanjski, I., Gautama, S., 2015. Smart city mobility application—gradient boosting trees for mobility prediction and analysis based on crowdsourced data. *Sensors* 15 (7), 15974-15987.
- Shafique, M.A., Hato, E., Year. A comparison among various classification algorithms for travel mode detection using sensors' data collected by smartphones. In: *Proceedings of the International Conference on Computers in Urban Planning and Urban Management*.
- Shaon, M.R.R., Qin, X., Chen, Z., Zhang, J., 2018. Exploration of contributing factors related to driver errors on highway segments. *Transportation Research Record* 2672 (38), 22-34.
- Sigelman, L., Zeng, L., 1999. Analyzing censored and sample-selected data with Tobit and heckit models. *Political Analysis* 8 (2), 167-182.
- Simons-Morton, B.G., Guo, F., Klauer, S.G., Ehsani, J.P., Pradhan, A.K., 2014. Keep your eyes on the road: Young driver crash risk increases according to duration of distraction. *Journal of Adolescent Health* 54 (5), S61-S67.
- Şimşekoğlu, Ö., 2015. How do attitudes, personality traits, and driver behaviors relate to pedestrian behaviors?: A Turkish case. *Traffic Injury Prevention* 16 (1), 84-89.
- Singh, S., 2015. Critical reasons for crashes investigated in the national motor vehicle crash causation survey.
- Smith, P., Shah, M., Da Vitoria Lobo, N., Year. Monitoring head/eye motion for driver alertness with one camera. In: *Proceedings of the Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, pp. 636-642.
- Sorensen, A., 2000. Land readjustment and metropolitan growth: An examination of suburban land development and urban sprawl in the Tokyo metropolitan area. *Progress in Planning* 53 (4), 217-330.
- Sossa, H., Zamora, E., Year. Vision-based blind spot warning system by deep neural networks. In: *Proceedings of the Mexican Conference on Pattern Recognition*, pp. 185-194.
- Soucy, P., Mineau, G.W., Year. A simple KNN algorithm for text categorization. In: *Proceedings of the Proceedings 2001 IEEE International Conference on Data Mining*, pp. 647-648.
- Stanton, N.A., Salmon, P.M., 2009. Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science* 47 (2), 227-237.

- Stavrinos, D., Byington, K.W., Schwebel, D.C., 2009. Effect of cell phone distraction on pediatric pedestrian injury risk. *Pediatrics* 123 (2), e179-e185.
- Stavrinos, D., Byington, K.W., Schwebel, D.C., 2011. Distracted walking: Cell phones increase injury risk for college pedestrians. *Journal of Safety Research* 42 (2), 101-107.
- Tan, H., Zhao, F., Hao, H., Liu, Z., Amer, A.A., Babiker, H., 2020. Automatic emergency braking (AEB) system impact on fatality and injury reduction in china. *International Journal of Environmental Research and Public Health* 17 (3), 917.
- Tefft, B.C., 2013. Impact speed and a pedestrian's risk of severe injury or death. *Accident Analysis & Prevention* 50, 871-878.
- Tefft, B.C., 2014. Prevalence of motor vehicle crashes involving drowsy drivers, United States, 2009-2013. CiteSeer.
- Thomas, P., Morris, A., Talbot, R., Fagerlind, H., 2013. Identifying the causes of road crashes in Europe. *Annals of Advances in Automotive Medicine* 57, 13.
- Tian, R., Li, L., Chen, M., Chen, Y., Witt, G.J., 2013. Studying the effects of driver distraction and traffic density on the probability of crash and near-crash events in naturalistic driving environment. *IEEE Transactions on Intelligent Transportation Systems* 14 (3), 1547-1555.
- Tideman, M., Van Der Voort, M.C., Van Arem, B., Tillema, F., Year. A review of lateral driver support systems. In: *Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference*, pp. 992-999.
- Treat, J.R., Tumbas, N., McDonald, S., Shinar, D., Hume, R.D., Mayer, R., Stansifer, R., Castellan, N., 1979b. Tri-level study of the causes of traffic accidents: Final report. Executive summary. Indiana University, Bloomington, Institute for Research in Public Safety.
- FHWA, U.S. DOT. 2009. Advance dilemma-zone detection system.
- Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. *Accident Analysis & Prevention* 59, 309-318.
- Wali, B., Ahmed, A., Ahmad, N., 2018. An ordered-probit analysis of enforcement of road speed limits. In: *Proceedings of the Institution of Civil Engineers-Transport*, pp. 225-234.
- Wali, B., Greene, D.L., Khattak, A.J., Liu, J., 2018b. Analyzing within garage fuel economy gaps to support vehicle purchasing decisions—a copula-based modeling & forecasting approach. *Transportation Research Part D: Transport and Environment* 63, 186-208.
- Wali, B., Khattak, A.J., 2020. Harnessing ambient sensing & naturalistic driving systems to understand links between driving volatility and crash propensity in school zones—a generalized hierarchical mixed logit framework. *Transportation Research Part C: Emerging Technologies* 114, 405-424.

- Wali, B., Khattak, A.J., Bozdogan, H., Kamrani, M., 2018c. How is driving volatility related to intersection safety? A bayesian heterogeneity-based analysis of instrumented vehicles data. *Transportation research Part C: Emerging Technologies* 92, 504-524.
- Wali, B., Khattak, A.J., Karnowski, T., 2019. Exploring microscopic driving volatility in naturalistic driving environment prior to involvement in safety critical events—concept of event-based driving volatility. *Accident Analysis & Prevention* 132, 105277.
- Wali, B., Khattak, A.J., Karnowski, T., 2020. The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. *Analytic Methods in Accident Research*, 100136.
- Wali, B., Khattak, A.J., Khattak, A.J., 2018d. A heterogeneity based case-control analysis of motorcyclist's injury crashes: Evidence from motorcycle crash causation study. *Accident Analysis & Prevention* 119, 202-214.
- Wali, B., Khattak, A.J., Xu, J., 2018e. Contributory fault and level of personal injury to drivers involved in head-on collisions: Application of copula-based bivariate ordinal models. *Accident Analysis & Prevention* 110, 101-114.
- Wang, L., Zhong, H., Ma, W., Abdel-Aty, M., Park, J., 2020. How many crashes can connected vehicle and automated vehicle technologies prevent: A meta-analysis. *Accident Analysis & Prevention* 136, 105299.
- Wang, X., Khattak, A.J., Liu, J., Masghati-Amoli, G., Son, S., 2015. What is the level of volatility in instantaneous driving decisions? *Transportation Research Part C: Emerging Technologies* 58, 413-427.
- Waseem, M., Ahmed, A., Saeed, T.U., 2019. Factors affecting motorcyclists' injury severities: An empirical assessment using random parameters logit model with heterogeneity in means and variances. *Accident Analysis & Prevention* 123, 12-19.
- Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. *Statistical and econometric methods for transportation data analysis* CRC press.
- Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. *Statistical and econometric methods for transportation data analysis* Chapman and Hall/CRC.
- Watanabe, Y., Ito, S., 2007. Influence of vehicle properties and human attributes on neck injuries in rear-end collisions. In: *Proceedings of the Proceedings of the 20th International Technical Conference on the Enhanced Safety of Vehicles (CD-ROM)*. Washington, DC: National Highway Traffic Safety Administration.
- Wierwille, W., Hanowski, R., Hankey, J., Kieliszewski, C., Lee, S., Medina, A., Keisler, A., Dingus, T., 2002. Identification and evaluation of driver errors: Overview and recommendations.

- Williams, R., 2005. Gologit2: A program for generalized logistic regression/partial proportional odds models for ordinal variables. Retrieved May 12, 2005.
- Williams, R., 2016. Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology* 40 (1), 7-20.
- Wooldridge, J.M., 2010. *Econometric analysis of cross section and panel data* MIT press.
- Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accident Analysis & Prevention* 40 (4), 1320-1329.
- Yang, W., Zhang, X., Lei, Q., Cheng, X., 2019. Research on longitudinal active collision avoidance of autonomous emergency braking pedestrian system (AEB-P). *Sensors* 19 (21), 4671.
- Yannis, G., Papadimitriou, E., Karekla, X., Kontodima, E., 2010. Mobile phone use by young drivers: Effects on traffic speed and headways. *Transportation Planning and Technology* 33 (4), 385-394.
- Ye, F., Lord, D.J.T.R.R., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: Multinomial logit, ordered probit, and mixed logit. *2241* (1), 51-58.
- Yeo, J., Lee, J., Jang, K., 2020. The effects of rainfall on driving behaviors based on driving volatility. *International Journal of Sustainable Transportation*, 1-9.
- Young, M.S., Mahfoud, J.M., Walker, G.H., Jenkins, D.P., Stanton, N.A., 2008. Crash dieting: The effects of eating and drinking on driving performance. *Accident Analysis & Prevention* 40 (1), 142-148.
- Zhang, M., Khattak, A.J., Liu, J., Clarke, D., 2018. A comparative study of rail-pedestrian trespassing crash injury severity between highway-rail grade crossings and non-crossings. *Accident Analysis & Prevention* 117, 427-438.

VITA

Numan Ahmad was born in Peshawar, Pakistan. After passing his higher secondary school certificate from Islamia College Peshawar (Pakistan), he started his Bachelor of Science in Civil Engineering at the University of Engineering and Technology Peshawar (Pakistan). After completing his BSc in Civil Engineering in 2013, he started his master's in Transportation Engineering at the National University of Sciences and Technology Pakistan. His master thesis relates to exploring factors associated with crash severity on motorways in Pakistan. After completing his MS in Transportation Engineering in 2016, he joined the provincial Transport and Mass Transit Department at Peshawar (Pakistan) where he worked for one year as a Highway Engineer. In 2018, Mr. Ahmad went to the United States to pursue his in Civil and Environmental Engineering Transportation and Master in Statistics at The University of Tennessee, Knoxville. His research covers a wide range of transportation-related topics which include transportation safety, Intelligent Transportation Systems, connected and automated vehicles, electric and alternative fuel vehicles, and safety of vulnerable road users (pedestrians, bicyclists, and motorcyclists). His research interests are mainly application of advanced statistical and economic models, artificial intelligence or machine learning methods, and more sophisticated heterogeneous ensembles (combination of both statistical and machine learning methods) to investigate research issues related to transportation safety, connected and automated vehicles, electric and alternative fuel vehicles, and safety of vulnerable road users. Mr. Ahmad has authored/co-authored 10 journal articles and 17 peer-reviewed conference papers to date.

As a Ph.D. student, Mr. Ahmad received several scholarships and awards including the Lifesavers Scholarship, Graduate Student Senate Travel Award, Collaborative Sciences Center for Road Safety (CSCRS) award for being an author of one of the top three papers presented at the CSCRS summit in North Carolina. He has been an active ITE student member and has served as a secretary for the ITE student chapter at UTK in 2018-2019. Mr. Ahmad has also been serving as a reviewer for various scientific journals including the Journal of Transportation Research Record, Accident Analysis and Prevention, and Safety Science.