Behavior-based Predictive Safety Analytics – Pilot Study

APRIL 2019 | Final Report

SAFE-D
SAFETY THROUGH DISRUPTION

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

VT | TRANSPORTATION INSTITUTE
VIRGINIA TECH

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 02-020	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Behavior-based Predictive Safety Analytics – Pilot Study		5. Report Date April 2019
		6. Performing Organization Code:
7. Author(s) Johan Engström Andrew Miller Wenyan Huang Susan Soccolich Sahar Ghanipoor Machiani Arash Jahangiri Felix Dreger Joost de Winter		8. Performing Organization Report No. Report 02-020
9. Performing Organization Name and Address: Safe-D National UTC Virginia Tech Transportation Institute 3500 Transportation Research Plaza Blacksburg, VA 24060 San Diego State University		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/Project 02-020
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT)		13. Type of Report and Period Final Research Report 14. Sponsoring Agency Code
15. Supplementary Notes This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program.		
16. Abstract This report gives an overview of the main findings from the Behavior-based Predictive Safety Analytics – Pilot Study project. The main objective of the project was to investigate the possibilities of developing statistical models predicting individual driver crash involvement based on individual driving style, demographic and behavioral history variables, using large sets of naturalistic driving data. The project was designed as a pilot project with the objective of providing the basis for a future more comprehensive research effort. Based on Second Strategic Highway Research Program (SHRP2) data, a subset of behavior and crash data including 2,458 drivers was created for analysis. The data were analyzed to investigate to what extent these drivers were differentially involved in crashes and near crashes, to what extent this was associated with individual characteristics, and if it is possible to predict individual drivers' crash and near crash involvement based on variables representing individual characteristics. The results clearly demonstrated the presence of differential crash and near crash involvement and showed significant associations between enduring personal factors and crash involvement. Moreover, logistic regression and random forest classifiers were relatively successful in predicting crash and near crash involvement based on individual characteristics, but the ability to specifically predict involvement in crashes was more limited.		
17. Key Words Crashes, driver behavior, driver characteristics, data mining, driving style, risk, naturalistic data	18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.	

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 94	22. Price $0

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Abstract

This report gives an overview of the main findings from the Behavior-based Predictive Safety Analytics – Pilot Study project. The main objective of the project was to investigate the possibilities of developing statistical models predicting individual driver crash involvement based on individual driving style, demographic and behavioral history variables, using large sets of naturalistic driving data. The project was designed as a pilot project with the objective of providing the basis for a future more comprehensive research effort. Based on Second Strategic Highway Research Program (SHRP2) data, a subset of behavior and crash data including 2,458 drivers was created for analysis. The data were analyzed to investigate to what extent these drivers were differentially involved in crashes and near crashes, to what extent this was associated with individual characteristics, and if it is possible to predict individual drivers' crash and near crash involvement based on variables representing individual characteristics. The results clearly demonstrated the presence of differential crash and near crash involvement and showed significant associations between enduring personal factors and crash involvement. Moreover, logistic regression and random forest classifiers were relatively successful in predicting crash and near crash involvement based on individual characteristics, but the ability to specifically predict involvement in crashes was more limited.

Acknowledgements

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program.

The authors would like to thank David Forney of SmartDrive for his time, guidance, and expertise, and Jeff Hickman of VTTI for his thorough review and support of this report.

Table of Contents

INTRODUCTION 1

BACKGROUND 2

CONCEPTUAL FRAMEWORK, DATA AND MODELING APPROACHES 4

Conceptual Framework 4

Data 6

Dataset Construction 6

Modeling 7

Independent and Dependent Measures 7

Dependent Measures – Crash and Near Crash Involvement 9

Statistical Models 9

RESULTS 10

Analysis of Differential Crash Involvement 10

Association between Self-Reported Driver Behavior/Personality and Crash Involvement 13

Classification of Crash Involved Drivers Based on Enduring Personal Factors 13

DISCUSSION 15

CONCLUSIONS AND RECOMMENDATIONS 18

ADDITIONAL PRODUCTS 18

Education and Workforce Development Products 18

Curriculum for a Course Module on Differential Crash Involvement and Behavior-based Predictive Analytics ..18

Technology Transfer Products 19

Data Products 19

REFERENCES 20

APPENDIX A. STATE OF THE ART REVIEW 22

APPENDIX B. PUBLICATION MANUSCRIPTS 57

APPENDIX C. COURSE CURRICULUM..... 89

Curriculum for a Course Module on Differential Crash Involvement and Behavior-based Predictive Analytics ..89

APPENDIX D. DATA DICTIONARY OF VARIABLES 91

List of Figures

Figure 1. Conceptual framework for differential crash involvement. 6

Figure 2. Overview of the study design with independent variables (blue) and dependent variables (red). 6

Figure 3. Kinematic events, consisting of hard starts and hard stops at different levels (purple and red lines, respectively) based on time series data, acceleration X. 8

Figure 4: AIC values of logistic regression models between CNC and each driving style measure at 46 g-force thresholds. 8

List of Tables

Table 1. Summary Table of Proportions of Drivers Accounting for a Major (70%, 80%, 90%, 95%) Proportion of Risk 11

Table 2. Comparisons between the High and Low Risk Groups in Phase I 11

Table 3. The Prediction Performance of Logistic Regression Models 14

Table 4. The Confusion Matrices of Logistic Regression Models 14

Table 5. Prediction Performance of Random Forest Models 14

Table 6. Confusion Matrices of random Forest Models 15

Introduction

The general aim of this project was to investigate the possibilities of developing statistical models to predict individual driver crash involvement based on driving style, demographic, and behavioral history data. Such models have a range of applications, in particular in the areas of fleet safety management and insurance. Although the relationship between individual driver characteristics and road safety has a long research history, the advent of large sets of naturalistic driving data, which include a significant number of crashes as well as driver behavior and demographics data, allows for exciting new research possibilities.

Today, there are two main sources of naturalistic driving data that contain a sufficient amount of crash and behavioral data to be used for the present purposes: (1) the Second Strategic Highway Research Program (SHRP 2) dataset and (2) commercial fleet- and behavior change management programs offered by in-vehicle monitoring systems (e.g., Lytx and SmartDrive systems), which collect tens of thousands of crashes annually. SHRP 2 was a large-scale project, carried out by the Virginia Tech Transportation Institute (VTTI) and funded by the Transportation Research Board, that collected over 30 million passenger vehicle miles of continuous video and vehicle sensor data, and recorded over 2,000 crashes and 7,000 near-crashes (Dingus et al., 2015; Hankey et al., 2016). By contrast, the data collected by Lytx and SmartDrive (Lytx, 2017, SmartDrive, 2017) is event-triggered, consisting of 10–20 s epochs of video and sensor data. Of key importance for this project, SmartDrive also collects exposure data (driving time and mileage) for each driver. The total mileage amounts to billions of miles, resulting in the capture of tens of thousands of crashes and even more near-crashes and non-conflict events. For each event, a range of driver behaviors and other safety relevant observations are manually coded by trained data reductionists.

The main objective of this project was to conduct a pilot investigation into how such large sets of naturalistic crash and behavior data can be used to establish predictive models of crash involvement. The work included a thorough state-of-the-art review of previous work (See Appendix A) in this area, including the development of a conceptual framework for relating behavior to crash causation and risk. Based on this, different proof-of-concept analyses and modeling efforts were conducted. This work also involved collaboration with Mr. Felix Dreger and Dr. Joost de Winter from Delft University in The Netherlands. Mr. Dreger visited VTTI for 1 month during the course of the project. The work subsequently resulted in three publications (two accepted and one to-be-submitted), co-authored by the entire research team. These are described in more detail in subsequent sections of this report.

A further objective of the project was to investigate to what extent the commercial data collected by Lytx and SmartDrive could be made available for academic research. To this end, members of the research team visited both companies in San Diego. Both companies were generally interested in collaborating around this topic and, and as long as the data can be completely

anonymized, there are no fundamental barriers for making this data available to academic researchers. This initial contact eventually resulted in a collaborative modeling effort with SmartDrive, which was conducted as part of the project. This was not included in the original work plan, but was conducted through an extension of the present project. This work was conducted separately from the main part of the project as it required a non-disclosure agreement (NDA) between SmartDrive, VTTI, and San Diego State University. Due to the NDA, and the fact that this analysis was not completed at the time of this writing, the results are not reported here. The intention is to publish these results in an academic journal pending SmartDrive's approval of the disclosed content.

A final goal was to develop a curriculum for undergraduate and graduate studies on behavior-based predictive safety analytics with a module in a graduate-level course. However, due to resource constraints, and since it was not clear what graduate course might be the target for this module, the course material was not developed in this pilot project. Thus, the main educational component is a general curriculum for such a module, which is included at the conclusion of this report.

The pilot project was designed with the objective of providing the basis for more comprehensive research efforts in the future. The analyses conducted in this project barely scratched the surface of what can be done in this area and, as further described below, a range of open issues were identified that could be addressed in a future project. This report provides an overview of the main results obtained relative to the main research questions stated in the project work plan.

Background

Drivers are the contributing factor in the majority of road crashes and understanding the relationship between individual driver characteristics and crash involvement has been a long-standing goal in road safety research (e.g., Elander et al., 1993; Guo et al., 2010; McKenna, 1983). It is well known that a small proportion of drivers often account for a major proportion of crashes (Sagberg et al., 2015), a phenomenon often referred to as the Pareto principle or the 80–20 rule, that has also been observed in many other domains. Thus, it is of great value to be able to identify these risky drivers before crashes happen. For example, is a driver who is regularly speeding and/or tailgating, and/or has a history of traffic violations, more likely to crash than a driver adopting a less aggressive driving style who has received no tickets in the past? If so, can such risky drivers be reliably identified based on individual driver characteristics, such as observed driving style, demographics, personality screening, or behavioral history.

Such behavior-based predictive safety analytics (BPSAs) have a wide range of applications. In particular, in the commercial fleet safety management and auto insurance domains, a range of commercial applications are already used to capture risky driving styles (e.g., in terms of the number of hard braking event per vehicle mile traveled). Moreover, in order to counter driver

shortage in the US trucking industry, current investigations are being made into the possibility of recruiting younger drivers provided they pass a screening of crash-predictive personal characteristics (Boris & Luciana, 2017).

However, the success of BPSAs ultimately hinges on the establishment of models able to reliably relate individual driver characteristics to actual crash risk; this relationship is currently poorly understood. Traditionally, the main reason for this has been the lack of data containing a sufficient number of detailed crash recordings and recorded driving behavior, demographics and screening data collected over an extensive time period before the crash. In recent years, this picture has started to change due to the advent of naturalistic driving studies. However, existing naturalistic driving analyses have typically focused on the relationship between the engagement in potentially distracting secondary tasks (or other driver behaviors/states) and crash risk, with the primary goal to identify risky tasks/behaviors/states (e.g., Dingus et al., 2016). Such studies are valuable for informing human-machine interaction design and distraction policy. However, the present effort focuses on the relationship between individual driver characteristics and crash involvement, with the primary goal of identifying risky drivers. As mentioned above, this analysis has its key applications in the context of fleet safety and insurance.

To obtain a deeper understanding and develop reliable predictive models of the relationship between individual driver behavior and crashes requires big datasets that include a large number of drivers, driving exposure for each driver, records of behaviors in non-conflict situations and, critically, a large number of crashes for the same driver population. Today, such datasets do exist. The SHRP 2 study, which is the largest publicly funded naturalistic driving effort to date, involved over 30 million vehicle miles, 3,000 vehicles, over 3,500 drivers and collected about 2,000 crashes (Dingus et al., 2015; Hankey et al., 2016). Although this dataset may be useful to explore the possibilities of establishing behavior-risk mappings at the individual level, the number of crashes is still relatively limited when broken down into specific categories, or when crashes of lower severity (e.g., curb strikes) are removed. Ideally, in order to establish relationships between individual behavior and crash risk, even larger naturalistic crash datasets are needed. Today, such datasets are becoming available as more vehicles are equipped with video logging systems. In particular, commercial programs for improving driver behavior, such as those offered by the companies Lytx and Smartdrive, can generate large sets of naturalistic crash and behavior data. Both companies have equipped on the order of hundreds of thousands of vehicles with event video data recorders and collected tens of thousands of crashes per year along with millions of annotated behavioral events and billions of miles driven. This type of video-based naturalistic crash data will proliferate even further in the near future[1] and, when

[1] For example, providers of traditional fleet management services such as Omnitracs are now offering video-based recording of safety-critical events.

combined with appropriate analytics methods, will become a true game changer for road safety. However, these datasets are proprietary and subject to ethical, legal and business constraints.

The objective of the present pilot project was to address a set of general research questions in order to lay the foundation for a larger effort on predictive crash analytics using large naturalistic crash and behavior data sets. In particular:

1. **How can the relationship between individual driver behavior/driving style and crashes be conceptualized?** For example, how are the concepts of "behavior" and "driving style" best defined for purposes of behavior-based predictive safety analytics? How can we best think about the mechanisms whereby these behaviors may produce crashes?
2. **What are the minimum requirements on behavioral and crash data for enabling predictive crash analytics on a larger scale?** For example at what level of detail do behaviors need to be coded? Is a generic "event-based" coding sufficient, or is a more detailed time-series coding required?
3. **What statistical methods are most appropriate for modeling the relationship between individual driver behaviors/driving styles and crashes based on naturalistic data?** Are existing statistical modeling techniques sufficient or do they need to be further developed to utilize naturalistic data properly?
4. **How can commercial naturalistic data be made available for analysis while respecting legal, ethical and business constraints?** For example, who owns the data? What degree of anonymization is needed to protect the data? What would be the motivation of commercial entities to share this data?

This final report provides an overview of the main results from the project. Details can be found in the specific publications found in Appendix B (de Winter et al., 2018; Huang et al., in review).

Conceptual Framework, Data and Modeling Approaches

Conceptual Framework

To guide the development of predictive models for individual crash involvement, a conceptual framework was developed defining a set of key terms and concepts useful for the present purposes. The framework is presented in more detail in the state-of-the-art review developed in the project (Engström et al., 2017; Appendix A).

Early work addressing the role of individual driver factors in the causation of (mainly non-traffic, work-related) crashes introduced the concept of "accident proneness" to account for the common observation that road crashes, or incidents in other domains, are typically not distributed among individuals in a way that could be explained by chance alone. The adequacy of this concept has been debated over the years (see review by McKenna, 1983) and today, at least

in the context of traffic safety, it has largely been abandoned in favor of the "differential crash involvement" concept. The basic idea underlying differential crash involvement is that some drivers have certain personal characteristics that make them more likely to become involved in crashes. For example, this may be related to a stronger propensity for risk taking among certain drivers, leading these drivers to look away from the road for longer periods and more frequently than the average driver, increasing the risk for an off-road glance to co-occur with an unexpected event (e.g., a lead vehicle braking), leading to a crash. However, drivers' behavior, off road glances in this example, is also determined by more temporary driver factors, such as a strong motivation to send a text message, or situational factors, such as driving in dense traffic requiring frequent mirror checks.

Knipling (2009) suggests a general distinction between personal and situational risk factors. The former refers to factors related to things "inside" the driver (explained further below) while situational factors refer to all other types of factors "outside" the driver, such as traffic, roadway, weather and the vehicle.

Personal factors may be further divided into temporary and enduring factors. The former relates to things that typically change from day-to-day or hour-to-hour, such as illness, sleepiness, and mood. In contrast, enduring factors, also referred to as constitutional personal driver factors (Knipling, 2009), refer to long term or permanent characteristics, such as gender, age, personality, driving experience, physical/sensory-motor abilities, skills, medical conditions and health, psychiatric and behavioral disorders, etc. This project was mainly concerned with enduring personal factors, since these are the fundamental factors underlying differential crash involvement. However, crashes often involve an interaction between enduring personal factors, temporary personal factors, and situational factors, and the relative contributions of these different types of factors to crash genesis are often difficult to disentangle (Elander et al., 1993).

Figure 1 provides a general illustration of these concepts. Current behavior, both driving and non-driving, is influenced by situational factors as well as temporary and enduring personal factors. This results in behavioral outcomes that may be successful (i.e., the situation played out as expected) or unsuccessful, leading to crashes, traffic conflicts (e.g., near crashes), violations, and convictions. These events constitute the driver's behavioral history and certain events may become recorded in crash databases, naturalistic driving data, inspection records (for commercial vehicles), legal records, etc. Importantly, all this takes place in a sociocultural context, such as a national or workplace culture (Sagberg et al., 2015). Enduring personal factors will be reflected in recurring observable behavioral patterns as well as behavioral history, although both are also influenced by temporary personal and situational factors. Some of these recurring behaviors, such as tailgating, speeding, or distraction, may be associated with increased crash risk. Hence, by developing statistical models that map from direct measurements of enduring personal factors (e.g., personality tests), observable driving style, and/or behavioral history to crash involvement, it may be possible to identify unsafe drivers before they become involved in crashes.

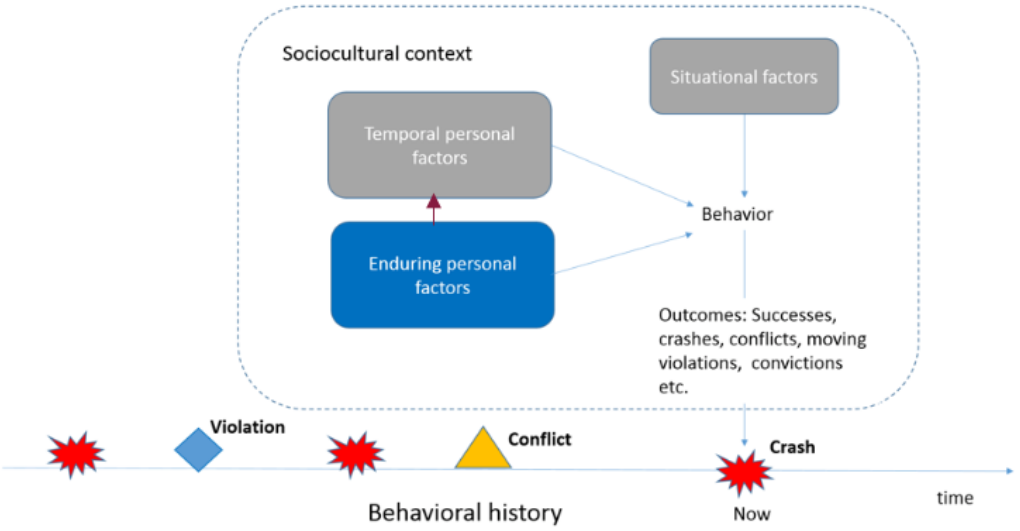

Figure 1. Conceptual framework for differential crash involvement.

Data

Dataset Construction

A subset of the SHRP 2 data was used to construct the datasets used for the analyses in the project. The datasets used for each analysis differed somewhat in the details (as outlined in the individual publications), but were generally constructed as described in the following.

For each individual driver, six consecutive calendar months were extracted beginning from the second month of data collection (study period, months 2–7) to account for any first-month effects, such as the observer effect (see Figure 2). This 6-month data interval was used to calculate driving style measures and crash/near crash involvement. In addition, questionnaire data for each participant, collected prior to the start of the SHRP 2 data collection, was retrieved.

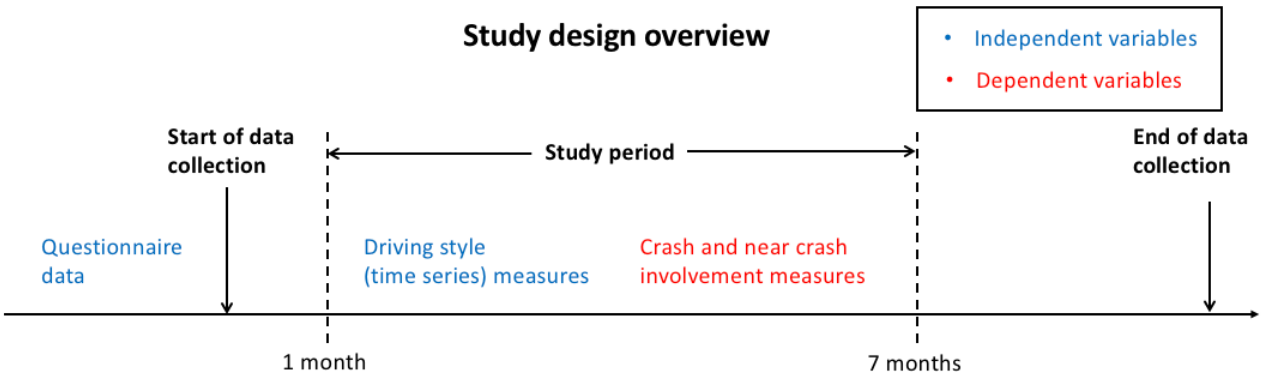

Figure 2. Overview of the study design with independent variables (blue) and dependent variables (red).

In addition, a range of further inclusion criteria were applied. In particular, drivers selected for the present analysis were required to have participated in SHRP 2 data collection for at least 7 months, and to have driven more than 1,000 miles in the 6-month study period (see Huang et al.,

in review for a more detailed description of these criteria). This resulted in a dataset of 2,458 drivers and 3.91 million trips, amounting to a total of 27.16 million miles of driving distance and 0.69 million driving hours.

Modeling

Independent and Dependent Measures

The following independent measures were included in the analysis.

Demographics: A Driver Demographic Questionnaire was used in SHRP 2 to investigate a variety of participant demographic information. The present analysis selected two variables from this questionnaire: age and gender. Age was stratified into three age groups: younger than 25 years, between 25 and 55 years, and older than 55 years. This was done for two reasons. First, SHRP 2 oversampled younger and older drivers, so the sample size was equivalent between the age groups. Second, prior research shows that drivers between 25 and 55 years of age have comparable crash risks (Ryan et al., 1998).

Driving history: A Driving History Questionnaire was used to obtain self-reported driving history information about participants, including driving experience, past violations and crashes, and training received. The present analysis included two variables from this questionnaire: self-reported violations and self-reported crashes in the past 3 years. These two independent variables were recoded into binary variables indicating whether the participant had at least one violation (or crash) in the last 3 years or not.

Driving style: Driving style here refers to persistent driving patterns characteristic for individual drivers (Sagberg et al., 2015). In the present study it was operationalized, based on Simons-Morton et al. (2013), in terms of the rates (numbers per mile) of six types of kinematic events calculated in the study period based on specific thresholds for each dependent variable. The six kinematic events were hard starts, stops, left turns, right turns, left yaw movement, and right yaw movement. Multiple events were counted as one if the interval between them was less than 1 second and events were removed if their event duration was less than half a second. The analysis built logistic regressions between each dependent variable and each driving style measure across different g-force thresholds (e.g., see Figure 3) and, for each driving style variable, selected the specific g-force level thresholds with the minimum Akaike Information Criterion (AIC) value, indicating the highest quality of statistic model (see Figure **4**). AIC is an estimator of the relative quality of statistical models or the relative information lost when a given model is used, and minimum AIC was used to ensure the goodness of fit by maximizing likelihood as well as to prevent overfitting by minimizing the number of parameters. These thresholds were identified across all drivers who had relevant data in the 6-month time period.

An example demonstrating the calculation of two of the driving style measures, hard starts and hard stops, is shown in Figure 3. This plot includes data from one trip (on May 21, 2011) and one driver. In this example, purple lines (+0.24g and +0.33g) result in different numbers of hard

starts (1 and 0) and red lines (-0.24g and -0.29g) result in different numbers of hard stops (4 and 2).

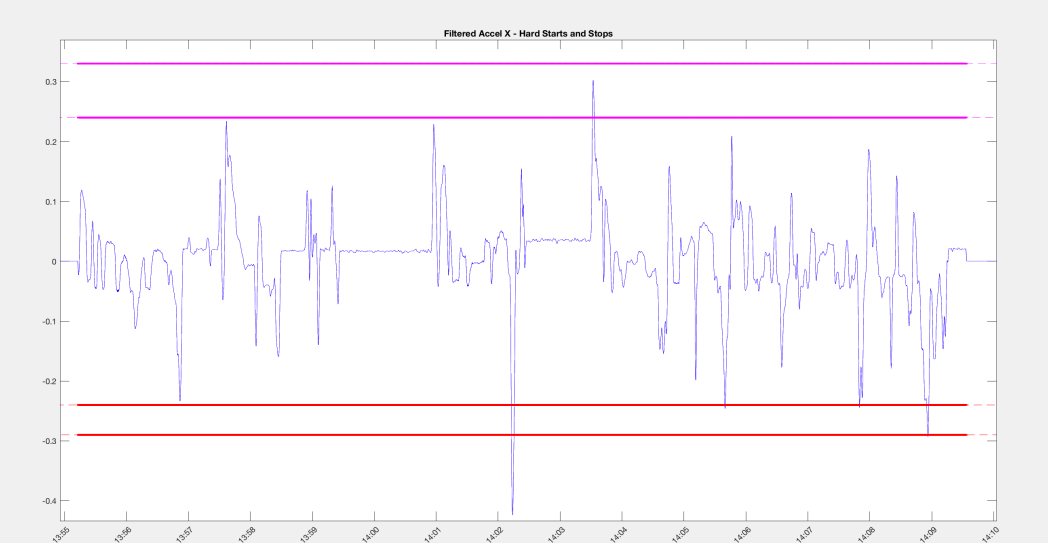

Figure 3. Kinematic events, consisting of hard starts and hard stops at different levels (purple and red lines, respectively) based on time series data, acceleration X.

The full set of thresholds identified for crash and near crash (CNC)-involved drivers that were used in this analysis is presented below in Figure 4.

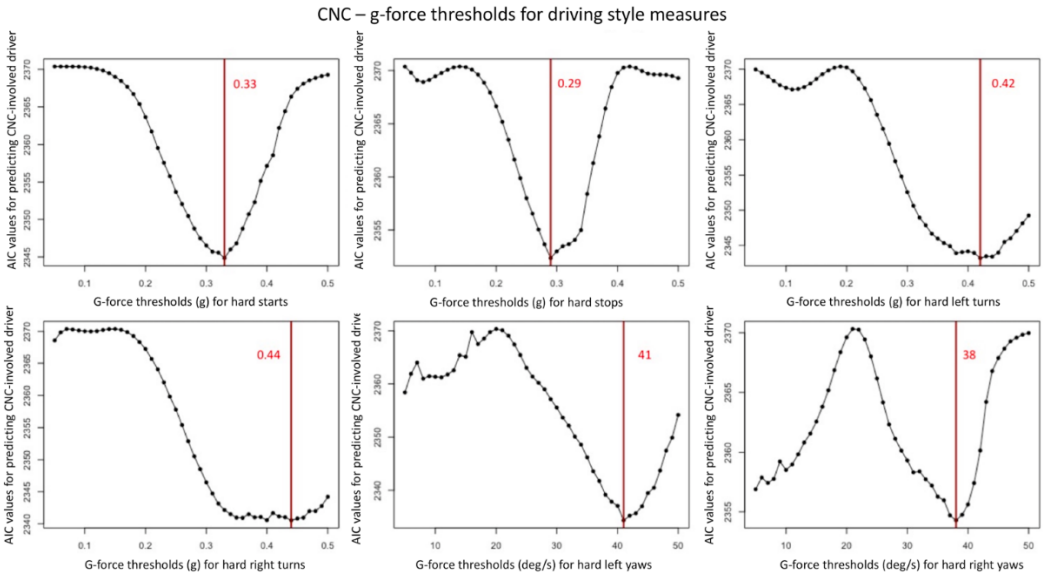

Figure 4: AIC values of logistic regression models between CNC and each driving style measure at 46 g-force thresholds.

Personality: Three self-report questionnaires were included from the SHRP 2 study: a modified version of the Manchester Driver Behavior Questionnaire (M-DBQ) containing 24 items answered on a 6-point Likert scale (Reason et al., 1990; Lajunen and Summala, 2003), the

Sensation Seeking Scale-form V (SSS-V) with 40 items (Zuckerman, 1994; Jonah, 1997), and a risk-perception questionnaire (Dingus et al., 2015). A principal component analysis (PCA) was used to group by the highest factor loadings and the components were interpreted as (1) slips, (2) violations, and (3) lapses. The mean score of each scale was calculated and used in further analyses.

The SSS-V is a self-report survey where respondents selected one of two choices that better described their feelings or likes (Zuckerman, 1994). The present analysis used the total score of the SSS-V to indicate the degree to which the participant engaged in sensation seeking behavior.

A risk-perception questionnaire was created for the SHRP 2 data collection (Dingus et al., 2015). The questions assessed the perceptual risk with driving behaviors on a seven-point Likert scale ranging from "No Greater Risk" to "Much Greater Risk." Most items provided little variance across drivers, so only one item was selected for inclusion. The item selected was "If you were to engage in changing lanes suddenly to get ahead in traffic, how do you think that would affect your risk of a crash?"

Dependent Measures – Crash and Near Crash Involvement

In the SHRP 2 dataset, safety critical events (SCEs) (i.e., crash, near crash, crash-relevant, non-conflict, subject conflict) were manually validated and coded by trained data reductionists. Only CNCs, as defined as in Hankey et al. (2016), were used in the present analysis. For this analysis, crashes of all severity levels were used. Two dependent variables represented CNC involvements of individual drivers: CNC is a binary variable indicating whether the participant was involved in zero or at least one CNC event in the study period; crash is a binary variable indicating whether the participant was involved in zero or at least one crash event in the study period. A driver was labeled as a CNC- or crash-involved driver if they had at least one CNC or crash in the study period.

Statistical Models

Two types of classification models were investigated with the goal of identifying CNC- (or crash-) involved drivers based on the independent measures described above: logistic regression and random forest (RF) classification.

Logistic regression is a statistical classification method that was used in this analysis to model the probability of a SHRP 2 participant being a CNC- or crash-involved driver. First, logistic regression was used to model the probability of a participant being a CNC- or crash-involved driver (for a similar use, see Guo and Fang, 2013). The dependent variable (CNC or crash) is a binary variable and is assumed to follow a Bernoulli distribution with a probability (p_i). This probability is associated with a set of covariates by a logit link function where the set of covariates are all potential independent variables:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i,1} + +\beta_2 x_{i,2} + \cdots + +\beta_k x_{i,k} = \boldsymbol{X_i \beta} \qquad (1)$$

where X_i is the matrix of predictors for individual i, and β is the vector of regression parameters. Both forward and backward variable selections were performed and the best model was selected based on the minimum AIC value. A driver is predicted as a CNC- or crash-involved driver if this probability is greater than a predefined threshold value (e.g., $p_0 = 0.5$). The Odds Ratio ($OR_j = \exp(\beta_j)$) is the change in probability of being a CNC-involved driver versus not being a CNC-involved driver associated with a variable j.

Second, RF classification, proposed by Breiman (2001), was used for classification and regression. This method creates a series of decision trees (i.e., forest), each of which is used to solve the classification problem individually. The final result is obtained based on the majority vote across all decision trees. The decision tree algorithm, introduced by Breiman in 1984, uses a recursive binary splitting approach to grow a tree by selecting a predictor and a cut point for that predictor to split the data into two parts. This procedure is iterated at several steps to create a dendrogram type of structure (i.e., a decision tree). At each splitting step, different criteria can be used to identify the best split (i.e., best classification).

The prediction performance of the models was evaluated in terms of the recall rate (or sensitivity), precision (positive predictive value), and accuracy. In the context of this study, the recall rate is the number of correctly predicted CNC- or crash-involved drivers divided by the total number of CNC- or crash-involved drivers. The precision is the number of correctly predicted CNC- or crash-involved drivers divided by the total number of drivers predicted by a model to be CNC- or crash-involved. Finally, the accuracy is the fraction of all drivers correctly classified as either CNC- (or crash-) involved or not involved.

Results

Three main analyses were conducted on the SHRP 2 data and reported in separate publications. First, a descriptive analysis was conducted to investigate to what extent SHRP 2 drivers were differentially involved in crashes and CNCs (Huang et al., 2018). Second, a correlational analysis of the association between self-reported driver behavior/personality scores and crash involvement was conducted (de Winter et al., 2018). The goal of the third analysis, which was the main analysis in this project, was to investigate to what extent it is possible to predict drivers' crash and/or CNC involvement based on the independent variables described above (Huang et al., in review).

Analysis of Differential Crash Involvement

In this analysis, the study period of 6 months (see Figure 2) was further divided into two phases: Phase I (months 2–4) and Phase II (months 5–7). The purposes of analyzing 6 months and dividing into two phases were three-fold: (1) it allowed for prediction of Phase II metrics based on Phase I variables; (2) the computational undertaking of mining continuous data was time-prohibitive; and (3) including a 6-month time period maximized the number of participants involved while maintaining a suitable amount of data for analyses. In each phase, the number of

CNC events per 1,000 miles driving distance was calculated and drivers were divided into high- and low risk groups depending on whether they accounted for 80% of the total CNC rate (total risk). The cut-off percentage was also varied to investigate the sensitivity to this criterion. Thus, drivers who accounted for 80% of the total CNC rate (total risk) in the respective phase were classified as high-risk drivers in that phase. The remaining drivers were classified as low-risk drivers. Table 1 summarizes the results of this classification. The table shows that the data strongly indicates the presence of differential CNC involvement. That is, a small proportion of drivers (25.4% in Phase 1) account for the great majority (80%) of CNCs, results that are very similar to the classic 80–20 rule.

Table 1. Summary Table of Proportions of Drivers Accounting for a Major (70%, 80%, 90%, 95%) Proportion of Risk

Proportions of low/high risk drivers in Phase I	Proportions of low risk drivers in Phase II	Proportions of high risk drivers in Phase II	Relative Risk
Low-risk drivers (74.6%)	**1107 (61.2%)**	**243 (13.4%)**	**2.23 (1.90, 2.61)**
High-risk drivers (25.4%)	**275 (15.2%)**	**184 (10.2%)**	

The main goal of the analysis was then to investigate to what extent this differential crash involvement could be explained in terms of enduring personal factors. To this end, Chi-squared and Wilcoxon rank sum statistical tests were used to examine the differences between the high and low-risk groups as classified during Phase I with respect to the independent variables described above. The results of these comparisons are shown in Table 2. As the table shows, most of the personal characteristics were statistically significantly associated with participants' classifications as high or low-risk drivers.

Table 2. Comparisons between the High and Low Risk Groups in Phase I

Variable	Type of data	Tests of group differences	Sample size	*p*-value	Effect size	Effect description	
Age	Categorical	Chi-squared test	1799	0.011	0.13	⇓	High-risk drivers have a higher proportion of drivers in the younger age groups and lower proportion of drivers in the middle and senior age groups.
Violations	Categorical	Chi-squared test	1807	0.018	0.06	⇑	High-risk drivers have a higher proportion of drivers with at least one self-reported violation in the past three years.
DBQ 1 - slips	Continuous	Wilcoxon rank sum test	1800	0.009	0.08	⇑	High-risk drivers have a significantly higher average DBQ1 score than low-risk drivers

Variable	Type of data	Tests of group differences	Sample size	*p*-value	Effect size	Effect description	
DBQ 2 - violations	Continuous	Wilcoxon rank sum test	1800	< 0.001	0.13	⇑	High-risk drivers have a significantly higher average DBQ2 score than low-risk drivers.
SSQ total score	Continuous	Wilcoxon rank sum test	1801	< 0.001	0.13	⇑	High-risk drivers have a significantly higher average sensation seeking score than low-risk drivers.
Risk-perception – lane change	Continuous	Wilcoxon rank sum test	1781	0.008	0.08	⇓	High-risk drivers have a lower average risk-perception score on sudden lane changes than low-risk drivers.
Number of hard starts per mile in Phase I	Continuous	Wilcoxon rank sum test	1809	< 0.001	0.26	⇑	High-risk drivers have a significantly higher average hard start rate than low-risk drivers.
Number of hard stops per mile in Phase I	Continuous	Wilcoxon rank sum test	1809	< 0.001	0.35	⇑	High-risk drivers have a significantly higher average hard stop rate than low-risk drivers.
Number of hard left turns per mile in Phase I	Continuous	Wilcoxon rank sum test	1809	< 0.001	0.25	⇑	High-risk drivers have a significantly higher average hard left turn rate than low-risk drivers.
Number of hard right turns per mile in Phase I	Continuous	Wilcoxon rank sum test	1809	< 0.001	0.22	⇑	High-risk drivers have a significantly higher average hard right turn rate than low-risk drivers.
Number of hard left yaws per mile in Phase I	Continuous	Wilcoxon rank sum test	1809	< 0.001	0.12	⇑	High-risk drivers have a significantly higher average hard left yaw rate than low-risk drivers.
Number of hard right yaws per mile in Phase I	Continuous	Wilcoxon rank sum test	1809	< 0.001	0.18	⇑	High-risk drivers have a significantly higher average hard right yaw rate than low-risk drivers.

Note: The phi coefficient was used to compute the effect size of categorical variables and Cliff's delta was used to compute the effect size of continuous variables. Gender, Number of Previous Crashes, and DBQ 3 – lapses were included in this analysis but were non-significant.

An analysis was also conducted on the consistency of the high- and low-risk classification (based on the 80% criterion) of drivers across the two phases. As shown in Table 1, 10.2% of the drivers were consistently classified as high-risk drivers in Phase I and II, and 28.6% (15.2% + 13.4%) of drivers' statuses were inconsistent. This analysis indicated the consistency of the high/low classification was moderate, at 71.4% across low- and high-risk drivers. However, only 40% of the high-risk drivers in Phase I were still classified as high-risk drivers in Phase II. Despite this, the relative risk of being classified as a high-risk driver in Phase II given high-risk classification in Phase I was significant at 2.23 [95% Confidence Interval (CI) = (1.90, 2.61)].

These results clearly demonstrate the phenomenon of differential CNC involvement, although it was only partially explained by the enduring personal factors included in the present analysis. CNC involvement was at least somewhat persistent over time for individual drivers. Taken together, these results show that enduring personal factors play a role in CNC involvement, although CNCs are also influenced to a large extent by other factors unrelated to enduring individual driver characteristics (or other enduring personal factors not included in the present analysis).

Association between Self-Reported Driver Behavior/Personality and Crash Involvement

A second analysis was specifically focused on the association between scores on the DBQ and the SSS, other involvement in crashes of different severity levels, and involvement in near crashes, driving style variables and demographics (age and gender). The results were published in de Winter et al. (2018) as commentary to an earlier paper by Martinussen et al. (2017), which analyzed correlations between the DBQ and recorded violations and crashes. The original study (Martinussen et al., 2017) found a moderate association between the DBQ violation score and actual traffic offences, but no significant association between DBQ scores and crashes. The present analysis (de Winter et al, 2018), which used the SHRP 2 dataset (but using the full data rather than the 6-month periods used in the other two analyses), generally confirmed these findings. DBQ-violations and SSS scores showed moderate correlations (around 0.2) with near-crashes, certain driving style measures (hard turns) and age, but low (0.02-0.1) correlations with crashes, depending on crash severity and at fault classification.

Classification of Crash Involved Drivers Based on Enduring Personal Factors

The third analysis aimed to develop statistical models for classifying drivers involved in CNCs based on variables representing enduring personal characteristics.

Two types of classification models were investigated with the goal of identifying CNC- (crash-) involved drivers based on the independent self-reported and driving style variables described above: logistic regression and RF classification. As described above, separate logistic regression was initially conducted for each of the driving style measures to determine optimal g-force threshold values based on minimum AIC.

The models' prediction performance was evaluated in terms of the recall rate (or sensitivity), precision (positive predictive value), and accuracy. In the context of this study, the recall rate is the number of correctly predicted CNC- or crash- involved drivers divided by the total number of CNC- (crash-) involved drivers. The precision is the number of correctly predicted CNC- (crash-) involved drivers divided by the total number of drivers predicted by a model to be CNC- (crash-) involved. Finally, the accuracy is the fraction of all drivers correctly classified as either CNC- (crash-) involved or not involved.

The full dataset of 2,458 drivers was randomly partitioned into two balanced groups: a training set (70%, 1,720 drivers) and a test set (30%, 738 drivers).

The prediction performance results for the logistic regression models (LR_Crash and LR_CNC) are shown Table 3 and Table 4. Table 3 shows that LR_CNC (CNC involvement) has high recall rates of about 70% and good accuracy rates of about 60% for the training and test sets. Table 3 also shows that LR_crash (crash involvement) has very low recall rates of about 0 for both sets, which may be the result of imbalanced datasets between drivers with or without crash events in the study period.

Table 3. The Prediction Performance of Logistic Regression Models

Models	Training Set			Test Set		
	Recall	Precision	Accuracy	Recall	Precision	Accuracy
LR_CNC	0.691	0.613	0.589	0.723	0.644	0.619
LR_Crash	0.010	0.429	0.817	0.000	0.000	0.802

Table 4. The Confusion Matrices of Logistic Regression Models

Models	Training Set				Test Set			
LR_CNC			Predicted				Predicted	
			0	1			0	1
	Actual	0	358 (20.8%)	414 (24.1%)	Actual	0	157 (21.3%)	166 (22.5%)
		1	293 (17.0%)	655 (38.1%)		1	115 (15.6%)	300 (40.6%)
LR_Crash			Predicted				Predicted	
			0	1			0	1
	Actual	0	1403 (81.6%)	4 (0.2%)	Actual	0	592 (80.2%)	1 (0.1%)
		1	310 (18.0%)	3 (0.2%)		1	145 (19.7%)	0

The corresponding results for the RF models (RF_Crash and RF_CNC) are shown in Table 5 and Table 6. Table 5 shows that RF_CNC has a high recall and a good accuracy rates that are very close to the results of logistic regression models for the test set. Also, as shown by the confusion matrices in Table 6, the model performance on predicting crash involvement is very similar to the logistic regression model, where almost all drivers are classified as not crash-involved, indicating that the model failed to learn to recognize crash-involved drivers. Further results are reported in Huang et al. (in review).

Table 5. Prediction Performance of Random Forest Models

Models	Training Set	Test Set

	Recall	Precision	Accuracy	Recall	Precision	Accuracy
RF_CNC	1.000	1.000	1.000	0.745	0.652	0.633
RF_Crash	1.000	1.000	1.000	0.014	0.333	0.800

Table 6. Confusion Matrices of random Forest Models

Models	Training Set				Test Set			
RF_CNC			Predicted				Predicted	
			0	1			0	1
	Actual	0	772 (44.9%)	0	Actual	0	158 (21.4%)	165 (22.3%)
		1	0	948 (55.1%)		1	106 (14.4%)	309 (41.9%)
RF_Crash			Predicted				Predicted	
			0	1			0	1
	Actual	0	1407 (81.8%)	0	Actual	0	589 (79.8%)	4 (0.5%)
		1	0	313 (18.2%)		1	143 (19.4%)	2 (0.3%)

Discussion

The main objective of this pilot project was to address a set of general research questions in order to lay the foundation for a larger effort on predictive crash analytics using large naturalistic crash and behavior datasets. In addition, a set of specific analyses were conducted with the purpose of demonstrating how this type of analysis can be carried out using large-scale naturalistic driving data and identifying open issues that can be addressed in the envisioned follow-up project.

The results of these analyses clearly demonstrate the presence of differential CNC involvement in line with the Pareto 80–20 rule. Moreover, this is at least partly related to enduring personal factors associated with individual drivers (Huang et al., 2018). The results from the classification models (Huang et al, in review) further showed that a driver's CNC involvement can be predicted with some degree of accuracy, precision, and recall based on measures representing enduring personal factors, such as driving style, demographics, personality, behavioral history, etc. However, these enduring personal factors seem more weakly correlated with crashes than with near crashes (de Winter et al., 2018) and the present classification models were unsuccessful in predicting crash involvement as opposed to CNC involvement. As further discussed in Huang et al. (in review), the fact that the training set contained more examples of near crashes than crashes likely led to better classification of the former (see de Winter et al., 2018). Another potential reason for the different model performances could be that the current predictor variables mainly capture individual characteristics related to aggressive driving and

aggressive driving may be more strongly associated with near crashes than with crashes. By contrast, crashes may, to a larger extent than near crashes, be associated with driver inattention combined with rare/unexpected circumstances. This is supported by existing naturalistic driving analyses of driver inattention and crash/near crash involvement, which typically found eyes-off-road to be more common in crashes than in near crashes (Klauer et al., 2006; Victor et al., 2015). Thus, one interesting avenue of further research would be to investigate to what extent driver inattention is associated with enduring personal factors (e.g., individual drivers may differ consistently in their willingness to engage in secondary tasks) and whether independent measures of secondary task engagement would bear a stronger relationship with crash involvement than the present variables.

In any case, it would be premature to dismiss the possibilities of predicting crash involvement from enduring personal factors solely based on the present results, and there are several ways the classification models may be improved and there are other ways of analyzing these data that may shed further light on the relationship between enduring personal factors and crash involvement. In particular, there is much room for developing more sophisticated driving style indicators that may have a stronger relationship to crash involvement. Some potential candidates include jerk (Bagdadi and Varhelyi, 2011) and various measures based on speeding and close following (see Sagberg et al., 2015). Moreover, as previously mentioned, the current predictors are mainly related to aggressive driving; including indirect or direct measures of "inattention propensity" may help to improve model predictions for crashes. These are all examples of possibilities that could be further investigated in a follow-up project.

Following, the more general research questions posed in the work plan are discussed based on the project results.

How can the relationship between individual driver behavior/driving style and crashes be conceptualized?

A conceptual framework was outlined based on existing literature, in particular Knipling (2009), and described in detail in Engström et al. (2017). A key idea is that driving style—persistent driving patterns characteristic of an individual driver (Sagberg et al, 2015)—can be understood as the manifestation of enduring personal factors related to demographics, personality, etc. Enduring personal factors combine with temporal personal factors, such as distraction and fatigue, and situational factors to produce a behavioral history of crashes, violations, etc. Accordingly, crashes generally occur due to interactions between enduring and temporal personal factors and situational factors, and can thus potentially be predicted, at least to some extent, based on variables that reflect individual characteristics (i.e., enduring personal factors). This framework proved very useful in guiding the present analyses and it is recommended that it be adopted in the potential follow-up project.

What are the minimum requirements on behavioral and crash data for enabling predictive crash analytics on a larger scale?

To enable the type of predictive models addressed by the present study, the data needs to contain detailed information on driver behavior as well as a large number, at least on the order of thousands, of crashes for the same drivers. The behavior observation data should include driving exposure data in mileage and/or driving hours and preferably be recorded continuously in order to obtain true rates of behavioral events. Driver screening (questionnaire) data on personality, self-reported driving history, etc. is also very useful as a complement to the observational data. Moreover, crashes and near crashes should ideally be coded with respect to severity and type.

The only currently existing naturalistic dataset meeting all these criteria is the SHRP 2 dataset used here. However, as discussed above, the number of crashes in SHRP 2 was still found to be too small for the present analysis, especially if lower severity crashes, such as tire strikes are taken out. The fact that the present, albeit limited, classification models only achieved successful performance on CNC data (dominated by near crashes) may be at least partially explained by the relatively small number of crashes. Commercial naturalistic datasets (e.g., SmartDrive and Lytx) contain larger numbers of crashes but are, due to their event-based nature, typically limited with respect to the information available on behavior in non-conflict situations. The rapid development of data logging technologies, in particular video analytics techniques able to automatically detect a variety of driver behaviors in normal (non-conflict) driving, offer great promise in generating novel types of large-scale naturalistic driver behavior and crash datasets that can be used for behavior-based predictive analytics.

What statistical methods are most appropriate for modeling the relationship between individual driver behaviors/driving styles and crashes based on naturalistic data?

The two classification models employed in the present analysis, logistic regression and RF, yielded very similar results, indicating that, at least for this classification task, it was more the data than the type of model that influenced the results. In addition, a number of other statistical approaches were tested in the project, such as the prediction of individual crash and CNC rates based on Poisson and negative binomial regression. However, these attempts were largely unsuccessful, possibly due to the fact that the crash/CNC data contained a large point mass of zero values (i.e., drivers with no crashes or near crashes), leading to distributions very different from the Poisson or negative binomial distributions.

How can commercial naturalistic data be made available for analysis while respecting legal, ethical, and business constraints?

The discussions with Lytx and SmartDrive showed that there is a general interest from both companies to participate in academic research projects like the present one, and the collaboration with SmartDrive further showed that, as long as the data is completely anonymized, there are no fundamental barriers for sharing the data for academic research. A non-disclosure agreement was

needed in order to protect the company's proprietary information, but this did not preclude the publication of the results in academic journals.

Conclusions and Recommendations

As described above, the results from the pilot project clearly demonstrated an association between individual enduring personal factors and the involvement in crashes and near crashes, while the prediction of crash involvement was less successful. This is likely due to a combination of the relatively low proportion of crash-involved drivers in the training data and a weak association between the currently used predictor variables and individual crash involvement. However, the current driving style variables were relatively simple and, in a follow-up project, there is clearly much room for exploring whether other types of metrics representing individual characteristics, such as close following and speeding behaviors and "inattention proneness" may be more strongly associated with crash involvement. It would also be interesting to analyze more thoroughly why the present models were able to predict individual involvement in near crashes, but not crashes. It is also recommended that future work involve a more in-depth exploration of alternative analytics models and different ways of organizing the behavioral and CNC data. Finally, it would be interesting to see to what extent the same model yields similar results on different datasets, such as SHRP 2 versus SmartDrive data.

To conclude, this pilot project represented an initial exploration of applying predictive analytics models to identify unsafe drivers based on naturalistic driving data. The project generated some interesting and promising results but barely scratched the surface with regard to the possibilities of performing this type of analysis. These possibilities could be explored in a follow-up project with a larger budget and scope.

Additional Products

Since this was a pilot project with limited scope and budget, the additional products are limited to the master dataset produced and a curriculum for a course module on Behavior-based Predictive Analytics, further described below. The final project dataset can be found on the Safe-D Dataverse.

Education and Workforce Development Products

Curriculum for a Course Module on Differential Crash Involvement and Behavior-based Predictive Analytics

A curriculum for a small course module on behavior-based predictive analytics is outlined in Appendix C. This is mainly intended as a guide and the module scope and content needs to be adapted to the context where it is to be applied (e.g., a larger graduate-level course on traffic safety).

Technology Transfer Products

The project has resulted in two publications and one presentation to date:

De Winter, J. C. F, Dreger, F. A., Huang, W., Miller, A., Soccolich, S., Ghanipoor Machiani, S., & Engstrom, J. (2018). The relationship between the Driver Behavior Questionnaire, Sensation Seeking Scale, and recorded crashes: A brief comment on Martinussen et al. (2017) and new data from SHRP 2. *Accident Analysis and Prevention, 118,* 54-56.

Huang, W., Engstrom, J., Miller, A., Jahangairi, A., Ghanipoor Machiani, S., Dreger, F. A., Soccolich, S., & de Winter, J. C. F. (2018). Modeling Differential Crash Involvement Based on SHRP 2 Naturalistic Driving Data. *Accident Analysis and Prevention.* Manuscript submitted for publication.

Huang, W., Engstrom, J., Miller, A., Dreger, F. A., Soccolich, S., de Winter, J. C. F., & Ghanipoor Machiani, S. (2018). Analysis of Differential Crash and Near-Crash Involvement Based on Naturalistic Driving Data. Presented at the 7th International Symposium on Naturalistic Driving Research. Blacksburg, Virginia.

Data Products

The uploaded dataset from this project to the Safe-D collection on VTTI Dataverse contains 2,800 drivers from the SHRP 2 data collection. Data include questionnaire factors on driver behaviors and risk perception, exposure metrics based on time, hours, and trips, crash-related data, and driver behavior variables mined from the 6-month study period. All data is at the driver-level and is continuous.

Appendix D contains the data dictionary used for the dataset, including the variables, variable labels, data types, and minimum/maximum values.

References

1. Bocanegra, J., Hickman, J.S., and Hanowski, R. (2016). Comparative Analysis of the Large Truck Crash Causation Study and Naturalistic Driving Data. FMCSA-RRR-13-018.

2. Boris and Luciana (2017). Developing a Younger Driver Assessment Tool. Technical Memorandum #1. American Transportation Research Institute (ATRI).

3. de Winter, J.C.F., Dreger, F.A., Huang, W., Miller, A., Soccolich, S., Machiani, S.G., Engström, J., (2018). The relationship between the Driver Behavior Questionnaire, Sensation Seeking Scale, and recorded crashes: A brief comment on Martinussen et al.(2017) and new data from SHRP 2. Accid. Anal. Prev. 118, 54–56.

4. 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. https://doi.org/10.1073/pnas.1513271113

5. Dingus, T.A., Hankey, J.M., Antin, J.F., Lee, S.E., Eichelberger, L., Stulce, K.E., McGraw, D., Perez, M., Stowe, L., (2015). Naturalistic driving study: Technical coordination and quality control.

6. Engström, J. Ghanipoor Machiani, S. Miller, A., Huang, W. and Soccolich, S. (2017). Behavior-based Predictive Safety Analytics, Deliverable 2.1: State-of-the-art review. Project Deliverable, Behavior-based Predictive Analytics, Safe-D (Safety Through Disruption) University Transportation Center.

7. Elander, J., West, R., & French, D. (1993). Behavioral correlates of individual differences in road traffic crash risk: An examination of methods and findings. *Psychological Bulletin*, *113*, 279–294.

8. Guo, F., S. G. Klauer, J. M. Hankey, and T. A. Dingus. 2010. Near Crashes as Crash Surrogate for Naturalistic Driving Studies. *Transportation Research Record*, 2147, 66-74.

9. 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.

10. Huang, W., Engström, J., Miller, A., Jahangiri, A., Machiani, S.G., Dreger, F., Soccolich, S and de Winter, J. (2018). Modeling differential crash involvement based on SHRP 2 naturalistic driving data. Manuscript in review.

11. Lytx. (2017). Industry insights: Beyond telematics: How video predicts risky behavior. Lytx White Paper.

12. Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations: a real distinction? *Ergonomics*, 33, 1315-1332.

13. Ryan, G. A., Legge, M., Rosman, D. (1998). Age rated changes in drivers' crash risk and crash type. *Accident Analysis and Prevention 30*(3), 379-387.

14. Sagberg, F., Selpi, S., Piccinini, G. B., Engstrom, J. (2015). A Review of Research on Driving Styles and Road Safety. *Human Factors, 57*(7), 1258-1275.

15. SmartDrive. (2017). *Measuring driver risk with video-based analytics.* San Diego, California: SmartDrive Systems.

Appendix A. State of the Art Review

Behavior-based Predictive Safety Analytics, Deliverable 2.1: State-of-the-art review

Johan Engstrom[1], Sahar Ghanipoor Machiani[2], Andrew Miller[1], Wenyan Huang[1] and Susan Soccolich[1]

[1] Virginia Tech Transportation Institute, Blacksburg, VA

[2] San Diego State University, San Diego, CA

Abstract

The report reviews the current state of the art in the field of Behavior-based Predictive Safety Analytics (BPSA). BPSA, as conceived here, addresses the relationship between drivers' personal characteristics, associated recurring behaviors and crash involvement. It is today well established that drivers are differentially involved in crashes and that these individual differences are associated with enduring personal factors such as demographics, health, personality and acquired skills. Moreover, the individual characteristics are reflected in recurring patterns of observable driver behaviors and records of behavior history (e.g., past traffic violations). Based on screening of personal characteristics and recording of their behavioral manifestations, statistical models can be developed that predict crash involvement for individual drivers. Such models can be used to identify risky drivers proactively and have a range of industrial applications, in particular in the domains of vehicle fleet management and usage-based insurance.

This review starts by outlining some key concepts related to differential crash involvement. Existing research on crash involvement prediction for individual drivers is then reviewed, focusing on three general types of independent variables (1) enduring personal factors, (2) behavioral history and (3) observable behavior. The report also reviews the main statistical concepts and techniques that have been used to model differential crash involvement, as well existing industrial applications.

It is concluded that research on BPSA, and differential crash involvement in general, is relatively scattered with few links between academic research and industrial development. Also, for the two main application domains, fleet management and usage-based insurance (UBI), the development of predictive analytics models appears to proceed more or less in parallel. Moreover, somewhat surprisingly, existing large sets of naturalistic driving data has so far not been extensively used in this context, at least not in academic research. It is clear that there is a need for more systematic and targeted academic research on BPSA with a great potential to rapidly advance the state-of-the-art in the field.

Contents

ABSTRACT 24

INTRODUCTION 27

DIFFERENTIAL CRASH INVOLVEMENT: KEY CONCEPTS 28

Crash causation 28

Personal versus situational crash causation factors 28

At fault 31

Exposure 31

REVIEW OF EXISTING FACTORS AND MODELS PREDICTING INDIVIDUAL CRASH INVOLVEMENT 31

Enduring personal factors 32
- Age and experience 32
- Gender 33
- Health 33
- Personality 34

Behavioral history 35

Observable driving behavior 37

STATISTICAL MODELS FOR PREDICTING INDIVIDUAL CRASH INVOLVEMENT 43

Regression models 44

Classification models 44
- Unsupervised learning 44
- Supervised learning 45

APPLICATIONS 47

Fleet Management 47

Usage-Based Insurance (UBI) 49

DISCUSSION AND CONCLUSIONS 50

REFERENCES 52

Introduction

In-depth crash investigation studies (e.g., Treat et al., 1977; Craft and Preslopsky, 2009) as well as naturalistic driving studies (Dingus et al., 2006; Dingus et al., 2016) have consistently shown that driver factors play a role in the great majority of all crashes. Driver factors contributing to crashes include temporary behaviors or states such as driver distraction and fatigue but also more enduring personal characteristics such as (lack of) skills, acquired habits, health issues and personality-related factors (Knipling, 2009). This report focuses on the role of enduring personal characteristics and the degree to which crash involvement for an individual driver can be predicted based on such factors or their behavioral manifestations.

While the old concept of "accident proneness" (Greenwood and Woods, 1929) has been more or less abandoned today, there is substantial evidence for differential crash involvement for both commercial and private drivers (e.g., Hanowski et al., 2000; Knipling, 2009; Knipling et al., 2004; Soccolich, Hickman and Hanowski, 2011; Simons-Morton et al., 2012; Guo and Fang, 2013). That individual differences play a key role in crash involvement is also a view widely held by safety managers in the transportation industry (Knipling et al., 2004). These differences have been related to a range of personal factors, such as gender, age, personality and health and may be manifested in recurrent patterns of observable driving behavior (e.g., speeding, close following and secondary task engagement) as well in records of behavioral history such as past violations, convictions and crashes.

The main focus of the present report, and the Behavior-based Predictive Safety Analytics (BPSA) project as a whole, is on the relationship between observable recurring patterns of driver behavior, that is the *driving style* of the individual driver (Sagberg et al., 2015), and crash risk. The advent of large sets of naturalistic driving data, including both a large number of crashes and records of driving behavior, yields exciting new possibilities in creating predictive models mapping from individual behavior patterns to crash involvement. Compared the more traditional approach of predicting individual crash involvement based on drivers' crash, violation and conviction records which accumulate over years (e.g., Murray et al., 2005, 2006; Lueck and Murray, 2011), the behavioral data needed for predictive models based on driving style can be collected in weeks (Guo and Fang, 2013; Smartdrive, 2017). While the focus of this review is on observable behavior, risk prediction based on direct screening of individual characteristics as well as behavioral history is also covered. These variables can be combined with observable behavior variables in multifactorial risk prediction models (e.g., "big data" analytics techniques).

As further reviewed below, predictive safety analytics have many applications, in particular in the areas of driver selection, fleet management and usage-based insurance, and there is a strong current commercial interest in these techniques. Furthermore, an enhanced understanding of the relationship between individual driver behavior and crash risk is of key importance for

estimating potential safety benefits of automated driving (AD) systems, since one of the key expected benefits of AD is the elimination of unsafe driving behaviors.

The report is organized as follows. The next section introduces some key concepts relevant to the topic of differential crash involvement among drivers. Section 3 reviews existing literature on individual crash involvement prediction based on (1) the direct measurement/screening of enduring personal factors, (2) behavioral history and (3) observable driver behavior. Section 4 then reviews statistical modelling techniques commonly used in differential crash involvement research and Section 5 gives examples of existing applications. Finally, Section 6 provides a summary and some general conclusions.

Differential crash involvement: Key concepts

Crash causation

Crashes typically occur through interaction between a multitude of factors related to the driver, the vehicle and the environment or current traffic situation. A useful way to think about crash causation is in terms of the Swiss cheese model (Reason, 1990) which pictures crash *defenses* as layered slices of Swiss cheese with holes in them. The holes represent different limitations in the defenses and when they become aligned a crash occurs. Knipling (2009) conceptualizes the Swiss cheese model specifically in terms of the relationship between driver error and crashes. Here, the slices represent things like "behavior while driving", "attention" and "road and traffic events" whereas the holes represents potential crash causation factors such "tailgating", "distraction" and "car cutting in". As Knipling (2009) points out, one limitation of the Swiss cheese metaphor when applied to driving is that any single factor (hole) may cause a crash although, in practice, crashes are caused by the interaction of two or more factors. Based on these ideas, Knipling (2009) outlines the more concrete Crash Trifecta model, further developed by Dunn, Hickman and Hanowski (2015). The key idea is that many crashes can be generally described in terms of three general elements which may or may not occur in combination (1) *unsafe pre-incident behavior or maneuver* (e.g., speeding, tailgating, unsafe turn), (2) *transient driver inattention* (e.g., an off-road glance) and (3) *an unexpected traffic event* (e.g., unexpected braking by the vehicle ahead).

Personal versus situational crash causation factors

The key question for present purposes is how one might think about the role of individual driver factors in crash causation. Early work addressing this question was motivated by observations that, after controlling for non-personal factors, crashes (or accidents in other domains) are typically not distributed among individuals in a way that could be explained by chance alone. In order to account for this individual variation, the concept of "accident proneness" was introduced (e.g., Greenwood and Woods, 1919). Thus, accident proneness was essentially a statistical concept ignoring the actual mechanisms underlying the individual differences. The adequacy of

this concept has been debated over the years (see review by McKenna, 1983) and today it is largely abandoned in favor of the "differential crash involvement" concept. The basic idea underlying differential crash involvement is that some drivers have certain personal characteristics that make them more likely to become involved in crashes. For example, this may be related to a stronger propensity for risk taking among certain drivers, leading these drivers to look away from the road for longer periods and more frequently than the average driver, increasing the risk for an off-road glance to co-occur with an unexpected event (e.g., a lead vehicle braking), thus leading to a crash. However, importantly, drivers' behavior (off road glances in this example) is also determined also by more temporary driver factors (e.g., a strong motivation to send a text message) or situational factors (e.g., driving in dense traffic requiring frequent mirror checks).

Knipling (2009) suggests a general distinction between *personal* and *situational* risk factors. The former refers to factors related to things "inside" the driver while situational factors refer to all other types of factors "outside" the driver, such as traffic, roadway, weather and the vehicle.

Personal factors may be further divided into *temporary* and *enduring* factors. The former relates to things that typically changes from day-to-day or hour-to-hour like colds, sleepiness and mood. By contrast, enduring factors, also referred to as *constitutional personal driver factors* (Knipling, 2009), refer to long term or permanent characteristics of a person such as gender, age, personality, driving experience, physical/sensory-motor abilities, skills, medical conditions and health, psychiatric & behavioral disorders etc. The present report is mainly concerned with enduring personal factors, since these are the fundamental factors underlying differential crash involvement.

Enduring factors may manifest themselves in terms of observable driver behavior patterns (e.g., speeding, close following, hard braking, involvement in traffic conflicts, engagement in distraction, fatigue etc.) as well as in (driving and non-driving) behavioral history (e.g., the number of crashes or violations in the past 3 year, criminal records, credit history etc.). Driving behaviors may then contribute to crash causation in different ways. As mentioned above, and reviewed in further detail below, there is strong evidence that enduring personal factors influence crash involvement beyond mere chance. However, it is also clear that they often interact with situational factors and temporary personal factors in non-trivial ways in producing crashes. For example, while the occurrence of driver fatigue can be regarded a temporary factor, there is strong evidence that the *susceptibility* to fatigue is an enduring factor (Knipling et al., 2004). A similar argument may be made for the role of alcohol in crash causation, as the effect of alcohol on behavior may depend strongly on enduring personality-related factors (see review in Elander et al., 1993). Thus, crashes often occur through an interaction between enduring personal factors, temporary personal factors and situational factors and the relative contributions of these different types of factors to crash genesis are often difficult to disentangle (Elander et al., 1993).

Figure 1 provides a general illustration of these concepts. Current (driving and non-driving) behavior is influenced by situational factors as well as temporary and enduring personal factors. This results in behavioral outcomes that may be successful (i.e., the situation played out as expected) or unsuccessful, leading for example to crashes, traffic conflicts (e.g., near crashes), violations and convictions. These events constitute the driver's behavioral history and certain events may become recorded in crash databases, naturalistic driving data, inspection records (for commercial vehicles), legal records, etc. Importantly, all this takes place in a sociocultural context (e.g., a national or workplace culture; Sagberg et al., 2015). Enduring personal factors will be reflected in recurring observable behavioral patterns as well as behavioral history, although both are also influenced by temporary personal and situational factors. Some of these recurring behaviors (e.g., tailgating, speeding, distraction) may be associated with increased crash risk. Hence, by developing statistical models that map from direct measurements of enduring personal factors (e.g., personality tests), observable driving style and/or behavioral history to crash involvement, it may be possible to identify unsafe drivers and before they become involved in crashes. The specific focus of this review is on such predictive models. Thus, while the literature on the general relationship between enduring personal factors, observable behavior, behavioral history and crash involvement is extensive, the scope of the present review is limited to studies that developed actual predictive models with the purpose to identify unsafe drivers. General overviews of research on individual driver characteristics and road safety is given in Knipling et al. (2004) and Knipling (2009). Sagberg et al. (2015) provides an extensive review of research related to driving style and safety.

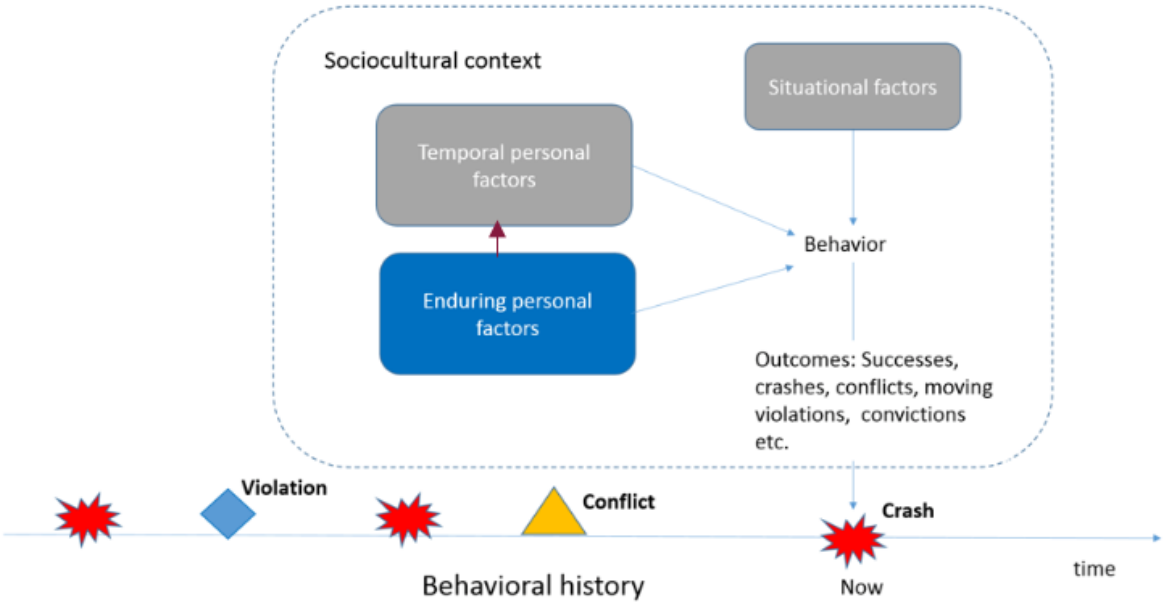

Figure 1. Conceptual framework for BPSA and differential crash involvement

At fault

In road safety research, a common distinction is that between “at-fault” and “not-at-fault crashes”. In the context of differential crash involvement research, such a distinction becomes particularly important as it would be expected that (both enduring and temporary) personal factors would play a stronger role in “at-fault” crashes. However, since “fault” refers to legal culpability rather than actual causal mechanisms alternative terms are preferable. For example, "active" versus "passive" crashes (Elander et al., 1993; West et al., 1992) refer to whether or not the behavior of the reporting driver played a role in the causation of the crash (and not whether the driver was legally responsible, which depends on the jurisdiction in the region where the crash took place). A study in which reported crashes were classified as active or passive showed that, whereas crash involvement in a 1-year period doubled the odds of crash involvement in a subsequent 2-year period, active crash involvement quadrupled the odds of a further active crash (West et al., 1992). Thus, more reliable measures of differential crash involvement could be obtained by focusing on active crashes only.

Exposure

A fundamental issue in differential crash involvement research is exposure (e.g., on what types of roads/traffic the driver is driving and how much or for how long). If exposure is not controlled for, it is impossible to determine whether individual differences in crash involvement is related to personal factors or just a result of some drivers simply being exposed to more risk than others (McKenna, 1983). As reviewed below, this has traditionally been a major issue in studies based on recorded or self-reported behavior history (crashes, violations etc.) where objective exposure data is typically not available and subjectively reported exposure is always subject to uncertainty (Elander et al., 1993). Thus, one of the key advantages of naturalistic driving data in differential crash involvement research is that precise estimates of exposure is typically readily available.

Review of existing factors and models predicting individual crash involvement

This section is structured based on the concepts introduced in the previous section (Figure 1). The next subsection addresses models predicting individual crash involvement directly on the basis of (measured or identified) ensuring personal factors such as age, gender, health and personality. Section 0 then reviews models based on behavioral history while section 0 focuses on predictive models based on observable driving behavior.

Enduring personal factors

Age and experience

Age is a key potential predictor of crash involvement. It is well-established that teenage drivers and older drivers are generally overrepresented in crash statistics (Massie et al., 1995; NHTSA, 2012). However, it is challenging to disentangle effects of age from effects of driving experience on crash risk where, traditionally, a key reason has been the lack of exposure information. For commercial vehicle drivers, even if exposure is available, the situation is further complicated by the fact that younger drivers are typically hired by different types of companies than older drivers and drive different types of vehicles (Knipling et al., 2004). These issues do not seem to apply to the same extent for young car drivers, and a recent review by McCartt et al. (2009) concluded that, for young car drivers, age and experience seem to affect crash risk independently of each other.

Effects of old age on crash risk are more uncertain. Some studies have found an overrepresentation of drivers older than 65 years in crashes per mile, but this may also be related to their typical low mileage (Elander, 1993). Older drivers also tend to self-regulate to limit their exposure to difficult driving situations (Ball et al., 1998), thus compensating for potential physical, perceptual or cognitive impairments. Knipling et al., (2004) suggests that there is little evidence of increased crash risk for older truck drivers, but a recent review concluded that crash risk among truck drivers starts increasing after the age of 63 (Duke, Guest and Boggess, 2010). A key issue here is that effects of age may be confounded due to better (safer) truck drivers staying longer on the job while the worst truck drivers (with a bad crash history) may get fired or change career voluntarily.

Statistical models using age as a predictor have yielded somewhat mixed results. Soccolich et al. (2011) investigated the extent to which age and a range of other anthropometric variables (see below) for truck drivers predicted their involvement in safety-critical events (crashes, near crashes, crash-relevant conflicts and non-intentional lane departures) in naturalistic commercial vehicle data. In a first step, drivers were grouped into three risk categories (Safe, Average and Risky) by means of statistical clustering. Next, differences in various demographics variables between the risk groups were tested (ANOVAs and Fisher's tests). It was found that the average age did not distinguish significantly between the risk groups.

In a study using naturalistic driving data collected from private passenger cars (the 100-car study; Dingus et al., 2005), Guo and Fang (2013) developed a statistical model for predicting crash and near crash (CNC) involvement in based on age, personality and critical incident rate. Results for the two latter factors are discussed further below. With respect to age, it was found that drivers under 25 years of age were strongly overrepresented in CNCs compared to drivers older than 25 years. In a negative binomial regression model, 'age under 25 years' versus '25-55 years' significantly predicted CNC rate. Moreover, similar to Soccolich et al. (2011), the drivers were classified into three risk categories (Low, Moderate, High risk) by means of cluster analysis

based on their CNC rates. Logistic regression models were then developed classifying high/moderate risk drivers vs. low risk drivers, or high risk drivers vs. moderate/low risk drivers. It was found that age was a significant predictor, but only in the former model. The discrepancy between the two studies may possibly be explained by the different driver populations, where teenage drivers most likely strongly contributed to the age effect in Guo and Fang (2013) while this age group was absent in the commercial vehicle dataset analyzed by Soccolich et al. (2011).

Abdel-Aty and Radwan (2000) used negative binomial models to test the effect of roadway conditions on accident occurrence. Researchers included permanent roadway features (such as lane, shoulder, and median widths and the roadway curvature, as well as several others). These researchers also considered how driver demographic variables, such as age or gender, affected the relationship between accident involvement and roadway features by building separate negative binomial models for each demographic variable level and comparing them. The models indicated female drivers experienced significantly more accidents than male drivers in the following roadway conditions: narrow lane width, reduced median width, larger number of traffic lanes, and heavy traffic volume. Age was also found to be associated with significant differences in accident rate by roadway features. Young drivers experienced more accidents on roadway curves and both young and older drivers had more accidents on roadways with reduced shoulder or median widths or heavy traffic. Speeding was associated with more accidents in males and young drivers.

To summarize, age appears to be a relatively reliable predictor of crash involvement, but mainly for passenger car driver populations involving young drivers, especially teenagers, and/or in combination with other variables such as roadway features.

Gender

Gender does not seem to be a strong predictor of crash risk. For example, in the analysis of Guo and Fang (2013), gender did not have a significant impact on CNC risk and was thus dropped from the model. In line with this Soccolich et al. (2011) did not find any differences due to gender between the three risk groups.

A recent white paper by CEI (2017) reports on the use of demographics information in a predictive safety analytics model developed for commercial vehicle fleets. The model includes age and gender as well as other demographics information such the industry in which the fleet operates and whether the driver assigned management role. However, no results on the predictive value of these variables are presented.

Health

Many studies have demonstrated a relationship between various anthropometric and health variables and involvement in crashes or other safety critical events. This includes obesity (Wiegand, Hanowski and McDonald, 2009; Anderson, Godava and Steffan, 2012), fatigue and sleep disorders (Howard et al., 2004), ADHD (Segal and Habinski, 2006).

Sagberg (2006) estimated relative crash risk in terms of odds ratios for a range of medical conditions and found significantly increased risk for the following conditions: non-medicated diabetes (OR = 3.08), a history of myocardial infarction (OR = 1.77), using glasses when driving (OR = 1.26), myopia (OR = 1.22), sleep onset insomnia (OR = 1.87), frequent tiredness (OR = 1.36), anxiety (OR = 3.15), feeling depressed (OR = 2.43), taking antidepressants (OR = 1.70) and for patients that had suffered a stroke (OR = 1.93).

However, these types of factors have seldom been used in predictive models intended to identify high risk drivers, one reason being that many of these conditions are relatively rare. One exception is Soccolich et al. (2011; see above) who, in their model based on naturalistic truck data, found that drivers who suffered head injury, inner ear problem, arthritis and motion sickness were significantly overrepresented in the high risk group.

Personality

Personality can be broadly construed as enduring psychological traits that affect behavior, or as "a style of interaction with the world" (Knipling, 2009). It is typically measured by means of subjective instruments representing different personality dimensions. One influential approach is the NEO inventory, also known as the "big-five" personality factors: neuroticism, extroversion, openness to experience, agreeableness and conscientiousness (e.g., Dahlen and White, 2006). Some evidence suggests that low agreeableness and low conscientiousness are the NEO dimensions most related to vehicle crash risk (Boris and Luciana, 2017). Schwebel et al. (2006) suggest that most personality dimensions proposed in the literature can be grouped into three main categories: (1) sensation seeking, (2) conscientiousness (here viewed as the antonym to impulsiveness) and (3) anger/hostility.

There is substantial evidence for a relationship between personality characteristics and unsafe driving behavior (Boris and Luciana, 2017; Jonah, 1997; Knipling et al., 2004; Knipling, 2009; Schwebel, 2006). Ulleberg and Rundmo (2003) further suggested, based on questionnaire data, that the relationship between personality traits and risky driving is mediated primarily through attitudes. On the other hand, Wilson and Greensmith (1983) did not find any difference in personality measures between groups of accident-involved and accident-free drivers, while they found several differences in driving behavior variables (see below). However, their personality inventories differed from those used in most other studies.[2]

Guo and Fang (2013) included the NEO five factor personality inventory in their statistical model developed to predict crashes and near crashes (CNC) in the 100-car study dataset based on age, personality and critical incident events (see above and further below). It was found that the five NEO variables correlated strongly and they were thus reduced to a single dimension by

[2] The personality screening instruments used by Wilson and Greensmith (1983) were the Eysenck Personality Inventory, Form A; the IPAT anxiety questionnaire and the Parry aggression and anxiety questionnaire.

means of PCA before being used in the model. In a negative binomial regression model the composite personality variable was found to be a significant predictor of CNC rate. However, the personality score was not a significant predictor in the logistic regression model classifying high vs. moderate/low risk drivers (or high/moderate vs. low risk drivers).

Behavioral history

One important line of research on individual crash risk has related drivers' behavioral history in terms of past inspection violations, convictions and crashes to future crash risk. A key advantage of this approach is the availability (at least in the US) of large amounts of violation, conviction[3] and historical crash data from government records (e.g., the Motor Vehicle Record, the Commercial Driver's License Information System (CDLIS) and the Motor Carrier Management Information System (MCMIS). It is also common practice among carriers to monitor drivers and evaluate them based on the number a crashes and or traffic violations in a certain time period (e.g., 3 years).

Studies have typically demonstrated strong relationships between behavioral history variables and crash involvement, both for private and commercial drivers. For example, West et al. (1992) reported that the odds of having a crash in a 2-year period were doubled for private car drivers who had had one or more crashes in the preceding year. Several commercial vehicle studies have demonstrated a significant relationship between traffic citation and conviction history and crash rates at the carrier level (Lantz and Blevins, 1997; Lantz, Loftus and Keane, 2004; Knipling, Olsen and Prailey; 2004). However, due to the high turnover of driver at fleets, carrier-level safety measures based on individual driver behavior may not be stable over time. For this reason, the American Transportation Research Institute (ATRI) conducted a study in 2005 with the goal to predict crash involvement from individual truck driver behavior history (Murray, Lantz and Kepler, 2005; Murray et al., 2006) which was more recently followed up with a similar study (Lueck and Murray, 2011). Both studies were based on behavior history data from CDLIS and MCMIS.

CDLIS is a nationwide source of commercial drivers' traffic conviction data and MCMIS is a centralized database of carrier-based information about accidents and roadside inspections of commercial motor vehicles and drivers, originally collected at the state-level and maintained by FMCSA. *Violations* are issued to drivers during roadside inspections when inspectors discover that a driver and/or vehicle is not in compliance with one or more of the Federal Motor Carrier Safety Regulations (FMCSRs) and are recorded in MCMIS. By contrast, *convictions*, recorded in CDLIS, represent cases where a driver has been issued a citation and being found guilty of the

[3] Convictions refer to the subset of citations that have gone through the adjudication process.

specific charge in court. Thus, a similar behavior may show up both as a violation and as a conviction but there is not necessarily a one-to-one relationship between them.

In the 2005 study (Murray et al., 2005; Murray et al., 2006), drivers that had a roadside inspection during the prior three months (February to April 2004) were identified in MCMIS. This lead to a sample of 540,750 US drivers. For these drivers, inspection violations, convictions and crashes that occurred during the period of February 2001-April 2004 were obtained from MCMIS and CDLIS. A Chi-square analysis (methodological details are not reported) showed that a variety of violations and convictions, as well as past crashes, were strongly associated with increased future crash probability. However, it is not clear from the report if the three-year data sample was split up in some way to test the predictability of behaviors recorded during the first period on crashes during the second period.

With respect to violations, the highest increase in crash likelihood was obtained for the following behaviors (percentages indicate the increase in crash likelihood):

- Reckless driving violation (325%)
- Improper turn violation (105%)
- Improper lane change violation1 (78%)
- Failure to yield right of way violation (70%)
- False/no log book violation (51%)

The top-five convictions associated with crash risk were

- Improper or erratic lane changes conviction (100%)
- Failure to yield right of way conviction (97%)
- Improper turn conviction (94%)
- Failure to keep in proper lane conviction (91%)

Finally, a past crash increased the crash likelihood by 87 percent. The authors also fitted a logistic regression model to predict crash risk by means of stepwise inclusion of the violation/conviction/crash variables.

In the 2011 replication of the 2005 ATRI study (Lueck and Murray, 2011), US drivers that received a roadside inspection or had been involved in a crash during January to March 2010 were included in the analysis, resulting in 587,772 drivers. Violations, convictions and crashes that occurred during the calendar year 2008 were then used to predict crash involvement during 2009. A Chi-square analysis similar to that performed in the 2005 study yielded the following results for violations:

- Improper passing violation (88%)
- Hours-of-Service violation (45%)
- False/no log book violation (42%)
- An Improper Lane Change violation (41%)

- A Following Too Close violation (41%)

The convictions most associated with an increase in crash risk were:

- A Failure to Use/Improper Signal conviction (96%)
- An Improper Passing violation (88%)
- An Improper Turn conviction (84%)
- An Improper or Erratic Lane Changes conviction (80%)
- An Improper Lane/Location conviction (68%)

A past crash was associated with an increased likelihood of 88%.

While the majority of the behaviors analyzed (73.5%) had a stable relationship to crash risk for the two ATRI studies, there were also several major differences. For example, half of the top-ten behaviors in the 2005 study were non-significant in the 2011 study. In general, the associations between behaviors and crash involvement were weaker in the 2011 study.

The authors discuss different possible explanations for these discrepancies including changes in how conviction data is handled, technological advances (e.g., replacing log books with electronic data recorders), a reduced overall prevalence of violations/convictions and crashes (reducing statistical significance) and the successful implementation of countermeasures addressing several of the most problematic behaviors during the six years that passed between the studies.

A recent white paper by the fleet management company CEI (CEI, 2017) describes a driver risk prediction model based on a 5-year behavior history data enhanced with driver demographics (e.g., age and gender) and geographic data, employing modern "big data" predictive analytics techniques. The company claims that this model outperforms traditional models based on behavior history only and is able to predict actual crash involvement with a high degree of accuracy. However, no details of the model are disclosed.

As mentioned above, and discussed in Knipling (2009), a general issue with this type of behavior history analysis is that exposure data is not available. It is thus theoretically possible that the obtained relationship between historical behaviors and future crash risk is at least partially confounded by driving exposure (i.e., the drivers who drove the most had the most crashes as well as violations and convictions). The same holds for studies based on self-reported crashes, although some of these have addressed this problem by also including self-reported driving exposure (Elander et al., 1993).

Observable driving behavior

There is a long-standing tradition of research indicating a relationship between individual characteristics of observable patterns of driving behavior and crash involvement (Sagberg et al., 2015). As noted above, a key advantage of using observable behavior data as input to crash prediction models is that behavioral events (such as speeding or high g-forces) occur more

frequently than, for example, violations and convictions, which enables predictions based on a shorter time window of data. This is important especially for evaluating newly licensed drivers and commercial drivers that are new to a company. Moreover, when naturalistic driving data is used to obtain observable behavior, exposure information is normally readily available.

In a classical study, Tillmann and Hobbs (1949) performed interviews with crash-involved and crash-free taxi drivers during taxi trips. They observed that taxi drivers with a high previous crash involvement…

> "…were easily distracted while driving. They tended to be readily annoyed at other motorists on the road, often criticizing their own driving mistakes in others. Horn honking and racing other cars away from a stop light were their specialties. (p. 325; cited by Sagberg et al., 2015).

By contrast, taxi drivers with low crash rate…

> "…were serious when driving and often refused to talk. They tended to be courteous to other drivers on the road and stated that they were conscious of the fact that the other driver might do the wrong thing. They appreciated the possible limitations of their vehicle." (p. 326; cited by Sagberg et al., 2015).

Based on these results, their oft-cited conclusion was "a man drives as he lives".

Probably the earliest attempts to identify risky drivers based on actual driving behavior measurements is the work by Greenshields and Platt (Greenshields, 1963; Greenshields and Platt, 1967) based on the "Drivometer", an early apparatus for automatically collecting driving data in an instrumented vehicle. An initial study (Greenshields, 1963) demonstrated that a group of drivers with a history of high crash involvement exhibited significantly more frequent accelerator pedal reversals than a control group of driving instructors. In a second study (Greenshields and Platt, 1967), drivers recruited based on insurance data were divided into four categories: (1) drivers with high previous crash involvement (but low number of violations), (2) drivers with a high number of traffic violations (but low number of crashes), (3) novice drivers and (4) a control group of ordinary drivers and driving instructors with low crash involvement. The participants drove a pre-defined route during which a variety of driving behavior measures were collected by vehicle sensors combined with real-time observer event annotation. A discriminant classifier which included the variables *running time* (the total time the vehicle was moving), *accelerator pedal reversals*, *brake applications*, *"gross" steering-wheel reversals*, and *"micro" steering-wheel reversals* was able to classify a set of novel participants (previously unseen by the model) into the four groups with an accuracy of 67-100%.

Wilson and Greensmith (1983) conducted a similar using the same "Drivometer" apparatus developed by Greenshields and Platt, but included a wider range of driving performance variables. One hundred drivers were categorized into six groups based on self-reported accident involvement, driving exposure (mileage per year) and gender. The driving variables used in the analysis included:

- Run time

- Speed changes
- Fine steering reversals
- Coarse steering reversals
- Accelerator applications
- Brake applications
- Moderate lateral acceleration events (> 0.15 g was reached on a bend for 1 s or longer)
- Strong lateral acceleration events (> 0.3 g was reached on a bend for 1 s or longer)
- Gear changes
- Mean clear speeds (the speed in free driving conditions at a certain location)
- Signals (all signals made by the subject whether by hand or mechanically)
- Frequency of overtaking
- Frequency of being overtaken

As mentioned above, the study also involved several personality inventories but the personality scores did not differ between any of the participant groups. The participants drove a pre-defined route of about 50 km in different road types and conditions (rural, secondary and main highway). It was found that compared to accident-involved drivers, accident free drivers had a higher frequency of fine steering wheel reversals, drove at lower mean clear speeds, overtook other drivers less frequently, and were overtaken more frequently. Discriminant functions were used to classify individual drivers into their respective groups and 76% of the accident free drivers and 68% of the accident-involved drivers were correctly classified (it appears that the model was tested on the same dataset used to train it). The general conclusion from the study was that accident-involved drivers to a greater extent continually moves about in traffic and drive at faster preferred speeds.

West et al. (1993) had 48 drivers driving a pre-define route under observation by a test leader and also collected self-reports on crash involvement during the past three years. A logistic regression analysis showed that preferred speed on the motorway, but not maximum speed, significantly predicted self-reported crash involvement.

It should be noted that the studies reviewed so far all were controlled experiments, using pre-defined routes, where some performance measures were only taken at pre-defined locations (e.g., preferred speed on a motorway section). This likely helped to isolate the effects of personal factors by at least partly controlling for variance induced by situational factors (e.g., variations in road layout, traffic density etc.). This probably greatly simplified the classification task but such luxury is clearly not available to applied crash prediction models that have to deal with real-world driving data.

A series of roadside naturalistic observation studies by Evans and Wasielewski (Evans and Wasielewski, 1982, 1983; Wasielewski, 1984) investigated the relationship between headway, speed, accident involvement as well as several other variables. Headways and free-driving speed on an urban highway were measured from the roadside and the drivers and the vehicles' license

plate were photographed. From the license plate number, information regarding the vehicle and the vehicle owner, including gender, age and driving record (previous accidents and violations) was obtained. The results demonstrated significant, but relatively weak, correlations between the drivers' accident involvement history and the adopted headway (Evans and Wasielewski, 1982, 1983) as well as between accident involvement history and free speed (Wasielewski, 1984). Furthermore, it was demonstrated that the adopted speeds and headways for the same driver showed some consistency across repeated observations. However, the correlations between the repeated measures were quite weak indicating a strong influence of situational factors. The studies also found correlations between accident involvement and other driver characteristics and behavior variables such age, gender, violation points and seat belt use.

af Wåhlberg (e.g., 2006) has developed and evaluated the "celeration behavior theory" suggesting that "…all speed changes denote a (very miniscule) risk of accident, and therefore predict that the sum of all such changes (celerations) of a driver will be equal to his/her accident record, when both variables have been standardized." The term "celeration" here refers to the sum of longitudinal and lateral, positive and negative, accelerations. The theory has been tested in several studies with Swedish transit bus drivers driving fixed routes, for which crash records were available from the bus company. The study described in af Wåhlberg (2006) was based on data from 3 years involving about 250 drivers**.** The association between celeration measurements and accident involvement was quantified in terms of Pearson correlations. The correlations between the number of accidents and mean longitudinal acceleration (which for practical reasons was used as a proxy for celeration in this study) were in the range of 0.2-0.3 depending on the driver sample used. The corresponding correlations for speed measures were slightly lower, around 0.2.

Bagdadi and Varhelyi (2011) investigated the relationship between *jerk* (i.e., the time derivative of acceleration) and self-reported crash involvement using an existing dataset of 166 car drivers in Sweden. Using Poisson regression modelling, they found jerk rate to be a significant predictor of self-reported crash involvement.

The studies reviewed so far have suffered from a number of issues, such as limited sample sizes, the use of pre-determined routes which restricts the influence of situational factors compared to everyday, naturalistic driving and/or the lack of exposure (mileage) information in the accident records or self-reported accident involvement (as noted above, this is also a problem in the behavioral history studies). Thus, a major breakthrough in the study of individual driver risk has come with the advent of large-scale naturalistic driving data. This allows individual driving behavior variables to be associated with observed crashes and near crashes in the same dataset, accounting for exposure in a precise way. Moreover, video-recordings allow for a detailed analysis of crash and near crash causation and risk.

Probably the first published application of naturalistic driving data to crash prediction at the individual level is the study by Simons-Morton et al. (2012), who analyzed the relationship

between gravitational-force (g-force) events and at fault crash/near crash (CNC) rate for 42 newly licensed teenage drivers. Data for each driver was collected for 18 months. Five different types of g-force events where calculated: rapid starts (>0.35 g), hard stops (<-0.45 g), hard left/right turns (<+-0.05 g) and yaw (6 degrees in 3 seconds). These were combined into a composite metric used in the statistical modelling (models retaining the individual metrics were also tested but did not improve performance over the models using the composite metric). The results showed a strongly significant Spearman correlation of 0.6 between the composite metric of g-force events and CNC rates during the 18 month period. Furthermore, a logistic regression model showed that the rate of g-force events in the prior month predicted CNC risk (for the presence of a crash) in the following month with a 76% prediction accuracy. A further analysis reported in Simons-Morton et al. (2013) found that the g-force rates for individual drivers during the first 6 months were similar to individual rates during the next two 6 month periods, demonstrating consistency in individual risky driving.

Guo and Fang (2013) conducted a similar analysis using the 100-car naturalistic driving study data. As reviewed above, they found that age and personality were significant predictors of CNC rate. However, the strongest predictor was the rate of critical-incidence events (CIEs). CIE were not directly measured but coded in the data based on: (1) Identification of events where the car sensors exceeded a specified value (e.g., brake response of >0.6 g (the exact kinematic criteria are not reported), (2) when the driver pressed an incident push-button and (3) through data reductionists' judgments when reviewing the video. As also mentioned above, a logistic regression model based on the three predictors successfully classified drivers into two risk categories defined by clustering the rate of CNC involvement.

While these naturalistic driving studies have significantly advanced the state of the art in predictive analytics modelling of individual risk, a key limitation is that the CNCs that the models are trained to predict are dominated by low-severity crashes and near-crashes. Today, the SHRP 2 naturalistic study, with more than 3500 vehicles, provides a much larger dataset with a significant number of crashes (+1500), many of which are relatively severe. However, to the knowledge of the authors, the SHRP 2 dataset has so far not been analyzed with respect to individual driver risk.

Even larger sets of video-recorded naturalistic events exists are today collected by private companies such as Lytx and SmartDrive who instrument hundreds of thousands of vehicles with event-triggered video data recorders as part of safety management services offered to commercial vehicle fleets. Videos of safety critical events identified by kinematic triggers (and more recently also by means of onboard sensors), are further analyzed by data reductionists in order to identify and annotate unsafe behaviors. The reduced data is then used by fleets to identify risky drivers and coach them towards safer driving. As a result, thousands of crashes and large quantities of behavioral events are collected each month. Behavior-based predictive safety analytics is a key part of the services offered by these companies. Although no published studies are available on these models, some general descriptions of methodologies and results are available in recent white papers.

SmartDrive (2017) offers a relatively detailed description of how this type of event-triggered naturalistic driving data can be used to estimate individual driver risk, although no technical details are disclosed. The SmartDrive white paper argues that historical collision data is inadequate and impractical for identifying risky drivers since (1) collisions are rare and (2) some drivers do not continue driving after they have a collision. The paper acknowledges the traditional approach (review in the previous section) of using motor vehicle records, traffic citations and roadside inspection to estimate a drivers' risk rate but argues that, since also this data is relatively sparse, it is better suited to evaluate risk at the fleet level. It is suggested that a better approach is to observe driver behaviors and link these behaviors to collision involvement by means of predictive analytics models. This shorter timescale also enables more effective interventions for risky drivers.

The white paper (SmartDrive, 2017) describes an example analysis using data from more than 27,000 drivers, sampled over two years, resulting in 6.3 million video recordings associated with 18.6 million driving hours. This data clearly confirms the phenomenon of differential crash involvement: A small proportion of the drivers typically account for a large proportion of the crash risk. SmartDrive calculates the driver risk in terms of a safety score based on behavioral observations in triggered video events. These events may be triggered by strong g-forces but also by other information such as sensor information available from the vehicle's onboard data networks. It is argued that video observation is superior to merely analyzing driving performance data since the video helps to clarify whether the event is actually safety relevant. SmartDrive also collects continuous exposure information for the risk calculations. The behavior observation data is then correlated with past collisions through predictive algorithms (not further disclosed in the paper). These correlations are calculated separately for each industry (service, transit and trucking). SmartDrive claim to have identified 27 observable behaviors that occur at least 20% more frequently for collision-involved drivers than non-collision drivers. Of these, 14 behaviors occur at least 50% more frequently for collision drivers. Though all behaviors are not disclosed in the paper, some examples are included, such as talking on a hands free mobile phone (26 vs 17 observations per 1000 driving hours), an unfastened driver's seatbelt, (14 vs 12 observations per 1000 driving hours) and mobile phone texting/dialing (1.6 vs. 1.3 observations per 1000 driving hours).

The paper presents results showing that the safety scores for collision drivers are strongly increased compared to non-collision drivers for (70%, 49% and 40% for public transit, service and trucking industries respectively). Furthermore, the safety score correlates with collision rate. The key advantage, compared to the traditional behavior history approach, is that these predictions can be obtained based on a past time window of weeks rather than months or years.

Lytx has recently issued a similar white paper (Lytx, 2017), albeit somewhat less detailed than SmartDrive (2017). Like SmartDrive, Lytx argues for the strong value of video in understanding crash causation (as opposed to driving performance indicators typically calculated based on vehicle data commonly in "Telematics" fleet management devices; see the "Applications"

section below). The white paper describes results from a study based on 10,000 drivers from 50 fleets who triggered more than 150,000 events. Collision likelihood is presented for various behaviors such as not checking mirrors (2 times risk increase), running a stop sign (1.8 times risk increase), food/drink distraction (1.8 times risk increase) and driving at a following distance of 1.25-1.75 s (1.6 times risk increase). However, the white paper does not disclose any further information on how these numbers were obtained.

To summarize, the idea of predicting individual crash involvement based on observable driving behavior is not new and predictive models based on collected data dates back to the 1960's. However, until recently research in this domain has been subject to a number of challenges related to a lack of a common source of exposure, behavior and crash data. In recent years, this situation has changed due to the advent of large scale naturalistic driving data, including very large event-triggered datasets collected by the private sector from fleets of hundreds of thousands of vehicles). However, somewhat surprisingly, academic research using naturalistic driving data to study differential crash involvement is still relatively sparse although existing studies have shown very promising results (e.g., Simons-Morton et al., 2012; Guo and Fang, 2013).

Statistical models for predicting individual crash involvement

Research on predicting individual crash involvement has employed a wide variety of statistical techniques. This section provides an overview of the main types of models found in the literature. In addition, more advanced machine learning techniques such deep learning have recently become popular in the context of "big data analytics" with the increase in both available computational power and the amount of raw data (Goodfellow, Bengio, & Courville, 2016). Since these techniques are potentially applicable in the present context, they are reviewed here as well.

A general distinction can be made between regression and classification models. The key difference between regression and classification is the nature of dependent variable. Regression is the technique used to predict a numeric value of the dependent variable (e.g. the frequency of crash), on the basis of one or more independent variables. By contrast, classification models predict which of a set of categories (e.g. low and high risk driver) a new observation belongs to, on the basis of a training set of data containing observations whose category membership is known (supervised learning) or unknown (unsupervised learning). Moreover, regression analysis can also be used to investigate causal or correlated relationships between the independent and dependent variables.

Regression models

A commonly used statistical method in the field of driving safety and crash risk is the Poisson regression model. Poisson regression models predict the frequency of discrete events as a function of an exposure variable and at least one covariate. These properties make the model a natural fit for predicting discrete events in driving data, which may include crashes, drowsiness epochs, or other safety-related events (Knipling et al., 2004). Examples of exposure measures include hours or vehicle miles traveled. Because drivers can contribute to several discrete events, a driver's safety-related events are correlated. To account for this models typically include a random effects variable.

It is important to be aware of a key assumption of the Poisson distribution: the assumption that the mean and variance of the predicted variable must be equal in value. In data sets where this assumption has been violated, the data are said to be exhibiting under- or overdispersion (the variability of the data is less than or greater than expected if the data followed a Poisson distribution). In these cases, negative binomial regression models have been used and is now established as an industry standard (e.g., Guo and Fang, 2013; Shankar, Mannering, and Barfield, 1994; Abdel-Aty and Radwan, 2000).

Poisson and negative binomial models have become widely-accepted methods to evaluate the relationship between discrete events (e.g., crashes or near crashes) and driver-related factors, such as driver demographics or measures of personality. Example applications in the behavior-based predictive analytics field include Guo and Fang (2013) and Özkan et al. (2006). Interestingly, as mentioned above, these statistical techniques formed the basis for the early work on accident proneness in the early 20th century (Greenwood and Woods, 1929; see McKenna, 1983).

Classification models

As mentioned above, classification models can be further divided into those using unsupervised and supervised learning respectively. Examples of these techniques are reviewed below.

Unsupervised learning

Cluster analysis is a statistical method used to categorize individual drivers into different risk groups with similar meanings (homogeneous groups), which minimize within-group variation and maximize between-group variation (Constantinescu, Marinoiu, & Vladoiu, 2010). In the context of different crash involvement research, cluster analysis is typically used to divide subjects into "high risk" versus "low risk" groups, for example, based on their crash/SCE rate (e.g., Guo and Fang, 2013; Soccolich et al., 2011; reviewed above) or braking process features including maximum deceleration, average deceleration, and kinetic energy reduction (Wang et al., 2015).

Through cluster analysis, risk groups (high-, moderate-, and low-risk groups) can be identified. ANOVA and Fisher's tests can then be adopted to investigate the effects of each independent

variable. If the independent variable is continuous, ANOVA can be used to investigate whether means of that variable were significantly different across the different risk groups (Soccolich et al., 2011). If the independent variable is categorical, Fisher's test can be used to investigate whether that variable was equally distributed across different risk groups.

Principal Component Analysis (PCA) is a statistical method used to identify a smaller set of uncorrelated significant principal components (PCs) orthogonally transformed from a set of correlated observed variables, which reduces the dimensionality of the redundant and correlated variables and eliminates the multicollinearity issue in subsequent modeling. The first PC accounts for the highest proportion of the variability of the data, and each succeeding PC accounts for the highest possible variance among the remaining components (Constantinescu et al., 2010; Guo and Fang, 2013; Guo, Fang, & Antin, 2015).

As reviewed above, Guo & Fang (2013) used PCA to eliminate the multicollinearity among personality measures. In this case, the first PC (67.3% of the variability) was selected to represent the personality scores using the eigenvalue-one criterion. Guo, Fang, & Antin (2015) modelled the CNC rates based on 53 assessment metrics, which were divided into four categories including physical ability, visual ability, health, cognitive ability. In order to eliminate the multicollinearity within categories, PCA was performed for each category and significant components were selected to represent each category using the eigenvalue-one criterion. As a result, 16 PCs instead of 53 metrics were used to model CNC rate.

Another unsupervised technique is latent class analysis, which is useful for identifying unmeasured class membership and can be used to group drivers into categories based on their behaviors, similar to clustering or principle component analysis. Unmeasured class membership are represented by latent variables, variables that are not observed, but are underlying and inferred from other variables, such as driving style. Latent class analysis uses a probabilistic model to describe classes based on these latent variables, as opposed to Bayesian modeling and clustering which find similarities between observed cases with no inference on underlying variables. Roman et al. (2015) used a latent class growth model to predict unsafe behaviors in novice drivers over time by classifying them based on self-report behavioral questionnaires over a six-year study period.

Supervised learning

Supervised learning is used to train models to classify data (in the present context personal factors, behavior history or observable behavior) into a set of pre-defined target categories (e.g., high-risk versus low-risk drivers) for which training data (associated input and target pairs) is available.

Logistic regression is a statistical method used to model the probability of being a high-risk driver (Guo & Fang, 2013) or the probability of encountering a crash (Murray, Lantz, & Keppler, 2005). The dependent variable is a binary variable and is assumed to follow a Bernoulli

distribution with a probability (p_i). This probability is associated with a set of covariates by a logit link function.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i,1} + +\beta_2 x_{i,2} + \cdots + +\beta_k x_{i,k}$$

The Odds Ratio ($OR_j = \exp(\beta_j)$) is calculated to quantitatively evaluate the impact of each variable. The OR represents the relative odds of being a risky driver or encountering a crash for every one unit increase in a continuous variable, or relative risk between two levels of a categorical variable. For example, Guo & Fang (2013) concluded that CIE rate had a significant impact on the probability of being a high-risk driver, and that every one unit increase in CIE rate will increase the relative odds of being a high-risk driver by 41% (OR = 1.41).

Another classification technique that has been utilized in predicting individual crash risk is discriminant analysis. Discriminant analysis is used in a similar manner to logistic regression, but can predict two or more classes of events using one or multiple continuous or binary independent variables and may include non-linear relationships. As reviewed above, discriminant analysis was used by Greenshields and Platt (1967) in their early work on predicting crash involvement from recorded driver behavior patterns. More recently, Ba and authors (2017) use a linear and quadratic discriminant analysis and a supervised learning model with behavioral and physiological features to predict crashes in driving simulator data. Driver-related variables, including individual demographics, driving history, and driving behaviors related to vehicle control were included in the model to predict crash outcome. The authors found inclusion of drivers' states and traits as inputs into the model explained additional variance above vehicle-only dynamics in predicting crashes.

More advanced machine learning techniques have been used to predict crash risk among drivers based on their behaviors and characteristics, along with other factors that may influence crash risk including environmental variables. Zhu and authors (2017) created a model of the behavior-risk relationship using a hierarchical Bayesian network model, presenting each driver with an individual confidence level and each node in the neural network with its own distribution. Using GPS driving observations and driver characteristics, the authors found that the inclusion of driver behaviors and inputs, contextual information, such as environmental factors, and the interactive effects between driver and environment, lead to the best performing model, identifying certain driver behaviors and characteristics as leading to higher crash risk, such as age, gender, speed relative to other vehicles, and freeway merging speed. A second study created a Bayesian network modeling crash risk based on locations and similar driver-related parameters using simulated data (Gregoriades & Mouskos, 2013).

Deep learning, a technique that allows a computer system to improve with experience and data in prediction, is a kind of machine learning with a lot of flexibility and power by representing the world as a nested hierarchy of concepts (Goodfellow et al., 2016). Performing a deep learning

algorithm allows a computer to build complex representations out of simpler representations. For example, in the context of machine vision, computers are able to identify an object based on simple representations of data, like the inputted pixels, then examining the pictures for edges, corners and contours, and object parts. In the context of driver risk, the input of driver history, characteristics and demographics, behaviors, and environmental factors could lead to an accurate algorithm capable of predicting an individual's crash risk.

While some studies have used a machine learning approach (e.g. decision-tree modeling, Hu et al., 2017; elastic net regularized multinomial logistic regression, Arbabzadeh & Jafari, 2017) to determine crash risk based on driver behaviors and other variables, few have utilized deep learning algorithms to predict crash risk of individuals from driver behavior and characteristics. Yin and authors (2017) used convolutional neural network to reduce a large amount of naturalistic data using inputs they created based on lane deviation and following distance. They then used a trained Gradient Boosting Decision Tree to characterize vehicle operators as either good or bad drivers. The authors allude to their classification of bad drivers as having a higher crash risk, but were unable to test this model with current data, as there are no crashes to supervise the neural network.

Applications

This section reviews existing applications of behavior-based predictive safety analytics, focusing on general domains of (1) fleet management and (2) usage based insurance (UBI).

Fleet Management

Applications of predictive analytics for differential crash involvement using telematics and onboard safety monitoring systems have provided fleet management companies with a powerful decision-making tool for risk assessment. Data analytics could impact fleet operations in several aspects from accident prevention and driver selection, training, and retention to liability and maintenance.

Vehicle crashes are one of the leading causes of death in US (National Highway Traffic Safety Administration [NHTSA], 2015). Considering that fleet employees spend most of their time on the road, they are highly affected by these statistics. A portion of these accidents such as accidents caused by driver distraction, driving under influence, and speeding are preventable accidents if the driver in question does not fail to exercise every reasonable precaution to prevent the accident. Preventing these accidents is of vital importance for fleet managers not only in terms of saving lives but also due to liability exposure and financial losses. Accidents contribute to 14% of a fleet's total expenses (Suizo, 2015). Traditionally, addressing driver behavior and errors was a reactive task. With the advent of telematics and data collection devices on board, this task is now becoming more proactive. Advances in accident prevention technologies using data analytics techniques have made it possible to reduce number of preventable crashes

resulting in saving money and reducing injuries and fatalities. For example, CEI company ("CEI Fleet Driver Management," n.d.) reports an average of 15% reduction in accident rate using their fleet driver safety and risk management application.

Applications of onboard safety monitoring systems extend beyond crash prevention. It is applicable to avoiding litigation and mitigating liability; in an event of an accident, a provable history of dedication to safety matters (Omnitracs, 2016). For example, a record of interventions could help with avoiding negligent entrustment issues ("CEI Fleet Driver Management," n.d.) and exonerating the driver and protecting from liability. Data could also be used to assess involvement in the accident and estimate the potential payout with more accuracy and set reserve accordingly (Omnitracs, 2016). Monitoring telemetry data could also offer insights into vehicle preventive maintenance and productive maintenance scheduling by providing operational updates and identifying possible failures ("Designing a Connected Vehicle Platform on Cloud IoT Core | Solutions," n.d.).

Another application of telemetry data for fleet management is related to behavior-based safety (BBS) programs. BBS programs have been deployed across a variety of industries. However, application of BBS to changing behavior and reducing crashes in commercial fleets has challenges including observation difficulties, infrequency of crashes and violations, and delayed feedback and consequences that are usually tied to outcomes rather than to behavior. Onboard safety monitoring systems and predictive data analytics allow to overcome these challenges (Knipling & Hyten, 2015). Through the use of telematic data, techniques for identifying high-risk drivers has advanced significantly. Instead of traditional measures of high-risk drivers such as driving records, traffic violations, and accident report, real time behavior of drivers is monitored and analyzed to provide scores of risky behaviors. As noted above, companies such as Lytx, and SmartDrive, and more recently Omnitracs, offer event-based video analytics, where safety-critical events and associated unsafe behaviors are captured on video and annotated by human reductionists. Predictive analytics provides information on risk indicators (e.g., hard braking and speeding), safety guidelines violations, frequency of risky behaviors, and distracted or fatigued driver. Also, drivers being aware that their performance is being observed and rewarded influences driver behavior in a positive way (Omnitracs, 2016). Data could be leveraged to identify driving trends, verify existing training modules, develop targeted driver training ("Bendix Commercial Vehicles Systems LLC," n.d.), automatically notify driver and manager, assign remedial fleet driver safety training ("CEI Fleet Driver Management," n.d.), and understand driver learning curves ("Lytx, Inc.," n.d.). Hickman and Hanowski (2010) evaluated the effect of the Lytx driving behavior management service. Their results showed that the two subject carriers significantly reduced the mean rate of recorded safety-related events/10,000 VMT from baseline to intervention by 38.1 and 52.3 percent.

As reviewed in further detail above, Lytx and SmartDrive today offer predictive analytics as part of their driving behavior management service. These companies both collect a large number of crashes as part of their operations, which is used to develop predictive models mapping from

driver behaviors captured in the event videos to crash risk (Lytx, 2017; SmartDrive, 2017). These models can then be used to identify unsafe drivers based on their observable behavior, as part of the behavior management services offered to fleets.

Furthermore, in the trucking industry a key current motivation for the development of models relating personal characteristics to crash risk based is the current shortage of drivers. Current legislation prevents drivers younger than 21 years from operating across states due to their known over-involvement in crashes. If younger drivers in the 18-20 age range could be identified based on predictive analytics, they may potentially be incorporated into the work force (Boris and Luciana, 2017). However, due to the over-involvement of young drivers in crashes (Massie et al., 1995; NHTSA, 2012), there is a need for a model that can help predicting the safety of young drivers based on individual characteristics. ATRI has recently initiated a project with the goal to investigate these possibilities (Boris & Luciana, 2017).

Usage-Based Insurance (UBI)

In addition to fleet management and driver selection, an important application of onboard safety monitoring systems and big data analytics is Usage-Based Insurance (UBI). Group behavior-based factors such as credit scores, gender, age, marital status that are traditionally used to score driving risk and set insurance rates could result in undercharging or overcharging customers due to generalizations (Huetter, 2017). Monitoring driving behavior allows insurers to use true causal risk factors to assess risks and develop UBI rating plans (Karapiperis et al., 2015). UBI aims at personalizing insurance rates based on how and how much customer drives. Insurance policies such as pay-as-you-drive (PAYD), pay-as-you-drive-as-you-save (PAYDAYS), and pay-how-you drive (PHYD) are some examples of UBI programs. Data for UBI programs can be collected through a dongle plugged into the vehicle, a mobile application, or directly from the car itself (Huetter, 2017). An example of UBI service provider is Progressive Snapshot program ("Progressive Casualty Insurance Company," n.d.). Progressive's UBI technology collects information on risky driving behavior such as hard braking, fast acceleration, and phone usage while driving. The application provides driving tips such as avoiding weekend night time driving and reducing mileage by carpooling to save on insurance rates.

Although the value of individual-based driving data in determining accurate premium was recognized in the early days of automobile insurance history (Dorweiler, 1929), it was not practical due to the lack of technology. With telematics being introduced to the insurance market in the late 1990s, and evolution of onboard monitoring technology, the insurance market is moving toward telematics-based UBI programs (Karapiperis et al., 2015). Soon, UBI will become the primary means of automotive underwriting (Huetter, 2017), which has encouraged research and development in this area. For example, Händel et al. (2014) presented a framework for deployment of a smartphone-based measurement system for road vehicle traffic monitoring and UBI. Soleymanian et al. (2016) examined the effect of the UBI policy on changing the

customers' driving behavior. They concluded that safer drivers and adopters of UBI have higher retention rates. They also found that after UBI adoption, the drivers improved their driving behavior, resulting in lower risk of an accident (Soleymanian et al., 2016), which is expected considering that under UBI, policyholders are incentivized to adopt risk-minimizing behaviors (Karapiperis et al., 2015).

The actual behavior analytics models used in the existing UBI products are typically not disclosed, and many UBI products are still based on exposure (i.e., mileage) rather than observed driver behavior. Thus, development of UBI services seems rather disconnected from both the fleet management applications reviewed in the previous section and the academic research on differential crash involvement reviewed in Chapter 3.

Discussion and Conclusions

As the present review shows, predicting crash involvement for individual drivers based on driver characteristics and associated behavioral manifestations is a broad topic with a long history dating back at least 70 years. Due to the lack of detailed exposure, behavior and crash data from a single source, much research in this area has traditionally relied on self-reported data (e.g., driving style questionnaires and self-reports on past crashes; see e.g., Elander et al., 1993 and Sagberg et al., 2006). Early anecdotal evidence suggested that unsafe driving styles are strongly associated with drivers' general behavioral history (Tillman and Hobbs, 1949). More recently, it has been demonstrated that individual crash involvement can been predicted based on historical crash/violation/conviction data available in government records (e.g., Lueck and Murray, 2006).

As early as in the 1960's, successful attempts were made to distinguish crash-involved and crash-free drivers based on patterns of observable driving behavior (Greenshields and Platt, 1967), although, importantly, these results were based on data collected from pre-defined routes which strongly reduced the influence of non-personal situational factors. Nevertheless, this work identified a number of promising behavioral predictors of individual crash risk such as g-forces (acceleration jerk), preferred (free) speed, close following and steering reversals. More recent work has mainly focused on behavioral measures related to high g-forces (af Wahlberg, 2006; Bagdadi & Varhelyi, 2011; Simons-Morton et al., 2012), probably because these are somewhat less influenced by situational factors than speed, headway and steering measures.

In recent years, the advent of naturalistic driving data has offered new possibilities in studying differential crash involvement. First, this data enables precise estimation of driving exposure. Second, naturalistic data contains observable behavior as well as safety critical events for the same drivers and the larger naturalistic driving datasets, in particular SHRP 2, includes a significant amount of real crashes. Somewhat surprisingly, published academic studies on differential crash involvement using naturalistic driving data are relatively rare (Simons-Morton et al., 2012 and Guo and Fang, 2013, are two notable exceptions). The large quantities of

naturalistic behavior and crash data collected through commercial operations by fleet/driver management companies such as Lytx and SmartDrive opens up further opportunities to advance the understanding of differential crash involvement. Analytics models for predicting individual crash involvement based on tens of thousands of drivers, including previously unseen quantities of recorded behaviors and crashes (on the order of tens to hundreds of thousands), are being developed in-house by both companies (SmartDrive 2017, Lytx, 2017). However, it should be noted that these models are, so far, based mainly on manually annotated behaviors rather than actual patterns of driving. It is interesting to note that no existing academic study seems to have addressed individual risk prediction based on a larger set of naturalistic crashes and actual driving performance data.

The statistical methods applied in this domain have developed over the years, and today Poisson and negative binomial regression models are established as an "industry standard" for modelling the relationship between behavior and crash/incident rates. Logistic regression is the standard model used for classification. More advanced contemporary machine learning approaches such as Deep Learning, commonly used for other types of "big data analytics" have not been extensively used for the prediction of individual crash risk but it seems likely that this will be an important direction in which the field will develop in the coming years.

In general, academic research on this topic is rather scattered and often not connected to industrial applications. For example, there are few links between applied research on predicting crash involvement from violations and convictions for commercial vehicle drivers (e.g., Murray and Lueck, 2011) and more basic research on the relationship between driving style and safety (reviewed in Sagberg et al., 2015). There are several driver risk prediction applications already in us in the fleet management and insurance domains, offered by companies such as Ominitracs, SmartDrive, Lytx, Bendix, Progressive and StateFarm. However, these applications have typically been developed through in-house proprietary R&D and thus not disclosed or validated in published research.

References

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633–642.

af Wåhlberg, A. E. (2006). Speed choice versus celeration behavior as traffic accident predictor. *Journal of Safety Research*, *37*, 43–51.

Anderson, J.E., Godava, M., Steffan, T.K., et al. (2012). Obesity is associated with the future risk of heavy truck crashes among newly recruited commercial drivers, *Accident Analysis & Prevention,* 49, 378-84.

Arbabzadeh, N., & Jafari, M. (2017). A Data-Driven Approach for Driving Safety Risk Prediction Using Driver Behavior and Roadway Information Data. *IEEE Transactions on Intelligent Transportation Systems*, *PP*(99), 1–15. https://doi.org/10.1109/TITS.2017.2700869

Ba, Y., Zhang, W., Wang, Q., Zhou, R., & Ren, C. (2017). Crash prediction with behavioral and physiological features for advanced vehicle collision avoidance system. *Transportation Research Part C: Emerging Technologies*, *74*. Retrieved from https://trid.trb.org/view.aspx?id=1440935

Bagdadi O, Varhelyi A. 2011. Jerky driving—an indicator of accident proneness? Accid Anal Prev. 2011; 43(4):1359–1363.

Ball KK, Owsley C, Stalvey B, Roenker DL, Sloane ME, Graves M. Driving avoidance and functional impairment in older adults. *Accident Analysis and Prevention*. 1998;30:313–322.

Bendix Commercial Vehicles Systems LLC. (n.d.). Retrieved July 21, 2017, from http://www.bendix.com/en/

Boris and Luciana (2017). Developing a Younger Driver Assessment Tool. Technical Memorandum #1. American Transportation Research Institute (ATRI).

CEI Fleet Driver Management. (n.d.). Retrieved July 21, 2017, from http://www.ceinetwork.com/

CEI. 2017. Predictive analytics: Shedding new light on hidden high-risk fleet drivers. CEI White Paper.

Constantinescu, Z., Marinoiu, C., & Vladoiu, M. (2010). Driving Style Analysis Using Data Mining Techniques. *International Journal of Computers Communications & Control*, *5*(5), 654–663. https://doi.org/10.15837/ijccc.2010.5.2221

Craft, R. H., & Preslopsky, B. (2009). Driver distraction and inattention in the USA large truck and national motor vehicle crash causation studies. Paper presented at the First International Conference on Driver Distraction and Inattention (28-29 September)

Dahlen, E. R., White, R. P. (2006). The Big Five Factors, Sensation Seeking, and Driving Anger In the Prediction of Unsafe Driving. *Personality and Individual Differences*, 41(5), 903-915.

Designing a Connected Vehicle Platform on Cloud IoT Core | Solutions. (n.d.). Retrieved July 3, 2017, from https://cloud.google.com/solutions/designing-connected-vehicle-platform

Dimitris Karapiperis, Birny Birnbaum, Aaron Brandenburg, Sandra Castagna, Allen Greenberg, Robin Harbage, & Anne Obersteadt. (2015). *Usage-Based Insurance and*

Vehicle Telematics: Insurance Market and Regulatory Implications. National Association of Insurance Commissioners, Center for Insurance Policy and Research.

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. https://doi.org/10.1073/pnas.1513271113

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., Bucher, C., Doerzaph, Z. R., Jermeland, J., & Knipling, R.R. 2006. *The 100-Car Naturalistic Driving Study Phase II – Results of the 100-Car Field Experiment*. HNTSA DOT, Report No: HS 810 593

Dorweiler, P. 1929. Notes on Exposure and Premium Bases, *Proceedings of the Casualty Actuarial Society*, vol. XVI, no. 33 , pp. 319-343

Duke, J., Guest, M., Boggess, M. (2010). Age-related safety in professional heavy vehicle drivers: A literature review. *Accident Analysis & Prevention*, *42*(2), 364–371.

Dunn, N., Hickman, J. and Hanowski, R. 2015. Crash Trifecta: A Complex Driving Scenario for Describing Safety-Critical Event Causation. Paper presented at the 2015 TRB Annual Meeting.

Elander, J., West, R., & French, D. (1993). Behavioral correlates of individual differences in road traffic crash risk: An examination of methods and findings. *Psychological Bulletin*, *113*, 279–294.

Evans L. and Wasielewski P. 1982. Do accident-involved drivers exhibit riskier everyday driving behaviour? Accid. Anal. & Prev. 14, 51-64.

Evans L. and Wasielewski P. 1983. Risky driving related to driver and vehicle characteristics, Accid. Anal. & Pm. 15, 121-136.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Grace, & Suizo. (2015). Using Predictive Analytics to Improve Fleet Decisions. *Automotive Fleet*. Retrieved from http://www.automotive-fleet.com/channel/gps-telematics/article/story/2015/10/using-predictive-analytics.aspx

Greenshields, B. D. Driving behaviour and related problems. *Highway Research Record,* 1963, 25, 14-32.

Greenshields. B. D. and Platt. F. N. Development of a method of predicting high accident and high violation drivers. *Journal* of *Applied Psychology,* 1967, 51, 205-210.

Greenwood M. and Woods H. M., A report on the incidence of industrial accidents upon individuals with special reference to multiple accidents, 1919. Reproduced in W. Haddon, E. A. Suchman and D. Klein (Eds.), *Accident Research.* Harper & Row, New York, 1964.

Gregoriades, A., & Mouskos, K. C. (2013). Black spots identification through a Bayesian Networks quantification of accident risk index. *Transportation Research Part C: Emerging Technologies*, *28*, 28–43. https://doi.org/10.1016/j.trc.2012.12.008

Guo, F., & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, *61*, 3–9. https://doi.org/10.1016/j.aap.2012.06.014

Guo, F., Fang, Y., & Antin, J. F. (2015). Older driver fitness-to-drive evaluation using naturalistic driving data. *Journal of Safety Research*, *54*, 49.e29-54. https://doi.org/10.1016/j.jsr.2015.06.013

Händel, P., Ohlsson, J., Ohlsson, M., Skog, I., & Nygren, E. (2014). Smartphone-Based Measurement Systems for Road Vehicle Traffic Monitoring and Usage-Based

Insurance. *IEEE Systems Journal*, *8*(4), 1238–1248. https://doi.org/10.1109/JSYST.2013.2292721

Hanowski, R. J., Wierwille, W. W., Garness, S. A., and Dingus, T. A. 2000. *Impact of Local/Short Haul Operations on Driver Fatigue.* Final Report No. DOT-MC-00-203. Washington, DC, U.S. Department of Transportation, Federal Motor Carriers Safety Administration, September.

Hickman, J., & Hanowski, R. (2010). *Evaluating the Safety Benefits of a Low-Cost Driving Behavior Management System in Commercial Vehicle Operations*. Washington, DC: Federal Motor Carrier Safety Administration.

Howard, M. E., Desai, A. V., Grunstein, R. R., Hukins, C., Armstrong, J. G., Joffe, D., et al. (2004). Sleepiness, sleep-disordered breathing, and accident risk factors in commercial vehicle drivers. American Journal of Respiratory and Critical Care Medicine, 170(9), 1014-1021. doi: 10.1164/rccm.200312-1782OC

Hu, M., Liao, Y., Wang, W., Li, G., Cheng, B., & Chen, F. (2017). Decision Tree-Based Maneuver Prediction for Driver Rear-End Risk-Avoidance Behaviors in Cut-In Scenarios [Research article]. https://doi.org/10.1155/2017/7170358

Huetter, J. (2017). Progressive: Usage-based insurance is the future, likely assisted by OEM data. Retrieved from http://www.repairerdrivennews.com/2017/05/17/progressive-usage-based-insurance-is-the-future-likely-assisted-by-oem-data/

Jonah, B.A. (1997). Sensation seeking and risky driving: a review and synthesis of the literature. *Accident Analysis and Prevention, 29*(5), 651-665.

Knipling, R., & Hyten, C. (2015). *Commercial Vehicle Safety: Onboard Safety Monitoring as Part of Behavioral Safety Management*. AUBREY DANIELS INTERNATIONAL.

Knipling, R.R. 2009. Safety for the Long Haul; Large Truck Crash Risk, Causation, & Prevention.

Knipling, R.R., Hickman, J.S., and Bergoffen, G. 2004. Synthesis 1: Effective Commercial

Knipling, Ron, Erik C. B. Olsen, and Tammy D. Prailey. 2004. “Individual Differences and the

Lantz, Brenda M. and Michael W. Blevins. “An Analysis of Commercial Vehicle Driver

Lantz, Brenda, M., Jeff Loftus, and Tom Keane. “Development and Implementation of a

Lueck, M.D. and Murray, D.C. 2011. Predicting Truck Crash Involvement: A 2011 Update. American Transportation Research Institute, Arlington, VA.

Lytx, Inc. (n.d.). Retrieved July 21, 2017, from https://www.lytx.com/en-us/

Lytx. 2017. Industry insights: Beyond telematics: How video predicts risky behavior. Lytx White Paper.

Massie, D.L., Campbell, K.L., & Williams, A.F. (1995). Traffic accident involvement rates by driver age and gender. *Accident Analysis and Prevention, 27*(1), 73-87.

McCartt, A.T., Mayhew, D.R., Braitman, K.A., Ferguson, S.A., Simpson, H.M., (2009). Effects of age and experience on young driver crashes: review of recent literature. *Traffic Inj. Prev*. 10 (3), 209–219.

McKenna, F. P. (1983). Accident proneness: A conceptual analysis. *Accident Analysis and Prevention, 15,* 65-71.

Murray, D. C., Lantz, B., & Keppler, S. A. (2005). Predicting Truck Crash Involvement: Developing a Commercial Driver Behavior-Based Model and Recommended Countermeasures. Retrieved from https://trid.trb.org/view.aspx?id=771443

Murray, D., Turrentine, S., Lantz, B., Keppler, S., 2006. Predicting truck crash involvement: Developing a commercial driver behavior model and requisite enforcement countermeasures. Transportation Research Board 2006 Annual Meeting.

National Highway Traffic Safety Administration (2012): Traffic Safety Facts 2012 A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. Washington, D.C.: National Center for Statistics & Analysis.

National Highway Traffic Safety Administration (NHTSA). (2015). *Traffic Safety Facts*. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812203

Omnitracs. (2016). *White Paper: Critical Event Reporting + Omnitracs Analytics*. Dallas, Texas: Omnitracs, LLC.

Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D., & Summala, H. (2006). Cross-cultural differences in driving behaviours: A comparison of six countries. *Transportation Research Part F: Traffic Psychology and Behaviour*, *9*(3), 227–242.

Paul Dorweiler. (1929). Notes on Exposure and Premium Bases. Presented at the Proceedings of the Casualty Actuarial Society XVI, p. 319; reprinted PCAS LVIII, 1972, p. 59.

Reason, J. 1990. Human Error. Cambridge University Press

Roman, G. D., Poulter, D., Barker, E., McKenna, F. P., & Rowe, R. (2015). Novice drivers' individual trajectories of driver behavior over the first three years of driving. *Accident Analysis & Prevention*, *82*, 61–69. https://doi.org/10.1016/j.aap.2015.05.012

Sagberg, F. (2006). Driver health and crash involvement: a case-control study. *Accident Analysis and Prevention, 38*(1), 28-34.

Sagberg, F., Selpi, Piccinini, G. F. & Engström, J. 2015. A Review of Research on Driving Styles and Road Safety. *Human Factors* 57(7):1248-75.

Schwebel, D.C., Severson, J., Ball, K.K., Rizzo, M. (2006). Individual difference factors in risky driving: The roles of anger/hostility, conscientiousness, and sensation-seeking. *Accident Analysis and Prevention*. 2006 (38). 801–810

Segal, J.L. and Habinski, A. (2006). What we know about ADHD and driving risk: A literature review, meta-analysis and critique. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 15*(3), 105–125.

Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention*, *27*(3), 371–389. https://doi.org/10.1016/0001-4575(94)00078-Z

Simons-Morton B, Zhang Z, Jackson JC, Albert PS. 2012. Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes? Am J Epidemiol. 175(10):1075–1079.

Simons-Morton, B. G., Cheon, K., Guo, F. and Albert, P. 2013.Trajectories of Kinematic Risky Driving Among Novice Teenagers. Accid Anal Prev. March ; 51: 27–32.

SmartDrive. (2017). *Measuring driver risk with video-based analytics.* San Diego, California: SmartDrive Systems.

Soccolich, S. A., Hickman, J. S., & Hanowski, R. J. (2011). Identifying high-risk commercial truck drivers using a naturalistic approach. Retrieved from https://vtechworks.lib.vt.edu/handle/10919/23321

Soleymanian, M., Weinberg, C., & Zhu, T. (2016). The Value of Usage-Based Insurance beyond Better Targeting: Better Driving. Retrieved from https://research.chicagobooth.edu/~/media/8aeee2a8af83412c954d331dc412cc55.pdf

Tillmann, W. A., & Hobbs, G. E. (1949). The accident-prone automobile driver: A study of the psychiatric and social background. *American Journal of Psychiatry*, *106*, 321–331.

Treat, J. R., Tumbas, N. S., McDonald, S. T., Shinar, D., Hume, R. D., Mayer, R. E., Stanisfer, R. L. & Castellan, N. J. 1977. *Tri-level study of the causes of traffic accidents*. Report No. DOT-HS-034-3-535-77 (TAC).

Ulleberg, P., Rundmo, T., 2003. Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. Safety Science 41 (5), 427–443.

Wang, J., Zheng, Y., Li, X., Yu, C., Kodaka, K., & Li, K. (2015). Driving risk assessment using near-crash database through data mining of tree-based model. *Accident Analysis & Prevention*, *84*, 54–64. https://doi.org/10.1016/j.aap.2015.07.007

Wasielewski, P. 1984. Speed as a measure of driver risk: Observed speed versus driver and vehicle characteristics. Accident Analysis and Prevention, 16, 89-102.

West, R., Elander, J., & French, D. (1992a). Decision making, personality and driving style as correlates of individual crash risk (Contractor's Report No. 309). Crowthorne, England: Transport and Road Research Laboratory.

West, R., French, D., Kemp, R., & Elander, J. (1993). Direct observation of driving, self-reports of driver behavior, and accident involvement. *Ergonomics*, 36, 557–567.

Wiegand, D. M., Hanowski, R. J., McDonald, S. E. (2009). Commercial drivers' health: A naturalistic study of body mass index, fatigue, and involvement in safety-critical events. *Traffic Injury Prevention*, *10*(6), 573-579.

Wilson, R., & Greensmith, J. (1983). Multivariate analysis of the relationship between drivometer variables and drivers' accident, sex and exposure status. *Human Factors, 25,* 303-312.

Yin, S., Duan, J., Ouyang, P., Liu, L., & Wei, S. (2017). Multi-CNN and Decision Tree Based Driving Behavior Evaluation. In *Proceedings of the Symposium on Applied Computing* (pp. 1424–1429). New York, NY, USA: ACM. https://doi.org/10.1145/3019612.3019649

Zhu, X., Yuan, Y., Hu, X., Chiu, Y. C., & Ma, Y. L. (2017). A Bayesian Network model for contextual versus non-contextual driving behavior assessment. *Transportation Research Part C: Emerging Technologies*, *81*, 172–187. https://doi.org/10.1016/j.trc.2017.05.015

Appendix B. Publication Manuscripts

The relationship between the Driver Behavior Questionnaire, Sensation Seeking Scale, and recorded crashes: A brief comment on Martinussen et al. (2017) and new data from SHRP 2

De Winter, J. C. F, Dreger, F. A., Huang, W., Miller, A., Soccolich, S., Ghanipoor Machiani, S., & Engstrom, J. (2018). The relationship between the Driver Behavior Questionnaire, Sensation Seeking Scale, and recorded crashes: A brief comment on Martinussen et al. (2017) and new data from SHRP 2. *Accident Analysis and Prevention, 118,* 54-56.

J. C. F. de Winter[a], F. A. Dreger[b], W. Huang[c], A. Miller[c], S. Soccolich[c], S. Ghanipoor Machiani[d], J. Engström[c]

[a]Department of BioMechanical Engineering, Delft University of Technology, The Netherlands
[b]Department of Cognitive Robotics, Delft University of Technology, The Netherlands
[c]Virginia Tech Transportation Institute, U.S.A.
[d]San Diego State University, U.S.A.

The recently published paper by Martinussen et al. (2017) is a unique large-sample study (N = 3,683) on the relationship between the Driver Behavior Questionnaire (DBQ) and recorded violations and crashes.

There are two important findings. First, the authors found that 22.4% of participants who were classified into the 'violating unsafe drivers' group (based on a cluster analysis of self-reported answers to the DBQ and Driver Skill Inventory, DSI) were involved in a recorded traffic law offence. This percentage is 2.8 times as high as the average of the other three groups ('skilled safe drivers, 'unskilled safe drivers', and 'low confidence safe drivers'). This finding is consistent with a meta-analysis which showed that a moderate correlation (r = 0.24) exists between the DBQ violations score and recorded measures of speed/speeding (De Winter et al., 2015).

Second, the authors found that the four groups did not differ in recorded crash rates. It is important to emphasize, however, that only 1.1% of the participants were involved in a crash (despite the 6-year recording period). This low percentage means that the 'violating unsafe drivers' group contained only 6 or 7 crash-involved drivers (estimated from sample sizes reported in Martinussen et al., 2014). Considering that traffic violations correlate with crashes (Cooper, 1997; Factor, 2014) and young males are overinvolved in crashes (OECD, 2006), it would be inappropriate for one to conclude from their data that the 'violating unsafe drivers' group (consisting of 74% males with a mean age of 39 years) is equally safe as the other three groups (consisting overall of 47% males with a mean age of 54 years). With simulations, De Winter et al. (2015) showed that if crash rates are low, then correlations with crash involvement are necessarily small (see also Af Wåhlberg & Dorn, 2009).

Here, we report on DBQ-crash correlations in a newly accessed dataset from the Strategic Highway Research Program (SHRP 2) naturalistic driving study (Dingus et al., 2015). The dataset comprised 3,215 drivers. We removed drivers with less than 7 months of participation and drivers who drove less than 100 miles, leaving data for 2,790 drivers. The mean study length across drivers was 1.31 years (*SD* = 0.51 years). In case no more than two DBQ items were missing for a driver, then the scores for these items were replaced with the value from the single 'nearest neighbor' variable (1NN); otherwise, the DBQ data for that driver were discarded. Accordingly, DBQ data were available for 2,737 drivers. Participants' scores for the Sensation Seeking Scale Form V (SSS) were retrieved as well (*N* = 2,781). Whether the DBQ and SSS correlate with recorded crashes has been a much-debated topic (e.g., Af Wåhlberg, 2010; De Winter et al., 2015).

First, we applied principal component analysis on the 24-item DBQ. Inspection of the scree plot (see supplementary material, Figure S1) suggested that a three-component solution was appropriate. The three components were obliquely rotated (Promax) and interpreted as (1) slips, (2) violations, and (3) lapses (see Table S1 for loadings). Component scores were calculated using the regression method. Next, Spearman rank-order correlations were computed between the self-report scores (DBQ scores and SSS score) on the one hand, and relevant study variables (age, gender, crash involvement, driving style) on the other (Table 1).

The results in Table 1 confirm the well-known phenomenon that older drivers report fewer violations than younger drivers and that females report fewer violations but more errors than males. It is also found that DBQ errors and DBQ violations correlate with self-reported crashes in the past three years, and with objective crashes and near-crashes during the naturalistic driving study period. These correlations were overall small yet mostly statistically significant. The correlations were stronger for DBQ violations and SSS than for DBQ slips and lapses. For crashes of the highest severity level (airbag, injury, rollover), the correlation with DBQ violations was small ($\rho = 0.02$). Only 3% of drivers were involved in this type of crash. For all crashes, the correlation with DBQ violations was somewhat stronger ($\rho = 0.06$), and for near-crashes, the correlation with DBQ violations was moderate ($\rho = 0.20$). These findings support the previous assertion that correlations are smaller if the mean (and therefore the variance) of the number of crashes is higher (Af Wåhlberg & Dorn, 2009; De Winter et al., 2015).

Table 1 also shows that the DBQ violations score was associated with a more adverse driving style (hard starts, stops, and turns), with correlations between 0.04 and 0.24. Finally, it can be observed that the pattern of correlations for the SSS was similar to that for DBQ violations (Table 1; Figure S2). This is also reflected in the fact that the DBQ violations score was associated with the SSS score ($\rho = 0.36$), whereas the correlations between the SSS and DBQ slips and DBQ lapses were smaller ($\rho = 0.09$ and $\rho = 0.10$, respectively).

Table 1. Spearman rank-order correlations between Driver Behavior Questionnaire (DBQ) scores, Sensation Seeking Scale (SSS) scores, and study variables

Study variable	*M*	*SD*	ρ DBQ slips	ρ DBQ violations	ρ DBQ lapses	ρ SSS
Age group (1 = 16–19 years, 17 = 95–99 years)	6.1113	4.679	0.00	-0.33*	-0.10*	-0.43*
Gender (0 = male, 1 = female)	0.5219	0.4996	0.07*	-0.06*	0.19*	-0.15*
Distance driven in study period (miles)	10371.74	7283.22	0.04	0.18*	0.03	0.11*
Number of self-reported crashes in past 3 years (0, 1, 2+)	0.319	0.5815	0.10*	0.13*	0.09*	0.10*
Number of recorded crashes in study period	0.605	1.1488	0.04* (0.04)	0.06* (0.04*)	0.05* (0.05*)	0.10* (0.09)
Number of recorded near-crashes in study period	2.1846	3.35	0.04* (0.03)	0.20* (0.15*)	0.03 (0.02)	0.20* (0.18*)
Number of recorded at-fault crashes in study period	0.4989	1.0664	0.05* (0.04*)	0.05* (0.03)	0.05* (0.05*)	0.10* (0.09*)
Number of recorded at-fault near-crashes in study period	1.2885	2.3802	0.05* (0.04)	0.18* (0.15*)	0.02 (0.02)	0.20* (0.18*)
Number of recorded severity 1 crashes in study period	0.0333	0.1835	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)	0.02 (0.02)
Number of recorded severity 2 crashes in study period	0.0656	0.2711	0.02 (0.02)	0.06* (0.06*)	-0.01 (- 0.01)	0.05* (0.05*)
Number of hard starts per mile in a 6-month period	0.0458	0.0786	-0.02	0.07*	-0.01	0.10*
Number of hard stops per mile in a 6-month period	0.1312	0.1384	0.03	0.04*	0.04*	0.08*
Number of hard left turns per mile in a 6-month period	0.1665	0.1457	-0.03	0.21*	0.01	0.27*
Number of hard right turns per mile in a 6-month period	0.1629	0.1341	-0.03	0.24*	0.02	0.27*

Note. * $p < .05$. Correlations for the number of crashes per mile are reported in parentheses. Severity 1 crashes are defined as airbag/injury/rollover, high delta-V crashes (virtually all would be police reported). Severity 2 crashes are defined as police-reportable crashes (including police-reported crashes, as well as others of similar severity which were not reported) (Dingus et al., 2015). Hard starts, stops, and turns are defined as incidences where the acceleration exceeded 0.30 g (Jun et al., 2007). The sample sizes per cell are reported in the supplementary materials (Table S2).

Finally, although many of the correlations shown in Table 1 are statistically significant and theoretically interesting, we wish to caution that they are not necessarily practically significant. A boxplot of the SSS scores for non-crash-involved drivers and crash-involved drivers (Fig. 1, top) shows that there is a high degree of overlap of the SSS distributions of both groups, even though the difference was strongly significant, $t(2779) = 5.55$, $p = 3.07*10^{-8}$, Cohen's $d = 0.22$. For near-

crashes, the effect was somewhat stronger (Fig. 1, bottom), $t(2779) = 8.77$, $p = 2.97*10^{-18}$, Cohen's $d = 0.35$.

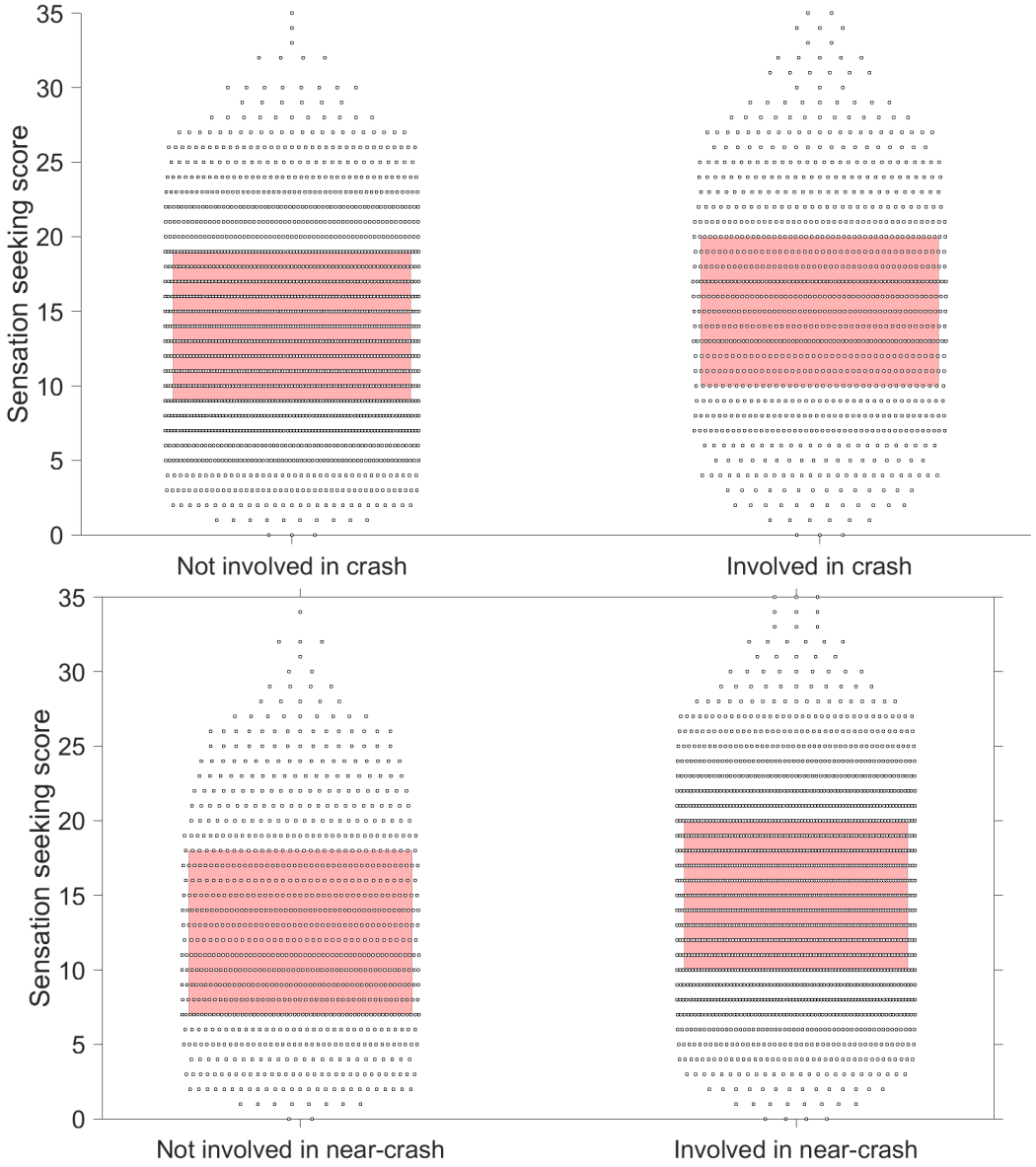

Figure 1. Top: Sensation Seeking Scale scores for drivers who are not involved in a crash (N = 1,766) and drivers who are involved in a crash (N = 1,015). Bottom: Sensation Seeking Scale scores for drivers who are not involved in a near-crash (N = 917) and drivers who are involved in a near-crash (N = 1,864). The red box shows the 25th and 75th percentiles, respectively. The markers represent the individual drives.

In conclusion, we support the findings and interpretations by Martinussen et al. (2017) and hope that the above points are a useful addendum. It appears that DBQ violations, as well as the SSS, exhibit small associations with crash involvement, and small to moderate associations with near-crash involvement and driving style. The predictive validity of DBQ errors (slips and lapses) appears to be weak. Future research should examine the validity of near-crashes as a proxy for crashes.

Acknowledgment

The authors would like to thank the Safe-D UTC program for the support regarding the preparation of the dataset.

References

Af Wåhlberg, A., & Dorn, L. (2009). Bus driver accident record: the return of accident proneness. *Theoretical Issues in Ergonomics Science, 10*, 77–91.

Af Wåhlberg, A. (2010). Social desirability effects in driver behavior inventories. *Journal of Safety Research, 41*, 99–106.

Cooper, P. J. (1997). The relationship between speeding behaviour (as measured by violation convictions) and crash involvement. *Journal of Safety Research*, *28*, 83–95. http://dx.doi.org/10.1016/S0022-4375(96)00040-0

De Winter, J. C. F., Dodou, D., & Stanton, N. A. (2015). A quarter of a century of the DBQ: Some supplementary notes on its validity with regard to accidents. *Ergonomics, 58*, 1745–1769. http://dx.doi.org/10.1080/00140139.2015.1030460

Dingus, T. A., Hankey, J. M., Antin, J. F., Lee, S. E., Eichelberger, L., Stulce, K. E., ... & Stowe, L. (2015). Naturalistic driving study: Technical coordination and quality control (No. SHRP 2 Report S2-S06-RW-1). Retrieved from http://onlinepubs.trb.org/onlinepubs/SHRP 2/SHRP 2_S06Report.pdf

Factor, R. (2014). The effect of traffic tickets on road traffic crashes. *Accident Analysis & Prevention*, *64*, 86–91. http://dx.doi.org/10.1016/j.aap.2013.11.010

Jun, J., Ogle, J., & Guensler, R. (2007). Relationships between crash involvement and temporal-spatial driving behavior activity patterns using GPS instrumented vehicle data. 86th Annual Meeting of the Transportation Research Board, Washington, DC. https://doi.org/10.3141/2019-29

Martinussen, L. M., Møller, M., & Prato, C. G. (2014). Assessing the relationship between the Driver Behavior Questionnaire and the Driver Skill Inventory: Revealing sub-groups of drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *26*, 82–91. http://dx.doi.org/10.1016/j.trf.2014.06.008

Martinussen, L. M., Møller, M., Prato, C. G., & Haustein, S. (2017). How indicative is a self-reported driving behaviour profile of police-registered traffic law offences? *Accident Analysis and Prevention, 99*, 1–5. http://dx.doi.org/10.1016/j.aap.2016.10.031

Organisation for Economic Co-operation and Development (OECD) (2006). *Young drivers: The road to safety*. Paris, France: OECD.

Supplementary material

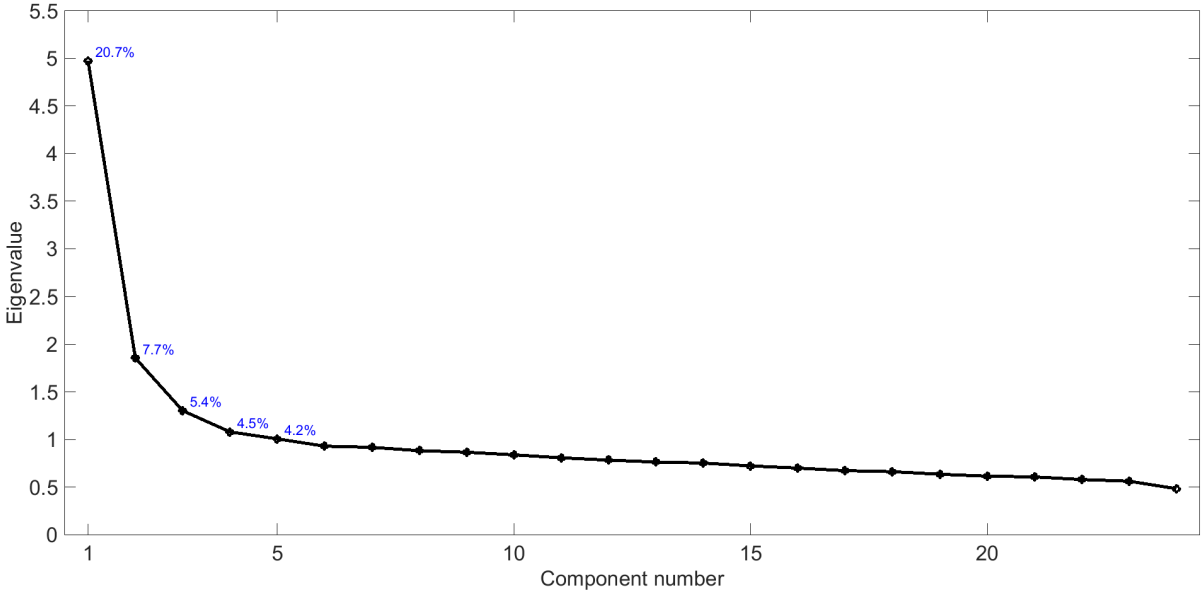

Figure S1. Eigenvalues of the correlation matrix of the 24 items of the Driver Behavior Questionnaire (DBQ), sorted in descending order ('scree plot'). Also shown are the percentages of variance explained (being proportional to the eigenvalue) for the first five components prior to rotation.

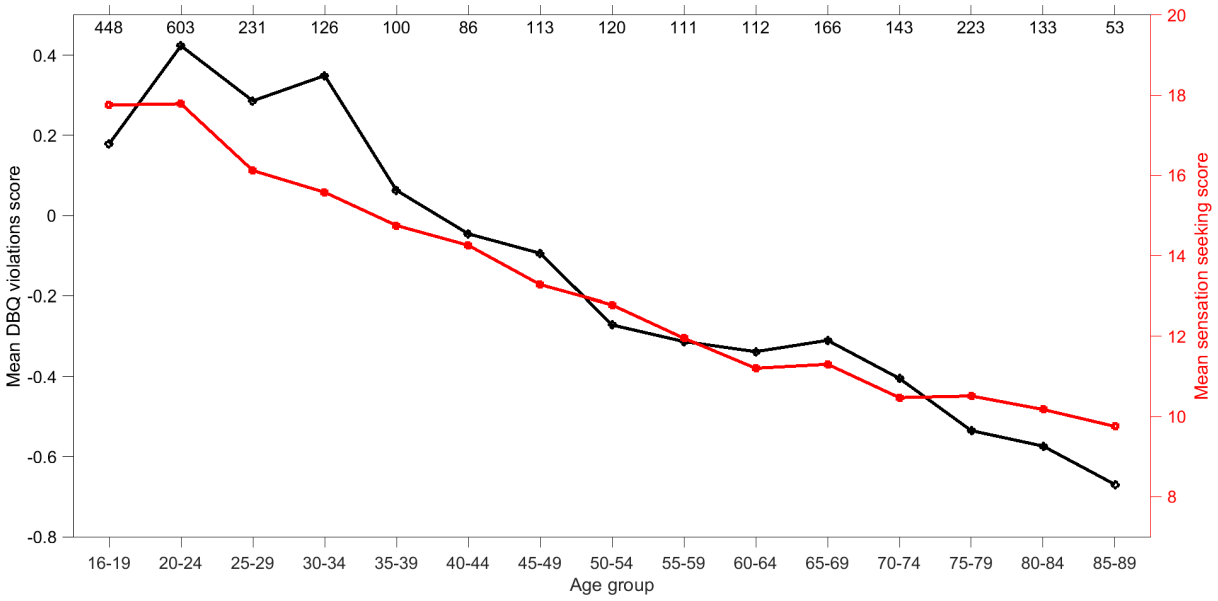

Figure S2. Mean Driver Behavior Questionnaire (DBQ) violations score and mean Sensation Seeking Scale score per age group. The sample sizes per age group are shown at the top of the figure. Results for 90–94 years and 95–99 years are not shown because of small sample size (n = 6 and 2, respectively). The age group was not known for 14 of 2,790 drivers.

Table S1. Principal component loadings of 24 Driver Behavior Questionnaire (DBQ) items, for the first three components after Promax rotation.

#	DBQ item	DBQ slips	DBQ violations	DBQ lapses
1	Attempt to drive away from traffic lights in the wrong gear	0.05	0.11	0.18
2	Become impatient with a slow driver in the fast lane and pass on the right	-0.25	0.74	0.21
3	Drive especially close to a car in front as a signal to the driver to go faster or get out of the way	-0.20	0.79	0.11
4	Attempt to pass someone that you hadn't noticed to be making a left turn	0.15	0.43	0.09
5	Forget where you left your car in a parking lot	-0.01	-0.04	0.65
6	Turn on one thing, such as your headlights, when you mean to switch on something else, such as the windshield wipers	0.26	-0.10	0.45
7	Realize that you have no clear recollection of the road along which you have just been traveling	-0.15	0.22	0.66
8	Cross an intersection knowing that the traffic lights have already changed from yellow to red	0.05	0.39	0.28
9	Fail to notice that pedestrians are crossing when turning onto a side street from a main road	0.49	0.03	0.18
10	Angered by another driver's behavior, you catch up to them with the intention of giving him/her "a piece of your mind."	0.27	0.51	-0.26
11	Misread the signs and turn the wrong direction on a one-way street	0.51	-0.16	0.19
12	Disregard the speed limits late at night or early in the morning	-0.09	0.62	0.19
13	When turning right, nearly hit a bicyclist who is riding along side of you	0.72	-0.06	-0.30
14	Attempting to turn onto a main road, you pay such close attention to traffic on the road you are entering that you nearly hit the car in front of you that is also waiting to turn.	0.48	0.03	0.12
15	Drive even though you realize you might be over the legal blood alcohol limit	0.01	0.36	-0.03
16	Have an aversion to a particular class of road user, and indicate your hostility by whatever means you can	0.37	0.38	-0.33
17	Underestimate the speed of an oncoming vehicle when attempting to pass a vehicle in your own lane	0.49	0.08	0.07
18	Hit something when backing up that you had not previously seen	0.52	-0.18	0.09
19	Intending to drive to destination A, you 'wake up' to find yourself on a road to destination B, perhaps because destination B is a more common destination.	0.06	0.05	0.57
20	Get into the wrong lane approaching an intersection	0.25	-0.06	0.42
21	Miss "Yield" signs, and narrowly avoid colliding with traffic having the right of way	0.73	-0.12	-0.01
22	Fail to check your rearview mirror before pulling out, changing lanes, etc.	0.31	0.11	0.17
23	Get involved in unofficial 'races' with other drivers	0.10	0.49	-0.15
24	Brake to quickly on a slippery road or steer the wrong way into a skid	0.45	0.01	0.12

Table S2. Sample sizes for each of the variables, and for each pair of variables.

Study variable		DBQ slips	DBQ violations	DBQ lapses	SSS
	N	**2,737**	**2,737**	**2,737**	**2,781**
Age group (1 = 16–19 years, 17 = 95–99 years)	**2,776**	2,723	2,723	2,723	2,767
Gender (0 = male, 1 = female)	**2,790**	2,737	2,737	2,737	2,781
Distance driven in study period (miles)	**2,790**	2,737	2,737	2,737	2,781

Study variable		DBQ slips	DBQ violations	DBQ lapses	SSS
Number of self-reported crashes in past 3 years (0, 1, 2+)	**2,781**	2,731	2,731	2,731	2,772
Number of recorded crashes in study period	**2,790**	2,737	2,737	2,737	2,781
Number of recorded near-crashes in study period	**2,790**	2,737	2,737	2,737	2,781
Number of recorded at-fault crashes in study period	**2,790**	2,737	2,737	2,737	2,781
Number of recorded at-fault near-crashes in study period	**2,790**	2,737	2,737	2,737	2,781
Number of recorded severity 1 crashes in study period	**2,790**	2,737	2,737	2,737	2,781
Number of recorded severity 2 crashes in study period	**2,790**	2,737	2,737	2,737	2,781
Number of hard starts per mile in a 6-month period	**2,779**	2,726	2,726	2,726	2,770
Number of hard stops per mile in a 6-month period	**2,779**	2,726	2,726	2,726	2,770
Number of hard left turns per mile in a 6-month period	**2,779**	2,726	2,726	2,726	2,770
Number of hard right turns per mile in a 6-month period	**2,779**	2,726	2,726	2,726	2,770

Note. DBQ = Driver Behavior Questionnaire, SSS = Sensation Seeking Scale.

Modeling differential crash involvement based on SHRP 2 naturalistic driving data

Huang, W., Engstrom, J., Miller, A., Jahangairi, A., Ghanipoor Machiani, S., Dreger, F. A., Soccolich, S., & de Winter, J. C. F. (2018). Modeling Differential Crash Involvement Based on SHRP 2 Naturalistic Driving Data. *Accident Analysis and Prevention.* Manuscript submitted for publication.

Wenyan Huang[a], Johan Engström[a], Andrew Miller[a], Arash Jahangiri[b], Sahar Ghanipoor Machiani[b], Felix Dreger[c], Susan Soccolich[a]

[a]Virginia Tech Transportation Institute, USA

[b]San Diego State University, USA

[c]Department of Cognitive Robotics, Delft University of Technology, The Netherlands

Abstract

It is well established that some drivers are more likely to become involved in crashes than others, a phenomenon known as differential crash involvement. The objective of the present analysis was to investigate, using the SHRP 2 dataset, to what extent it is possible to predict crash and/or near crash involvement for individual drivers based on enduring personal factors related to demographics, driving history, personality, and observed driving style. Two types of classification models, logistic regression and random forest, yielding similar results, were able to correctly identify 72–75% of the CNC-involved drivers (recall) and of those drivers predicted by the models to be involved in a CNC during the study period, 64–65% were correct predictions (precision). Therefore, the present results show that it is possible to predict CNC (mainly near-crash) involvement for individual drivers with some degree of accuracy, while this seems to be more difficult for crashes alone.

Keywords:

Differential crash involvement; Naturalistic driving study; Enduring personal factors;

1. Introduction

It is well established that individual drivers are differentially involved in crashes and that these individual differences are at least partly associated with enduring personal factors such as demographics, health, personality, and acquired skills (de Winter et al., 2018; Guo and Fang, 2013; Hanowski et al., 2000; Huang et al., 2018; Knipling, 2009, 2004; McKenna, 1983; Simons-Morton et al., 2012; Soccolich et al., 2011). Whether drivers' crash involvement can be reliably predicted from such enduring personal factors is an issue of great interest for vehicle fleet management, insurance industries, driver education, and enforcement.

Indeed, numerous studies have found a predictive relationship between various specific variables reflecting enduring individual factors and crash involvement. This includes demographics (e.g., age; McCartt et al., 2009), personality (e.g., Boris and Luciana, 2017; Dahlen and White, 2006), behavioral history such as traffic violations and convictions (e.g., Lueck and Murray, 2011; Murray et al., 2006) and recorded driving style (see review in Sagberg et al., 2015); see Engström et al. (2017) for a more detailed review of these and related studies.

Traditionally, crash involvement data in studies on individual crash involvement are based on self-reports (e.g., West et al., 1993) or crash databases that can be linked back to individual drivers (e.g., Lueck and Murray, 2011; Murray et al., 2006). As discussed in Knipling (2009), a general issue with traditional studies on individual differential crash involvement is that driving exposure is typically unavailable (although some studies have estimated exposure from self-reports; see e.g. West et al., 1993). It is thus possible that the observed relationships between individual characteristics (in particular related to driving style behavioral history) and crash risk are at least partially confounded by driving exposure. For example, the drivers who drove the most may have had the most crashes as well as the highest number of unsafe driving events, traffic violations, and convictions.

The advent of large sets of naturalistic driving data (e.g., Hankey et al., 2016) offers exciting new possibilities in creating predictive models which map individual behavioral patterns to crash involvement. The key advantage of using naturalistic driving data to study differential crash involvement is that it typically includes study participant demographic, behavioral history, and personality screening data, observed/recorded behaviors, crash and near-crash involvement as well as driving exposure in a single dataset. Moreover, the behavioral data needed for predictive models based on driving style can be collected in weeks as opposed to the months or years it takes for behavioral history data (such as convictions and violations) to accumulate (Guo and Fang, 2013; SmartDrive, 2017).

Studies of differential crash involvement using naturalistic driving data are so far relatively sparse. One of the earliest published application of naturalistic driving data to crash/near crash (CNC) involvement prediction at the individual level is the study by Simons-Morton et al. (2012), who analyzed the relationship between gravitational-force (g-force) events and at fault CNC rate for 42 newly licensed teenage drivers. Data for each driver was collected for 18 months. Five different types of g-force events were calculated: rapid starts (> 0.35 g), hard stops (< -0.45 g), hard left/right turns (< -0.05 g or > 0.05 g), and fast yaw rate (>2 degrees per second). These were combined into a composite metric used in the statistical modelling. The results showed a strongly significant Spearman correlation of 0.6 between the composite metric of g-force events and CNC rates during the 18-month period. Furthermore, a logistic regression model showed that the rate of g-force events in the prior month predicted CNC involvement in the following month with a 76% prediction accuracy. A further analysis reported in Simons-Morton et al. (2013) showed that the g-force rates for individual drivers during the first 6 months were similar to individual rates during the next two 6 month periods, demonstrating consistency in individual risky driving.

Guo and Fang (2013) conducted a similar analysis using the 100-car naturalistic driving study data (Dingus et al., 2006) and found that age, personality, and the rate of critical-incidence events (CIEs) were predictive of CNC involvement. CIEs were not directly measured but coded in the data based on: Identification of events (1) where the car sensors exceeded a specified value

(e.g., brake response greater than a set threshold g-force value, (2) when the driver pressed an incident push-button and (3) through human judgments when reviewing the video. A logistic regression model based on the three predictors successfully classified drivers into two risk categories defined by clustering the rates of CNC involvement.

While these naturalistic driving studies have significantly advanced the state of the art in predictive modelling of individual risk, a key limitation is that the CNCs that the models were developed to predict are dominated by low-severity crashes and near-crashes. Today, the The Second Strategic Highway Research Partnership (SHRP 2) Naturalistic Driving Study, with more than 3500 vehicles, provides a much larger dataset with a significant number of crashes (about 2000), many of which are relatively severe. However, to the knowledge of the authors, no published studies have so far analyzed the SHRP 2 dataset with respect to individual driver risk.

In recent analyses using the SHRP 2 data, we have established that there is an association, albeit a relatively weak one, between individual driver characteristics and CNC involvement (de Winter et al., 2018; Huang et al., 2018). The goal of the present analysis was to investigate to what extent it is possible to classify drivers into high/low-risk categories (defined by actual crash and near-crash involvement) based on demographics/personality screening data and basic driving style indicators similar to those employed by Simons-Morton et al. (2012) and typically used in current usage-based insurance (UBI) applications.

2. Method

2.1. The SHRP 2 data set

The SHRP 2 NDS collected data for over more than 30 million vehicle miles traveled, and recorded over 2,000 crashes and 7,000 near-crashes. The naturalistic driving data were collected automatically from key-on to key-off for every trip taken in one of the volunteer participants' vehicles. The SHRP 2 NDS included over 3,500 participants, aged 16 to 98, who resided near the following six site centers: Buffalo, NY; Tampa, FL; Seattle; Durham, NC; Bloomington, IN; and State College, PA. The expected duration of participation per driver was 12 months, but not all drivers completed the full study period. More information about the SHRP 2 NDS is available in Dingus et al. (2015) and Hankey et al. (2016).

2.2. Study design

For the present analysis, a subset of the SHRP 2 data was used for each individual driver, consisting of six consecutive months beginning from the second month of data collection, that is, months 2-7, referred to in the following as the study period (See Figure 1). For each participant, questionnaire data, collected prior to the start of the data collection, was retrieved from the SHRP 2 database, and both driving style measures from time series data and CNC involvements were calculated for the study period (see details in section 2.4).

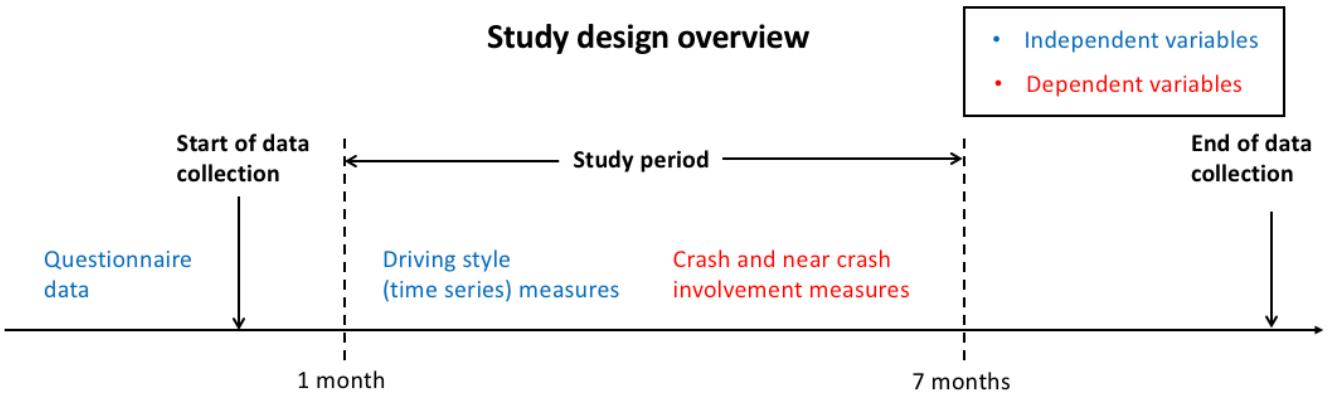

Figure 1. Overview of the study design with independent variables (blue) and dependent variables (red).

2.3. Data selection and inclusion criteria

Drivers included in this study had seven months of SHRP 2 data collection participation, at least 1,000 driving miles in the 2 to 7 month period, and complete questionnaire data. Trips selected for each driver had more than 10 seconds of moving time, a non-zero driving distance, and at least partial date and time information to be included in the analysis. These screening criteria applied in a step-by-step fashion resulted in a dataset of 2,458 drivers (see Table 1) across 3.91 million trips, amounting to a total of 27.16 million miles driving distance and 0.69 million driving hours.

Table 1. Criteria for step-by-step removal of drivers from the dataset used for analysis

Criterion	Number of drivers removed
1. The driver had less than 7 months participation	402
2. All trips for the driver were either less than 10 seconds moving time, had zero distance or lacked date and time information	13
4. The driver recorded less than 1000 miles distance in the study period	260
5. The driver had incomplete questionnaire data	82

2.4. Measures

2.4.1. Independent variables – questionnaire data

The self-reported questionnaire data used in the analysis represented personal characteristics of individual drivers and were grouped into three categories:

Demographics: A demographic questionnaire was used in SHRP 2 to investigate a variety of demographic information about the participant. The present analysis only selected two

variables from this questionnaire: age and gender. Age was stratified into three age groups: younger than 25 years, between 25 and 55 years, and older than 55 years.

Driving history: A driving history questionnaire was used to obtain self-reported driving history information about the participant, including driving experience, violations and crashes during the past three years, and training received. The present analysis included two variables from this questionnaire: self-reported violations and self-reported crashes in the past three years. These two independent variables were recoded into binary variables indicating whether or not the participant had at least one violation (or crash) in the last three years.

Personality: Three self-report questionnaires were included from the SHRP 2 study: a modified (Dingus et al., 2015) version of the *Manchester Driver Behavior Questionnaire* (M-DBQ; Reason et al., 1990; Lajunen and Summala, 2003) containing 24 items answered on a 6-point Likert scale, the *Sensation Seeking Scale-form V* (SSS-V; Zuckerman, 1994; Jonah, 1997) with 40 items (Dingus et al., 2015), and a *Risk-Perception Questionnaire* (Dingus et al., 2015).

The M-DBQ is a self-report driver behavior survey, participants indicate (from 0 = "Never" to 6 = "Nearly All the Time") how often they commit each described error (accidental) or violation (deliberate), where the concepts of errors and violations derive from the work of Reason (1990). A principal component analysis (PCA) with oblique rotation was applied on the 24-item M-DBQ . For the calculation of the subscales, the item 'Attempt to drive away from traffic lights in the wrong gear' was removed because of its non-applicability in the USA. In accordance with de Winter et al. (2018) a three-component solution was identified using a scree-plot. The items were grouped by the highest factor loadings and the components were interpreted as *(1) slips, (2) violations, and (3) lapses*. The mean score of each scale was calculated and used in further analyses. Ten items were assigned to slips, eight items to violations, and five items to lapses (see Annex).

The SSS-V is a self-report survey where respondents chose which of two choices better describes their feelings or likes (Zuckerman, 1994). The present analysis used the *total score of the SSS-V,* to indicate the degree to which the participant engages in sensation seeking behavior.

A risk-perception questionnaire created for the SHRP 2 data collection (Dingus et al., 2015) was also included. The questions assess the perceptual risk with driving behaviors on a seven-point Likert scale ranging from 'No Greater Risk' to 'Much Greater Risk'. Most items provided little variance across drivers, so only one item was selected for inclusion. The item selected was "If you were to engage in changing lanes suddenly to get ahead in traffic, how do you think that would affect your risk of a crash?"

2.4.2. Independent variables – recorded driving style

Driving style here refers to persistent driving patterns for individual drivers (Sagberg et al., 2015) identified using continuous time series driving data. In the present study, driving style was operationalized based on Simons-Morton et al. (2013) in terms of the rates (number per mile) of six types of kinematic events calculated for the study period based on specific g-force or yaw rate thresholds for each metric. These kinematic events consisted of hard starts, stops, left turns, right turns, left yaw movement, and right yaw movement. Multiple events were counted as one if the interval between them was less than one second and events were removed if their event

duration was less than half a second. The present analysis built logistic regressions between each dependent variable and each driving style measure at 46 different g-force thresholds and, for each driving style variable, selected the specific g-force level thresholds with the minimum Akaike Information Criterion (AIC) value, indicating the highest quality of statistic model. AIC is an estimator of the relative quality of statistical models (or the relative information lost when a given model is used), and minimum AIC was used to ensure the goodness of fit (by maximizing likelihood) as well as prevent overfitting (by minimizing the number of parameters).

An example illustrating the calculation of two of the driving style measures, hard starts and hard stops, is plotted below (Figure 2). In this example, purple lines (+0.24 g and +0.33 g) result in different numbers of hard starts (1 and 0) and red lines (-0.24 g and -0.29 g) result in different numbers of hard stops (4 and 2). This plot includes data from one trip (May 21st, 2011) and one driver.

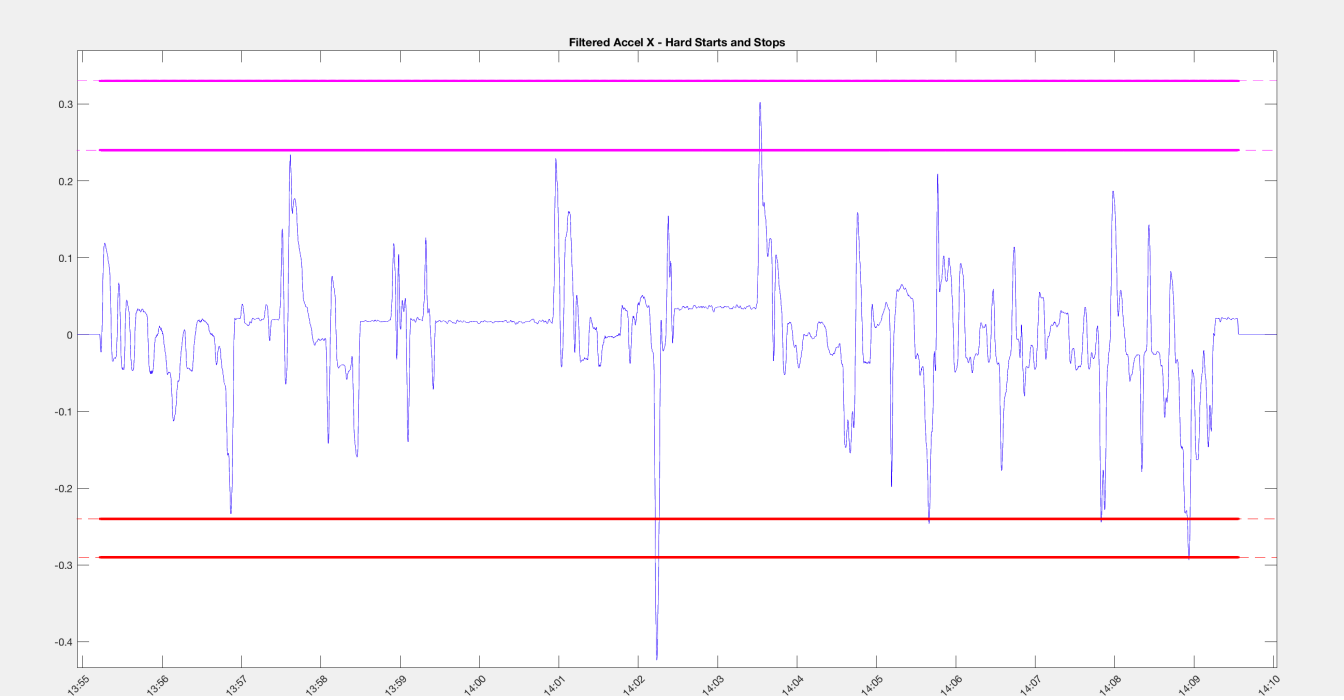

Figure 2. Kinematic events, consisting of hard starts and hard stops at different levels (purple and red lines, respectively) based on time series data, acceleration X.

2.4.3. Dependent variables – crash and near crash involvement measures

In the SHRP 2 dataset, safety critical events (SCEs) (i.e., crash, near crash, crash-relevant conflicts, subject conflicts) have been manually validated and coded by trained data reductionists. Only crashes and near crashes were used in the present analysis. A crash is defined as 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, and is coded at four levels of severity. A near crash is defined as any circumstance that requires a rapid evasive maneuver (e.g., steering, braking, accelerating, or combination of control inputs) by the subject vehicle or any other vehicle, pedestrian, cyclist, or animal to avoid a crash (see Hankey et al., 2016).

Crashes in the SHRP 2 dataset have been further coded into four severity levels (Hankey et al., 2016):

- *Level 1 (Severe Crash):* Any crash that includes an airbag deployment; any known injury of driver, pedal cyclist, or pedestrian (one sufficient to warrant a doctor's visit, including those self-reported and those apparent from video); a vehicle rollover; a high Delta-V; or vehicle damage requiring towing. A high Delta-V is defined as a change in speed of the subject vehicle in any direction during impact greater than 20 mph (excluding curb strikes) or (more commonly) acceleration on any axis greater than +/-2 g (excluding curb strikes).
- *Level 2 Crash Moderate Severity:* Not a level 1 crash; minimum of approximately $1,500 worth of damage as estimated from video. It also includes crashes that reach acceleration on any axis greater than +/-1.3 g (excluding curb strikes). Examples are most large animal and sign strikes.
- *Level 3 Crash Minor Severity:* Not a level 1 or 2 crash; the vehicle makes physical contact with another object or departs the road but sustains only minimal or no damage. This includes most road departures (unless criteria for a more severe crash are met), small animal strikes, all curb and tire strikes potentially in conflict with oncoming traffic, and other curb strikes with an increased risk element (i.e., the crash may have been worse if the curb had not been there).
- *Level 4 Crash Tire Strike, Low Risk:* Not a level 1, 2, or 3 crash; the tire is struck with little or no risk element (e.g., clipping a curb during a tight turn).

For this analysis, crashes of all severity levels were used. Two dependent variables represented crash and near crash involvements of individual drivers:

Crash or Near Crash (CNC) is binary variable, indicating whether the participant was involved in zero or at least one crash or near crash event in the study period.

Crash is a binary variable, indicating whether the participant involved zero or at least one crash event in the study period.

A driver was labelled as a CNC- or crash-involved driver if the driver had at least one CNC or crash in the study period.

2.5. Statistical models

Two types of classification models were investigated with the goal to identify CNC- (crash-) involved drivers based on the independent self-reported and driving style variables described above: logistic regression and random forest classification (as described above, separate logistic regression was initially conducted for each of the driving style measures to determine optimal g-force threshold values based on minimum AIC).

The prediction performance of the models was evaluated in terms of the recall rate (or sensitivity), precision (positive predictive value), and accuracy. In the context of this study, the recall rate is the number of correctly predicted CNC- or crash- involved drivers divided by the total number of CNC- (crash-) involved drivers. The precision is the number of correctly predicted CNC- (crash-) involved drivers divided by the total number of drivers predicted by a model to be CNC- (crash-) involved. Finally, the accuracy is the fraction of all drivers correctly classified as either CNC- (crash-) involved or not involved.

2.5.1. Logistic regression

Logistic regression is a statistical classification method that was used to model the probability of being a CNC- or crash-involved driver (for a similar use, see Guo and Fang, 2013). The dependent variable (CNC or Crash) is a binary variable and is assumed to follow a Bernoulli distribution with a probability (p_i). This probability is associated with a set of covariates by a logit link function where the set of covariates are all potential independent variables:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i,1} + +\beta_2 x_{i,2} + \cdots + +\beta_k x_{i,k} = \boldsymbol{X_i\beta} \qquad (1)$$

and $\boldsymbol{X_i}$ is the matrix of predictors for individual *i*, and $\boldsymbol{\beta}$ is the vector of regression parameters. Both forward and backward variable selections were performed and the best model was selected based on the minimum AIC value. A driver will be predicted as a CNC- or crash-involved driver if this probability is greater than a predefined threshold value (e.g., $p_0 = 0.5$). The Odds Ratio ($OR_j = \exp(\beta_j)$) is the change in probability of being a CNC driver versus not being a CNC involved driver associated with a variable *j*.

2.5.2 Random forest classification

Random forest (RF) classification, proposed by Breiman (2001), is an ensemble learning method that can be used for both classification and regression problems. This method creates a series of decision trees (i.e., forest), each of which is used first to solve the classification problem individually. Subsequently, the final result is obtained based on the majority vote across all decision trees. The decision tree algorithm, introduced by Breiman (1984), uses a recursive binary splitting approach to grow a tree by selecting a predictor and a cut point for that predictor to split the data into two parts. This procedure is iterated at several steps to create a dendrogram type of structure (i.e., a decision tree). At each splitting step, different criteria can be used to identify the best split (i.e., best classification). The criterion used in this study is the Gini Index *G* (Hastie et al., 2009) (see Table for definitions of the terms and indices):

$$G = \sum_{k=1}^{K} p_k^m (1 - p_k^m) \qquad (2)$$

At each step, the predictor that results in the highest decrease in the Gini Index is selected.

Table 2. Definitions of terms and indices in equation (2).

Variable	Definition
$\boldsymbol{K}$	Number of classes
$\boldsymbol{k}$	Class *k*
$\boldsymbol{x_i^m}$:	Predictor vector of i^{th} observation in node *m*
$\boldsymbol{y_i^m}$:	Target value of i^{th} observation in node *m*
$\boldsymbol{p_k^m}$:	Proportion of class *k* observations in node *m* ($\frac{1}{N^m}\sum_{x_i^m} I(y_i^m = k)$)

$I(y_i^m = k)$:	1 if $y_i^m = k$, and 0 otherwise
N^m	Number of observations received at node m

3. Results

3.1 Training and test data

The full dataset of 2458 drivers was randomly partitioned into two balanced groups: a training set (70%, 1720 drivers) and a test set (30%, 738 drivers). Table shows that (1) percentages of CNC- or crash-involved drivers in these three datasets were consistent; (2) the proportions of CNC and non-involved drivers were roughly equal, and (3) there were about four times more drivers that were not involved in a crash during the study period than drivers that had a crash during that period as operationalized using the crash variable.

Table 3. Frequency table for full dataset, training and test sets separately.

Percentage of drivers	Full dataset	Training set	Test set
	(2458 drivers)	(1720 drivers)	(738 drivers)
CNC = 1 (at least one CNC event)	55.5%	55.1%	56.2%
Crash = 1 (at least one crash event)	18.6%	18.2%	19.6%

3.2. Determining g-force thresholds for the driving style measures

In order to find suitable values for the thresholds of the driving style (g-force) measures, logistic regressions were built between each dependent variable (CNC and Crash) and each driving style measure at 46 different g-force threshold levels (see Figures 3 and 4). The thresholds were chosen based on the minimum AIC value for each variable.

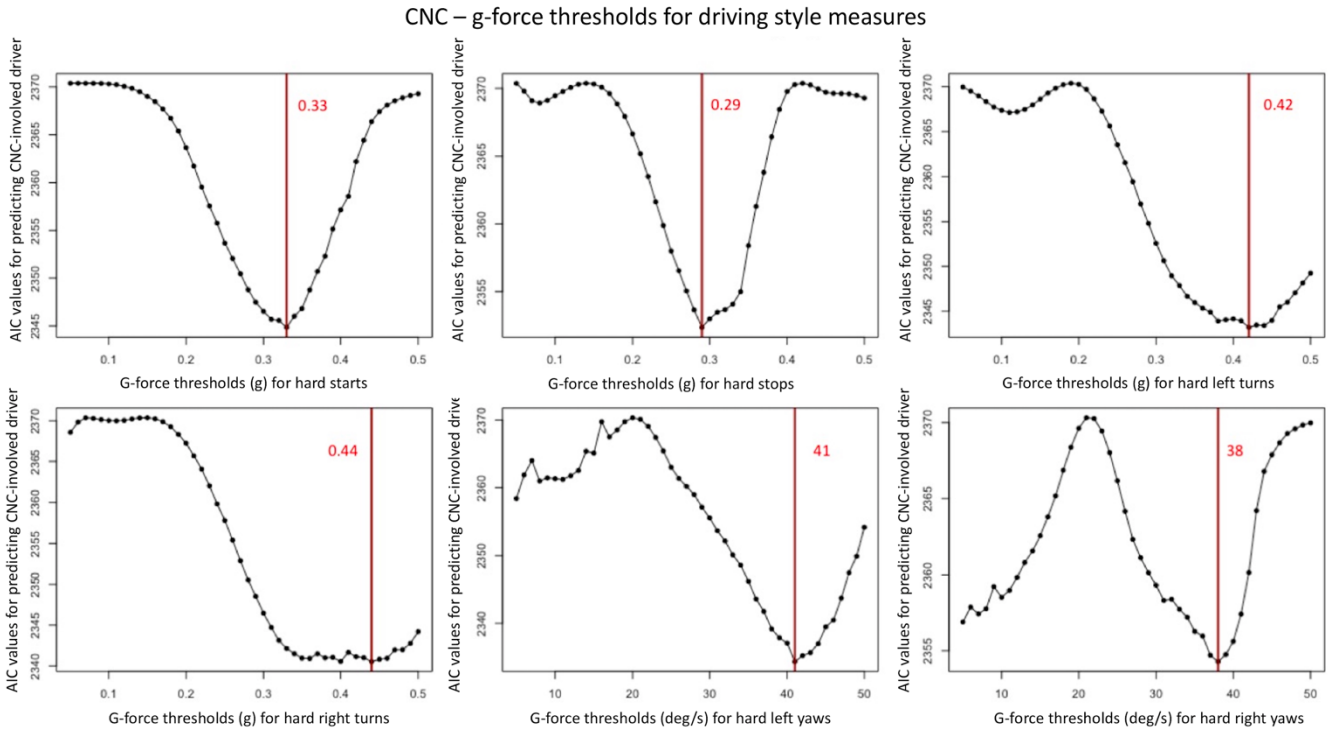

Figure 3. AIC values of logistic regression models between CNC and each driving style measure at 46 g-force thresholds.

Note. Red lines indicate the selected gravitational level with the minimum AIC value for each driving style measure.

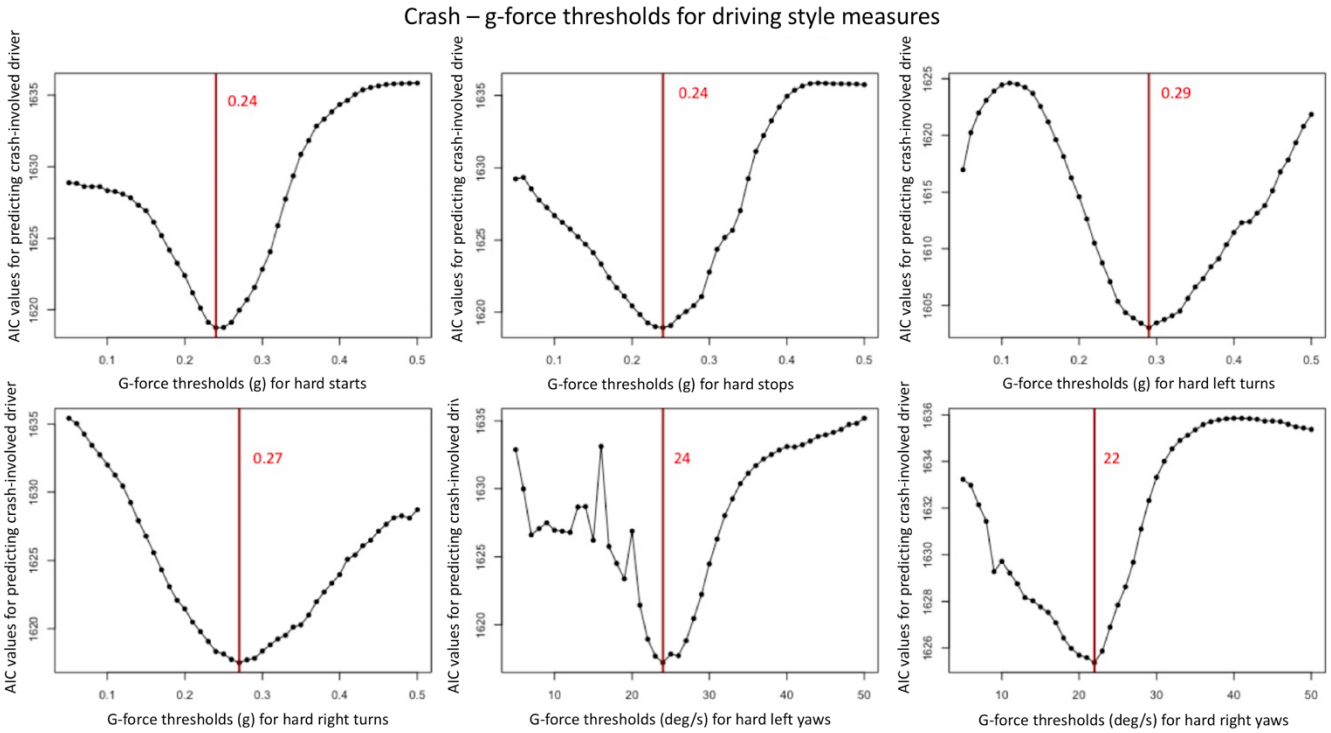

Figure 4. AIC values of logistic regression models between crash and each driving style measure at 46 different g-force or yaw rate thresholds.

Note. Red lines indicate the selected gravitational level with the minimum AIC value for each driving style measure.

The selected g-force levels for six driving style measures for each dependent variable are summarized in Table 4. Six driving style measures at specific g-force levels were selected as predictors in the logistic regression model to predict CNC- or crash-involved drivers.

Table 4. Thresholds of gravitational forces for each driving style measure for the two dependent variables (CNC and crashes)

Driving Style Measures	Time-Series Data Source	Gravitational Levels	Gravitational Levels
		CNC	Crash
Hard starts	Acceleration X	> + 0.33 g	> + 0.24 g
Hard stops	Acceleration X	< - 0.29 g	< - 0.24 g
Hard left turns	Acceleration Y	< - 0.42 g	< - 0.29 g
Hard right turns	Acceleration Y	> + 0.44 g	> + 0.27 g
Hard left yaws	Gyro Z	< - 41 deg/s	< - 24 deg/s
Hard right yaws	Gyro Z	> + 38 deg/s	> + 22 deg/s

3.3. Logistic regression models

Logistic regression modeling (LR) was implemented in the R software environment. The present analysis built a logistic regression model for each dependent variable, CNC-involvement and crash-involvement. The logistic regression model for predicting CNC involvement is henceforth referred to as LR_CNC while the model predicting crash involvement is referred to as LR_Crash. The variable selection results were the same regardless of whether forward or backward selection (with the minimum AIC value) was used.

The prediction performance results of logistic regression models are shown in Table 5 and Table 6. Table 6 shows that the LR_CNC model applied to the test data has a high recall rate (72%) and relatively good precision and accuracy rates (64% and 62% respectively). As also shown in Table 5, the LR_Crash model has a high accuracy (80%) on the test set. However, as is clear from the confusion matrices in Table 6, this is due to the model ignoring the crash-involved category, thus classifying all drivers but one as not crash-involved. Hence the accuracy trivially reflects the proportion of non-crash involved drivers in the data.

Table 5. The prediction performance of the logistic regression models.

Models	Training Set			Test Set		
	Recall	Precision	Accuracy	Recall	Precision	Accuracy
LR_CNC	0.691	0.613	0.589	0.723	0.644	0.619
LR_Crash	0.010	0.429	0.817	0.000	0.000	0.802

Table 6. The confusion matrices of the logistic regression models.

Models			Training Set						Test Set			
LR_CNC				Predicted						Predicted	
				0	1					0	1
			sum	651	1069			sum	272	466
	Actual	0	722	358 (20.8%)	414 (24.1%)		Actual	0	323	157 (21.3%)	166 (22.5%)
		1	948	293 (17.0%)	655 (38.1%)			1	415	115 (15.6%)	300 (40.6%)
LR_Crash				Predicted						Predicted	
				0	1					0	1
			sum	1713	7			sum	737	1
	Actual	0	1407	1403 (81.6%)	4 (0.2%)		Actual	0	593	592 (80.2%)	1 (0.1%)
		1	313	310 (18.0%)	3 (0.2%)			1	145	145 (19.7%)	0

The logistic regression results for the LR_CNC model are shown in Table 7 (the results for the LR_Crash model are not shown as this classifier failed to predict any crashes). The self-reported violations, hard start rate, M-DBQ 2 (violations), SSS-V total score, and hard left yaw rate show statistically significant effects on the probability of being a CNC-involved driver. However, it should be noted that these results are somewhat hard to interpret as the odds ratios vary significantly in size and confidence intervals for some of the driving style variables. This may be due to a relatively high threshold level resulting from the threshold ~~determination~~ procedure illustrated in Figures 3 and 4. This leads to very few instances of non-zero values for these variables, making the statistical results brittle.

Table 7. Logistic regression models outputs.

LR_CNC Model						
Effect	Estimate	Std. Error	z value	*p*-Value		Odds ratio [95% CI]
Intercept	-0.593	0.320	-1.850	0.064	.	0.553 [0.294, 1.033]
Self-reported violation [1-0]	0.426	0.112	3.804	<0.001	***	1.531 [1.230, 1.908]
SSS-V total score	0.022	0.008	2.714	0.007	**	1.022 [1.006, 1.039]
M-DBQ 2 - violations	0.233	0.137	1.694	0.090	.	1.262 [0.967, 1.656]
Risk-taking lane change	-0.055	0.035	-1.590	0.112		0.946 [0.884, 1.013]
Hard start rate	5.245	2.175	2.411	0.016	*	189.642 [3.750, 17864.073]
Hard stop rate	0.733	0.403	1.818	0.069	.	2.081 [0.959, 4.665]
Hard left yaw rate	23.252	8.598	2.704	0.007	**	1.25E+10 [3.59E+03, 7.62E+17]

Note: . = *p*-value<0.1; * = *p*-value <0.05; ** = *p*-value <0.01; *** = *p*-value <0.001

3.4. Random forest models

Random forest modeling (RF) was implemented in the R software environment using the "randomForest" package (Liaw and Wiener, 2002). Two tuning parameters needed to be determined to develop a random forest model, namely total number of trees and the number of randomly selected predictors to grow each tree. By experimenting different values, 500 and 4 were used for these parameters, respectively. Three random forest models were developed, and the results are shown in Tables 8 and 9.

Table 8 shows that RF_CNC has high recall rate and good accuracy rate that are very close to the results of logistic regression models for the test set. Also, shown by the confusion matrices in Table 9, the model performance on predicting crash involvement is also very similar to the logistic regression model, where almost all drivers are classified as not crash-involved, indicating that the model failed to learn to recognize crash-involved drivers.

Table 8. The prediction performance of random forest models.

Models		Training Set			Test Set	
	Recall	Precision	Accuracy	Recall	Precision	Accuracy
RF_CNC	1.000	1.000	1.000	0.745	0.652	0.633
RF_crash	1.000	1.000	1.000	0.014	0.333	0.800

Table 9. The confusion matrices of random forest models.

Models			Training Set					Test Set		
RF_CNC				Predicted				Predicted	
				0	1			0	1
			sum	772	948		sum	264	474
	Actual	0	772	772 (44.9%)	0	Actual	0	323	158 (21.4%)	165 (22.3%)
		1	948	0	948 (55.1%)		1	415	106 (14.4%)	309 (41.9%)
RF_crash				Predicted				Predicted	
				0	1			0	1
			sum	1407	313		sum	732	6
	Actual	0	1407	1407 (81.8%)	0	Actual	0	593	589 (79.8%)	4 (0.5%)

While developing the random forest model, predictor importance is internally calculated by measuring the Gini Index decrease caused by each predictor averaged across all trees. The results for the CNC model are shown in Figure 5 (again, results for the crash prediction model are not included). As can be seen from the plot, the most important variables in predicting CNCs are hard stop rate, hard start rate, hard left turn rate, hard right turn rate, hard left yaw rate, and hard right yaw rate. These results, where the driving style variables are indicated as the most important predictors, yield a somewhat different picture than the logistic regression model above, where also self-reported violation/crash and SSS-V total score showed up as significant predictors.

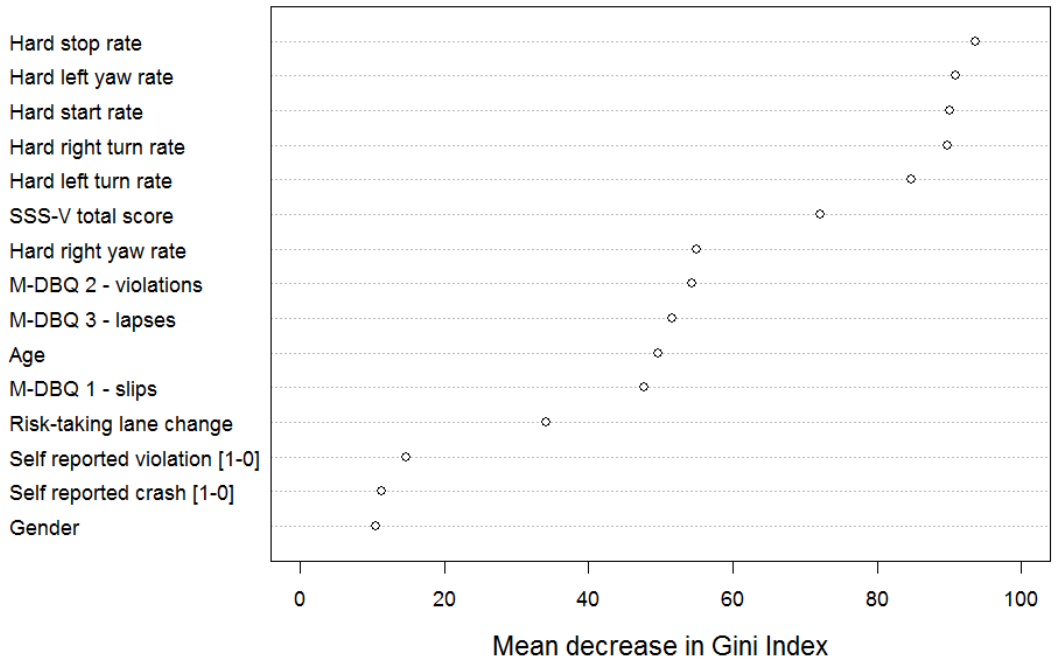

Figure 5. Importance of predictors in random forest model predicting CNCs based on mean decrease in Gini Index. High mean decreases in Gini reflect high importance

4. Discussion

The goal of the present analysis was to investigate, using the SHRP 2 dataset, to what extent it is possible to predict crash and/or near crash involvement for individual drivers based on enduring personal factors related to demographics, driving history, personality, and observed

driving style. Two types of classification models, logistic regression and random forest, were employed and yielded similar results.

In line with previous results based on naturalistic driving data (Guo and Fang, 2013; Simons-Morton et al., 2012), it was demonstrated that individual driver involvement in crashes *or* near crashes (CNC) in previously unseen (test) data can be predicted with some accuracy (62-63%) based on enduring personal factors. Moreover, the models were able to correctly identify 72-75% of the CNC-involved drivers (recall) and of those drivers predicted by the models to be involved in a CNC during the study period, 64-65% were correct predictions (precision). It should be noted that CNC involvement here mainly included involvement in near-crashes (55.5% of the drivers were involved in a CNC while only 18.6% were involved in a crash during the study period).

The present study is, to our knowledge, the first to, relate enduring personal factors specifically to crash (rather than CNC) involvement based on naturalistic driving data. However, the results here were somewhat less encouraging. Both models essentially failed to recognize crash-involved drivers and classified almost all driver as non-crash-involved. This result is likely a combination of the relatively low proportion of crash-involved drivers in the training data and a weak association between the currently used predictor variables and individual crash involvement. Thus, optimal classification performance could be obtained by trivially classifying all data points into the dominant category (i.e., non-involvement in crashes).

The present results thus show that it is possible to predict CNC (mainly near-crash) involvement for individual drivers with some degree of accuracy, while this seems to be more difficult for crashes alone. In a companion commentary, using a nearly identical dataset as in the present study, de Winter et al. (2018) looked specifically at correlations between, on the one hand, the Driver Behavior Questionnaire and the Sensation Seeking Scale and, on the other, crash and near crash involvement (as well as some of the g-force driving style measures investigated here). In line with the present results, the correlations between DBQ (Violations sub-scale) and SSS scores and near crashes were substantially higher (around 0.20) than the corresponding correlations for crashes (around 0.05-0.10). de Winter et al. (2018) also found relatively high correlations between the DBQ and the present g-force metrics.

Hence, taken together, these studies, as well as a complementary analysis presented in Huang et al. (2018), indicate that both near crash and crash involvement is associated with, and can to some extent be predicted from, enduring personal factors, but that the association is relatively weak. This indicates that the involvement in crashes and near crashes is also strongly influenced by more temporary personal as well as situational factors. This is in line with existing crash causation models, such as the Swiss cheese and the Crash Trifecta models (see Knipling, 2009) which suggest that multiple driver and situational factors typically have to align to produce a crash. This suggests that, rather than assuming a direct relationship between personal factors and crashes, the relationship is more complex and moderated/mediated by other temporal and situational factors. Hence, it might be interesting to explore more complex models including such factors as co-variates.

Moreover, the results indicate that enduring personal factors, at least those represented by the present predictor (independent) variables, appear to have a stronger association with near crashes than with crashes. It is difficult, based on the present results, to draw any strong

conclusions on why this may be the case but some speculation may be warranted. First, the fact that the training set contained more examples of near crashes than crashes is probably led to better classification of the former (see de Winter et al., 2018). Another potential reason could be that the current predictor variables mainly capture individual characteristics related to aggressive driving and that aggressive driving may be more strongly associated with near-crashes than with crashes. This notion is supported by the relatively high correlations between driving style and personality questionnaire ($DBQ_{violation}$ and SSS-V) scores and near crashes/g-force measures found by de Winter et al. (2018). By contrast, crashes may to a larger extent than near crashes be associated with driver inattention combined with rare/unexpected circumstances. This is supported by existing naturalistic driving analyses of driver inattention and crash/near crash involvement, which typically found eyes-off-road to be more common in crashes than in near crashes (Klauer et al., 2006; Victor et al., 2015). Driver inattention is most likely also strongly associated with enduring personal factors (e.g., individual drivers may differ consistently in their willingness to engage in secondary tasks), but such these characteristics may not be well-captured by the present predictor variables.

It is also possible that near-crashes recorded in naturalistic driving studies are to a greater extent than crashes related to the behavior of the observed driver. For example, near crashes where the subject vehicle is almost struck in the rear were typically not recorded in SHRP 2 (detecting such events would require rear proximity sensors which were not included in SHRP 2) while rear-end striking crashes are.

In any case, it would clearly be premature to dismiss the possibilities of predicting crash involvement from enduring personal factors solely based on the present results, and there are several ways the classification models may be improved and other ways of analyzing this data which may shed further light on the relationship between enduring personal factors and crash involvement.

First, while SHRP 2 is the largest naturalistic data collection to date, the number of crashes recorded is still relatively limited, especially more severe crashes (the present dataset is dominated by minor severity Level 3 and tire strike/low risk Level 4 crashes). Thus, it is possible that the results would have been different if a larger set of severe crashes were available. Such datasets do exist today in the commercial sector and have been used in related analyses (e.g., SmartDrive, 2017).

Second, the current driving style variables were relatively simple (g-force events), building on existing studies (Simons-Morton et al., 2012) and typically used in existing applications in the auto insurance domain (e.g., Drive Safe and Save by StateFarm and Snapshot offered by Progressive). Also, the current data-driven approach for finding optimal g-force threshold values based on AIC values of individual logistic regression models for each variable, was clearly associated with certain issues. In particular, this approach yielded high thresholds for some variables, with only few drivers having non-zero values, leading to brittle statistical results. There is clearly much room for developing more sophisticated driving style indicators that may have a stronger relationship to crash involvement. Some potential candidates include jerk (Bagdadi and Varhelyi, 2011) and various measures based on speeding and close following (see Sagberg et al., 2015). Moreover, as suggested above, the current predictors are mainly related to aggressive driving and including indirect or direct measures of "inattention propensity" may help to improve model predictions for crashes.

5. References

1. Bagdadi, O., & Varhelyi, A., 2011. Jerky driving—An indicator of accident proneness? Accid. Anal. Prev. 43 (4), 1359-1363. doi:10.1016/j.aap.2011.02.009

2. Boris and Luciana, 2017. Developing a Younger Driver Assessment Tool. Technical Memorandum #1. American Transportation Research Institute (ATRI).

3. Breiman, L. 1984. Classification and Regression Trees. Chapman & Hall/CRC.

4. Breiman, L., 2001. Random forests. Mach. Learn. 45 (1) , 5–32.

5. Dahlen, E. R., White, R. P., 2006. The Big Five Factors, Sensation Seeking, and Driving Anger In the Prediction of Unsafe Driving. Personality and Individual Differences, 41 (5), 903-915. doi:10.1016/j.paid.2006.03.016

6. de Winter, J.C.F., Dreger, F.A., Huang, W., Miller, A., Soccolich, S., Machiani, S.G., Engström, J., 2018. The relationship between the Driver Behavior Questionnaire, Sensation Seeking Scale, and recorded crashes: A brief comment on Martinussen et al. (2017) and new data from SHRP 2. Accid. Anal. Prev. 118, 54–56. doi:10.1016/j.aap.2018.05.016

7. Dingus, T.A., Hankey, J.M., Antin, J.F., Lee, S.E., Eichelberger, L., Stulce, K.E., McGraw, D., Perez, M., Stowe, L., 2015. Naturalistic driving study: Technical coordination and quality control.

8. 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., Bucher, C., Doerzaph, Z. R., Jermeland, J., & Knipling, R.R., 2006. The 100-Car Naturalistic Driving Study Phase II – Results of the 100-Car Field Experiment. HNTSA DOT, Report No: HS 810 593

9. Engström, J. Ghanipoor Machiani, S. Miller, A., Huang, W. and Soccolich, S., 2017. Behavior-based Predictive Safety Analytics, Deliverable 2.1: State-of-the-art review. Project Deliverable, Behavior-based Predictive Analytics, Safe-D (Safety Through Disruption) University Transportation Center.

10. Guo, F., Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. 61, 3-9. doi:10.1016/j.aap.2012.06.014

11. 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.

12. Hanowski, R. J., Wierwille, W. W., Garness, S. A., and Dingus, T. A., 2000. Impact of Local/Short Haul Operations on Driver Fatigue. Final Report No. DOT-MC-00-203. Washington, DC, U.S. Department of Transportation, Federal Motor Carriers Safety Administration, September.

13. Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning, in: The Elements of Statistical Learning. Springer, pp. 485–585.

14. Huang, W., Engström, J., Miller, A., Dreger, F., Soccolich, S., de Winter, J., and Ghanipoor Machiani, S., 2018. Analysis of differential crash and near-crash involvement based on naturalistic driving data. Abstract accepted for presentation at the Seventh International Symposium on Naturalistic Driving Research, Blacksburg, Virginia on August 28-30, 2018.

15. Jonah, B.A., 1997. Sensation seeking and risky driving: a review and synthesis of the literature. Accid. Anal. Prev. 29 (5), 651–665. doi:10.1016/S0001-4575(97)00017-1

16. 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. US Department of Transportation.report No. DOT HS 810 59.

17. Knipling, R.R., 2009. Safety for the long haul: Large truck crash risk, causation, & prevention. American Trucking Association Arlington, VA.

18. Knipling, R.R., 2004. Individual differences and the" high-risk" commercial driver. Transportation Research Board.

19. Lajunen, T., & Summala, H., 2003. Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. Transportation Research, Part F, 6, 97-107. doi:10.1016/S1369-8478(03)00008-1

20. Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18-22.

21. Lueck, M.D. and Murray, D.C., 2011. Predicting Truck Crash Involvement: A 2011 Update. American Transportation Research Institute, Arlington, VA.

22. McCartt, A.T., Mayhew, D.R., Braitman, K.A., Ferguson, S.A., Simpson, H.M., 2009. Effects of age and experience on young driver crashes: review of recent literature. Traffic Inj. Prev. 10 (3), 209–219. doi:10.1080/15389580802677807

23. McKenna, F.P., 1983. Accident proneness: A conceptual analysis. Accid. Anal. Prev. 15 (1) , 65–71. doi:10.1016/0001-4575(83)90008-8

24. Murray, D., Lantz, B., Keppler, S., Alliance, C.V.S., Board, T.R., 2006. Predicting truck crash involvement: Developing a commercial driver behavior model and requisite enforcement countermeasures, in: Transportation Research Board 85th Annual Meeting.

25. Reason J., 1990. Human Error. New York: Cambridge University Press.

26. Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K., 1990. Errors and violations: a real distinction? Ergonomics, 33 (10-11), 1315-1332. doi:10.1080/00140139008925335

27. Sagberg, F., Selpi, Bianchi Piccinini, G.F., Engström, J., 2015. A review of research on driving styles and road safety. Hum. Factors 57 (7), 1248–1275. doi:10.1177/0018720815591313

28. Simons-Morton, B.G., Cheon, K., Guo, F., Albert, P., 2013. Trajectories of kinematic risky driving among novice teenagers. Accid. Anal. Prev. 51, 27–32. doi:10.1016/j.aap.2012.10.011

29. Simons-Morton, B.G., Zhang, Z., Jackson, J.C., Albert, P.S., 2012. Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes? Am. J. Epidemiol. 175 10 , 1075–1079. doi:10.1093/aje/kwr440

30. SmartDrive., 2017. Measuring driver risk with video-based analytics. San Diego, California: SmartDrive Systems.

31. Soccolich, S.A., Hickman, J.S., Hanowski, R.J., 2011. Identifying high-risk commercial truck drivers using a naturalistic approach.

32. Truck and Bus Safety Management Techniques; A Synthesis of Safety Practice. TRB Commercial Truck & Bus Synthesis Program Project, ISSN 1544-6808, ISBN 0-309-08754-6.

33. Victor, T., Bärgman, J., Boda, C-N., Dozza, J., Engström, J., Flannagan, C. A., Lee, J. D., Markkula, G., 2014. Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk. SHRP 2 Research Report.

34. West, R., French, D., Kemp, R., Elander, J., 1993. Direct observation of driving, self reports of driver behaviour, and accident involvement. Ergonomics 36 (5), 557–567. doi:10.1080/00140139308967912

35. Zuckerman, M., 1994. Behavioral expressions and biosocial bases of sensation seeking. Cambridge university press.

Annex

	Component		
	[I] (Slips)	[II] (Violations)	[III] (Laps)
Item			
Attempt to drive away from traffic lights in the wrong gear	-0.103	-0.079	-0.148
Become impatient with a slow driver in the fast lane and pass on the right	0.214	**-0.732**	-0.168
Drive especially close to a car in front as a signal to the driver to go faster or get out of the way	0.153	**-0.788**	-0.046
Attempt to pass someone that you hadn't noticed to be making a left turn	-0.124	**-0.463**	-0.058
Forget where you left your car in a parking lot	0.059	0.016	**-0.714**
Turn on one thing, such as your headlights, when you mean to switch on something else, such as the windshield wipers	-0.266	0.089	**-0.449**
Realize that you have no clear recollection of the road along which you have just been traveling	0.121	-0.193	**-0.667**
Cross an intersection knowing that the traffic lights have already changed from yellow to red	-0.048	**-0.385**	-0.281
Fail to notice that pedestrians are crossing when turning onto a side street from a main road	**-0.527**	-0.011	-0.128
Angered by another driver's behavior, you catch up to them with the intention of giving him/her "a piece of your mind."	-0.235	**-0.497**	0.218
Misread the signs and turn the wrong direction on a one-way street	**-0.554**	0.153	-0.145

Disregard the speed limits late at night or early in the morning	0.102	**-0.625**	-0.187
When turning right, nearly hit a bicyclist who is riding along side of you	**-0.72**	0.083	0.282
Attempting to turn onto a main road, you pay such close attention to traffic on the road you are entering that you nearly hit the car in front of you that is also waiting to turn.	**-0.479**	-0.014	-0.114
Drive even though you realize you might be over the legal blood alcohol limit	-0.006	**-0.395**	0.039
Have an aversion to a particular class of road user, and indicate your hostility by whatever means you can	**-0.393**	-0.340	0.302
Underestimate the speed of an oncoming vehicle when attempting to pass a vehicle in your own lane	**-0.5**	-0.124	-0.002
Hit something when backing up that you had not previously seen	**-0.514**	0.197	-0.099
Intending to drive to destination A, you 'wake up' to find yourself on a road to destination B, perhaps because destination B is a more common destination.	-0.041	-0.051	**-0.593**
Get into the wrong lane approaching an intersection	-0.313	0.066	**-0.364**
Miss "Yield" signs, and narrowly avoid colliding with traffic having the right of way	**-0.733**	0.117	0.047
Fail to check your rearview mirror before pulling out, changing lanes, etc.	**-0.4**	-0.117	-0.071
Get involved in unofficial 'races' with other drivers	-0.035	**-0.52**	0.124
Brake to quickly on a slippery road or steer the wrong way into a skid	**-0.466**	-0.027	-0.077

* bold when assigned to component

Appendix C. Course Curriculum

Curriculum for a Course Module on Differential Crash Involvement and Behavior-based Predictive Analytics

Learning Objective

The objective of this course module is to provide students with an enhanced understanding of how drivers' involvement in crashes can be related to, and predicted based on, individual characteristics.

Content

The course will address theoretical concepts related to individual driving style and differential crash involvement, statistical techniques for analyzing and predicting individual crash involvement, as well as practical applications, for example in the context of driver education and screening, insurance, and fleet management.

The module will consist of three lectures, which may be reduced or expanded based on the general course in which the module is taught:

Lecture 1. The concept of differential crash involvement. This lecture will provide a historical background to the study of how individual characteristics are related to crashes and introduce key theoretical concepts commonly used in this context.

Example literature:

Elander, J., West, R., & French, D. (1993). Behavioral correlates of individual differences in road traffic crash risk: An examination of methods and findings. *Psychological Bulletin*, *113*, 279–294.

Knipling, R.R., 2009. Safety for the long haul: Large truck crash risk, causation, & prevention. American Trucking Association Arlington, VA.

McKenna, F. P. (1983). Accident proneness: A conceptual analysis. *Accident Analysis and Prevention, 15,* 65–71.

Sagberg, F., Selpi, Piccinini, G. F. & Engström, J. 2015. A Review of Research on Driving Styles and Road Safety. *Human Factors, 57*(7), 1248–75.

Lecture 2: Statistical modeling of differential crash involvement. This lecture will introduce the key statistical methods typically used in differential crash involvement research, including Poisson and negative binomial regression, logistic regression and machine learning techniques, and discuss example applications, such as modeling individual crash rates, classifying drivers into risk groups, etc.

Example literature:

Guo, F., & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, *61*, 3–9. https://doi.org/10.1016/j.aap.2012.06.014

Simons-Morton, B.G., Zhang, Z., Jackson, J.C., Albert, P.S., 2012. Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes? *Am. J. Epidemiol, 175*(10), 1075–1079. doi:10.1093/aje/kwr440

Lecture 3: Applications: This lecture will focus on real-world applications of research on differential crash involvement, in particular in the areas of driver education and evaluation, insurance, and commercial fleet management.

Example literature:

Boris and Luciana (2017). Developing a Younger Driver Assessment Tool. Technical Memorandum #1. American Transportation Research Institute (ATRI).

Huetter, J. (2017). Progressive: Usage-based insurance is the future, likely assisted by OEM data. Retrieved from http://www.repairerdrivennews.com/2017/05/17/progressive-usage-based-insurance-is-the-future-likely-assisted-by-oem-data/

Lytx. 2017. Industry insights: Beyond telematics: How video predicts risky behavior. Lytx White Paper.

Murray, D. C., Lantz, B., & Keppler, S. A. (2005). Predicting Truck Crash Involvement: Developing a Commercial Driver Behavior-Based Model and Recommended Countermeasures. Retrieved from https://trid.trb.org/view.aspx?id=771443

SmartDrive. (2017). Measuring driver risk with video-based analytics. San Diego, California: SmartDrive Systems.

Target audience

This course module mainly targets graduate students in the field of transportation human factors/traffic safety analysis, and could be part of a graduate course on these topics. However, given the strong industrial interest in this topic, it is possible that the course could also be given in an industrial context, targeting, for example, insurance or fleet management professionals.

Appendix D. Data Dictionary of Variables

Behavior-based Predictive Safety Analytics – Driver Behaviors and Outcomes
Data Dictionary of Variables

Variable Name	Variable Label	Brief Notation	Data Type	Minimum	Maximum
driverID	Driver ID	Driver ID	Discrete	3	5081966
DBQ1	Driver Behavior Questionnaire (DBQ) Factor 1: Errors	Questionnaire data	Continuous	1	3.1
DBQ2	DBQ Factor 2: Violations	Questionnaire data	Continuous	1	5.25
DBQ3	DBQ Factor 3: Slips/Lapses	Questionnaire data	Continuous	1	5.33
RP1_Minor	Risk Perception Factor 1: Minor Offenses	Questionnaire data	Continuous	1	7
RP2_Major	Risk Perception Factor 2: Major Offenses	Questionnaire data	Continuous	1	7
RP3_Dist	Risk Perception Factor 3: Distraction-related Behaviors	Questionnaire data	Continuous	1	7
studyLength	Length of study in months	Whole participation	Continuous	210	1055
percRemovedTrips	Percent of removed trips (due to various reasons)	Whole participation	Continuous	0.0291	0.9892
DistanceInMile	Distance driven in miles for duration of study	Whole participation	Continuous	7.923	59274.452
DurationInHour	Duration driven in hours for duration of study	Whole participation	Continuous	0.357	1317.747
numTrips	Number of trips for duration of study	Whole participation	Continuous	7	7156
DistanceInMile_six	Distance driven in miles for six month study period	Six months participation	Continuous	0	17905.208
DurationInHour_six	Duration driven in hours for six month study period	Six months participation	Continuous	0	537.811
numTrips_six	Number of trips driven for six month study period	Six months participation	Continuous	0	3220
Crash_six	Number of crashes for six month study period	Six months participation	Continuous	0	10

Variable Name	Variable Label	Brief Notation	Data Type	Minimum	Maximum
NearCrash_six	Number of nearcrashes for six month study period	Six months participation	Continuous	0	19
atFaultCrash_six	Number of at fault crashes for six month study period	Six months participation	Continuous	0	10
atFaultNearCrash_six	Number of at fault near crashes for six month study period	Six months participation	Continuous	0	15
Severity1Crash_six	Number of severity 1 crashes for six month study period	Six months participation	Continuous	0	1
Severity2Crash_six	Number of severity 2 crashes for six month study period	Six months participation	Continuous	0	2
RearEndStriking_six	Number of rear end strikings for six month study period	Six months participation	Continuous	0	14
atFaultCNC_six	Number of at fault crash and near crashes for six month study period	Six months participation	Continuous	0	16
CNC_six	Number of crash and near crashes for six month study period	Six months participation	Continuous	0	20
atFaultCNC_six_Bin	Presence or absence of an at fault crash / near crash during the six month study period	Six months participation; Binary DV	Continuous	0	1
CNC_six_Bin	Presence or absence of a crash / near crash during the six month study period	Six months participation; Binary DV	Continuous	0	1
atFaultCrash_six_Bin	Presence or absence of an at fault crash during the six month study period	Six months participation; Binary DV	Continuous	0	1
Crash_six_Bin	Presence or absence of a crash during the six month study period	Six months participation; Binary DV	Continuous	0	1
MRate_atFaultCNC_six	Rate of at fault crash / near crashes by mileage for six month study period	Six months participation; # events per mile	Continuous	0	0.017
MRate_CNC_six	Rate of crash / near crashes by mileage for six month study period	Six months participation; # events per mile	Continuous	0	0.017
MRate_atFaultCrash_six	Rate of at fault crashes by mileage for six month study period	Six months participation; # events per mile	Continuous	0	0.004

Variable Name	Variable Label	Brief Notation	Data Type	Minimum	Maximum
MRate_Crash_six	Rate of crashes by mileage for six month study period	Six months participation; # events per mile	Continuous	0	0.004
MRate_hardstart_six_24	Rate of hard starts by mileage for six month study period given the threshold	Six months participation; accelX > 0.24g	Continuous	0	3.166
MRate_hardstart_six_25	Rate of hard starts by mileage for six month study period given the threshold	Six months participation; accelX > 0.25g	Continuous	0	2.870
MRate_hardstart_six_30	Rate of hard starts by mileage for six month study period given the threshold	Six months participation; accelX > 0.30g	Continuous	0	1.979
MRate_hardstart_six_33	Rate of hard starts by mileage for six month study period given the threshold	Six months participation; accelX > 0.33g	Continuous	0	1.435
MRate_hardstop_six_24	Rate of hard stops by mileage for six month study period given the threshold	Six months participation; accelX < -0.24g	Continuous	0	3.247
MRate_hardstop_six_29	Rate of hard stops by mileage for six month study period given the threshold	Six months participation; accelX < -0.29g	Continuous	0	2.099
MRate_hardstop_six_30	Rate of hard stops by mileage for six month study period given the threshold	Six months participation; accelX < -0.30g	Continuous	0	2.301
MRate_hardleftTurn_six_29	Rate of hard left turns by mileage for six month study period given the threshold	Six months participation; accelY < -0.29g	Continuous	0	1.855
MRate_hardleftTurn_six_30	Rate of hard left turns by mileage for six month study period given the threshold	Six months participation; accelY < -0.30g	Continuous	0	1.637
MRate_hardleftTurn_six_33	Rate of hard left turns by mileage for six month study period given the threshold	Six months participation; accelY < -0.33g	Continuous	0	0.995
MRate_hardleftTurn_six_38	Rate of hard left turns by mileage for six month study period given the threshold	Six months participation; accelY < -0.38g	Continuous	0	0.580
MRate_hardleftTurn_six_42	Rate of hard left turns by mileage for six month study period given the threshold	Six months participation; accelY < -0.42g	Continuous	0	0.364

Variable Name	Variable Label	Brief Notation	Data Type	Minimum	Maximum
MRate_hardrightTurn_six_27	Rate of hard right turns by mileage for six month study period given the threshold	Six months participation; accelY > 0.27g	Continuous	0	2.896
MRate_hardrightTurn_six_29	Rate of hard right turns by mileage for six month study period given the threshold	Six months participation; accelY > 0.29g	Continuous	0	2.413
MRate_hardrightTurn_six_30	Rate of hard right turns by mileage for six month study period given the threshold	Six months participation; accelY > 0.30g	Continuous	0	2.413
MRate_hardrightTurn_six_35	Rate of hard right turns by mileage for six month study period given the threshold	Six months participation; accelY > 0.35g	Continuous	0	0.873
MRate_hardrightTurn_six_44	Rate of hard right turns by mileage for six month study period given the threshold	Six months participation; accelY > 0.44g	Continuous	0	0.280
MRate_hardleftYaw_six_24	Rate of hard ups by mileage for six month study period given the threshold	Six months participation; gyroZ < -24deg/sec	Continuous	0	3.233
MRate_hardleftYaw_six_46	Rate of hard ups by mileage for six month study period given the threshold	Six months participation; gyroZ < -46deg/sec	Continuous	0	0.527
MRate_hardleftYaw_six_41	Rate of hard ups by mileage for six month study period given the threshold	Six months participation; gyroZ < -41deg/sec	Continuous	0	0.760
MRate_hardrightYaw_six_19	Rate of hard downs by mileage for six month study period given the threshold	Six months participation; gyroZ > 19deg/sec	Continuous	0	3.861
MRate_hardrightYaw_six_22	Rate of hard downs by mileage for six month study period given the threshold	Six months participation; gyroZ > 22deg/sec	Continuous	0	2.413
MRate_hardrightYaw_six_37	Rate of hard downs by mileage for six month study period given the threshold	Six months participation; gyroZ > 37deg/sec	Continuous	0	0.372
MRate_hardrightYaw_six_38	Rate of hard downs by mileage for six month study period given the threshold	Six months participation; gyroZ > 38deg/sec	Continuous	0	0.351

*Note. A value of "NA" is used for missing data.