09-UT-006

The National Surface Transportation Safety Center for Excellence

Modeling 100-Car Safety Events: A Case-Based Approach for Analyzing Naturalistic Driving Data

Final Report

Feng Guo • Jonathan Hankey

Submitted: September 30, 2009

Lighting | Technology
Fatigue | Aging

Housed at the Virginia Tech Transportation Institute
3500 Transportation Research Plaza • Blacksburg, Virginia 24061

ACKNOWLEDGMENTS

The authors of this report would like to acknowledge the support of the stakeholders of the National Surface Transportation Safety Center for Excellence (NSTSCE): Tom Dingus from the Virginia Tech Transportation Institute, Richard Deering from General Motors Corporation, Carl Andersen from the Federal Highway Administration (FHWA), and Gary Allen from the Virginia Department of Transportation and the Virginia Transportation Research Council.

The NSTSCE stakeholders have jointly funded this research for the purpose of developing and disseminating advanced transportation safety techniques and innovations.

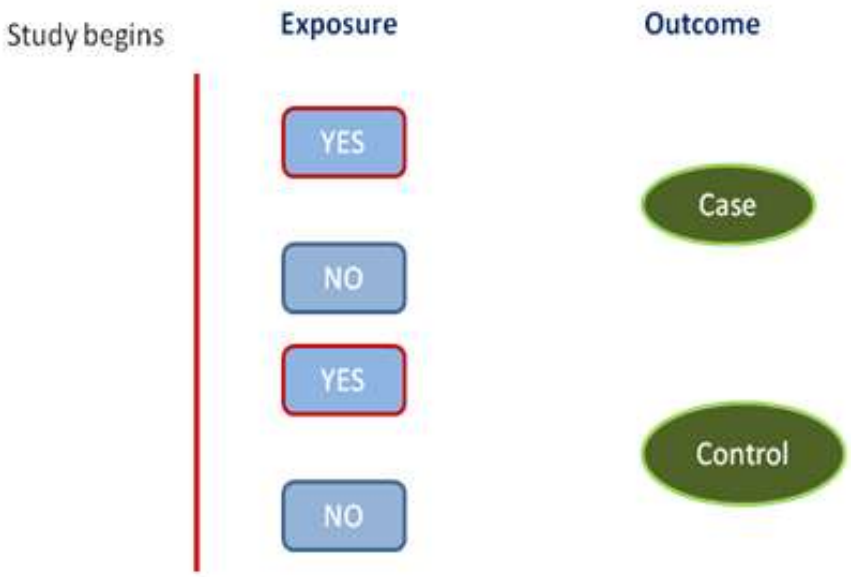
EXECUTIVE SUMMARY

Naturalistic driving study is an innovative way of investigating traffic safety and driving behaviors.[1] The method is characterized by instrumenting participant vehicles with data acquisition systems (DAS) that include cameras and various sensors to continuously monitor the driving process. This type of study can record detailed vehicle kinematic information as well as traffic conditions with advanced instruments such as radar. The rich information collected by naturalistic driving study provides numerous advantages over the traditional accident-database-based analyses or driving-simulator-based studies. However, the complicated data collection process also demands novel approaches for data analyses and modeling. This study developed an integrated framework for modeling the safety outcomes of naturalistic driving studies and addressed several critical methodological issues. Specifically, the following research questions were addressed: 1) how to extract exposure information for safety events and baselines (the study design), 2) how to measure and interpret safety risks, and 3) how to statistically model safety risks.

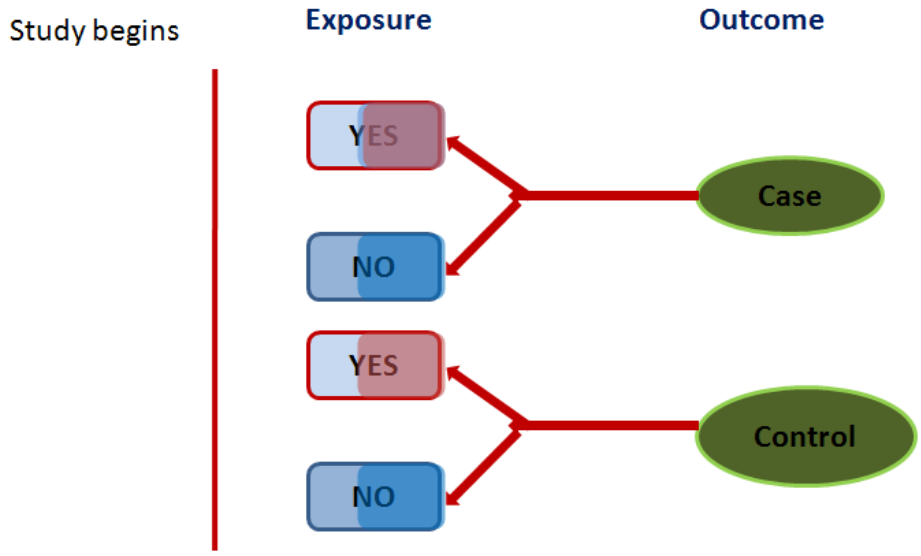
The proposed method was applied to the 100-Car Study.[1] A total random baseline sampling scheme was adopted with a sample size of 17,344. Two alternative statistical models, the generalized estimation equation and mixed-effect logistic regression, were used to incorporate driver-specific correlations and adjust for potential confounding effects. The results indicate a certain level of discrepancy between the model-based approaches and the crude odds ratios.

The Study Design and Measure of Risk

The study design is concerned with how the safety outcomes are identified and how exposure information is extracted. The data collection process of a naturalistic driving study is prospective and is similar to a cohort study; but safety event and baseline identification follows a case-control design. Thus the naturalistic driving study is analogous to the case-cohort type study as illustrated in figure 1. The case-cohort is a two-stage study design in which the first step is to collect and save all relevant data and the second step is to extract information from the saved data for analyses. The case-cohort method combines the merits of both cohort and case-control studies. It is less prone to bias than a case-control study but is more efficient than a cohort study. As shown by this research, the case-based approach is the preferred method for analyzing naturalistic driving data unless full automated data reduction techniques are available.



(a) Case-cohort study data collection step 1: exposure information not extracted



(b) Case-cohort study step 2: extract exposure information for case and control

Figure 1. Diagram. Case-cohort study.

There are three commonly used risk measures: risk ratio, odds ratio, and risk rate ratio (RRR). For most factors of interest, risk rate and RRR as measured by number of events per unit of exposure are most appropriate for naturalistic driving study. However, the RRR is difficult to calculate due to the high cost of accurately extracting exposure duration information. This study proposed to approximate RRR by odds ratio. The approximation requires a combination of appropriate baseline sampling methods and statistical models. It was shown that with a total

random baseline sampling method, the odds ratio will approximate the rate ratio. The sampling method is illustrated in figure 2, where PT+ and PT- is the exposure duration (person-time) for two exposure levels "+" and "-"; and B and D are the number of baseline line samples for each period. The random sampling method satisfies the condition $\frac{B}{D} = \frac{PT+}{PT-}$, which is the key to odds ratio to rate ratio approximation. A stratified random re-sampling method, which shares the same properties as the total random sampling method, was implemented to combine an existing baseline data set and a new data reduction to generate a total of 17,344 baseline samples.

Figure 2. Diagram. Random sampling scheme.

Statistical Modeling and Results

The statistical analyses focused on two issues: 1) to incorporate the correlation among observations from the same driver (the driver-specific correlation), and 2) to adjust for confounding effects through modeling. Two logistic-regression-based models, the Generalized Estimation Equation method (GEE) and mixed effect logistic regression, were adopted to address those issues. The models were applied to both the crashes and near-crashes. The modeling results are summarized in table 1 and table 2 and illustrated in figure 3 and figure 4.

Table 1. Modeling results for crashes.

	GEE Model			Random Effect Model			Contingency Table: Crude Odds Ratio		
Factors	**Odds Ratio**	**95% CI Low**	**95% CI High**	**Odds Ratio**	**95% CI Low**	**95% CI High**	**Odds Ratio**	**95% CI Low**	**95% CI High**
Drowsy	**6.35**	3.38	11.91	**6.31**	3.30	12.08	**7.12**	3.94	12.87
Weather: Inferior versus Normal	2.17	0.79	6.01	2.14	0.82	5.62	1.80	0.89	3.63
Road Surface: Other versus Dry	**4.81**	1.98	11.71	**4.79**	2.27	10.11	**3.10**	1.81	5.32
Lighting: Other versus Day	1.04	0.58	1.86	0.97	0.57	1.65	1.41	0.86	2.29
LOS B versus A	**0.42**	0.23	0.76	**0.42**	0.23	0.79	**0.42**	0.23	0.77
LOS C versus A	0.89	0.38	2.08	0.89	0.39	2.02	0.89	0.40	1.99
LOS DEF versus A	1.83	0.67	5.03	1.98	0.84	4.65	**2.47**	1.10	5.54
Distraction: Complex versus Non	**3.51**	1.18	10.41	**3.40**	1.36	8.51	**3.31**	1.4	7.82
Distraction: Moderate versus Non	0.65	0.25	1.64	0.68	0.32	1.42	0.64	0.31	1.32
Distraction: Simple versus Non	0.54	0.26	1.11	0.51	0.26	1.01	**0.49**	0.25	0.96
Junction versus Non-Junction	**5.89**	3.51	9.86	**6.05**	3.71	9.85	**5.38**	3.35	8.64

Table 2. Modeling results for near-crashes.

	GEE Model			Mixed Effect Model			Contingency Table: Crude Odds Ratio		
Factors	**Odds Ratio**	**95% CI Low**	**95% CI High**	**Odds Ratio**	**95% CI Low**	**95% CI High**	**Odds Ratio**	**95% CI Low**	**95% CI High**
Drowsy	**3.67**	2.69	5.01	**3.5282**	2.7040	4.6038	**4.08**	3.25	5.13
Weather: Inferior versus Normal	**1.94**	1.22	3.10	**1.8213**	1.1792	2.8129	1.06	0.82	1.39
Road Surface: Other versus Dry	**2.17**	1.38	3.40	**2.0419**	1.4356	2.9042	**1.43**	1.15	1.78
Lighting: Other versus Day	1.17	0.96	1.43	1.0372	0.8714	1.2344	1.12	1.03	1.40
LOS* B versus A	1.18	0.98	1.43	1.1790	0.9724	1.4295	1.18	0.98	1.41
LOS C versus A	**4.06**	3.17	5.20	**3.9243**	3.1674	4.8620	**4.07**	3.35	4.97
LOS DEF versus A	**4.99**	3.63	6.87	**5.3586**	4.0405	7.1068	**5.46**	4.23	7.04
Distraction: Complex versus Non	**2.02**	1.30	3.12	**1.9514**	1.3601	2.7997	**1.85**	1.34	2.57
Distraction: Moderate versus Non	**0.48**	0.37	0.63	**0.4844**	0.3757	0.6244	**0.46**	0.36	0.58
Distraction: Simple versus Non	**0.33**	0.25	0.42	**0.2976**	0.2332	0.3799	**0.30**	0.24	0.38
Junction versus Non-Junction	**3.36**	2.74	4.12	**3.4980**	2.9576	4.1371	**3.24**	2.78	3.78

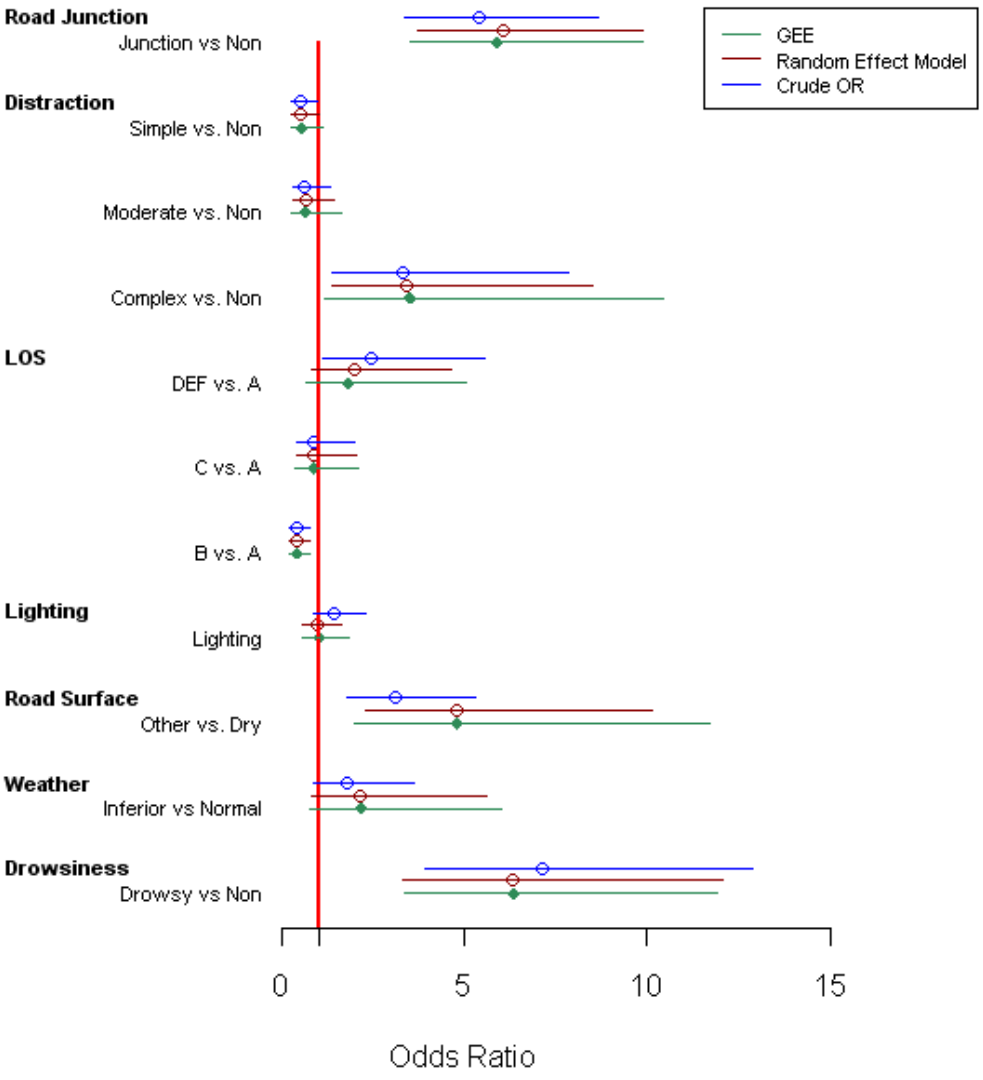


Figure 3. Graph. Crash odds ratios.

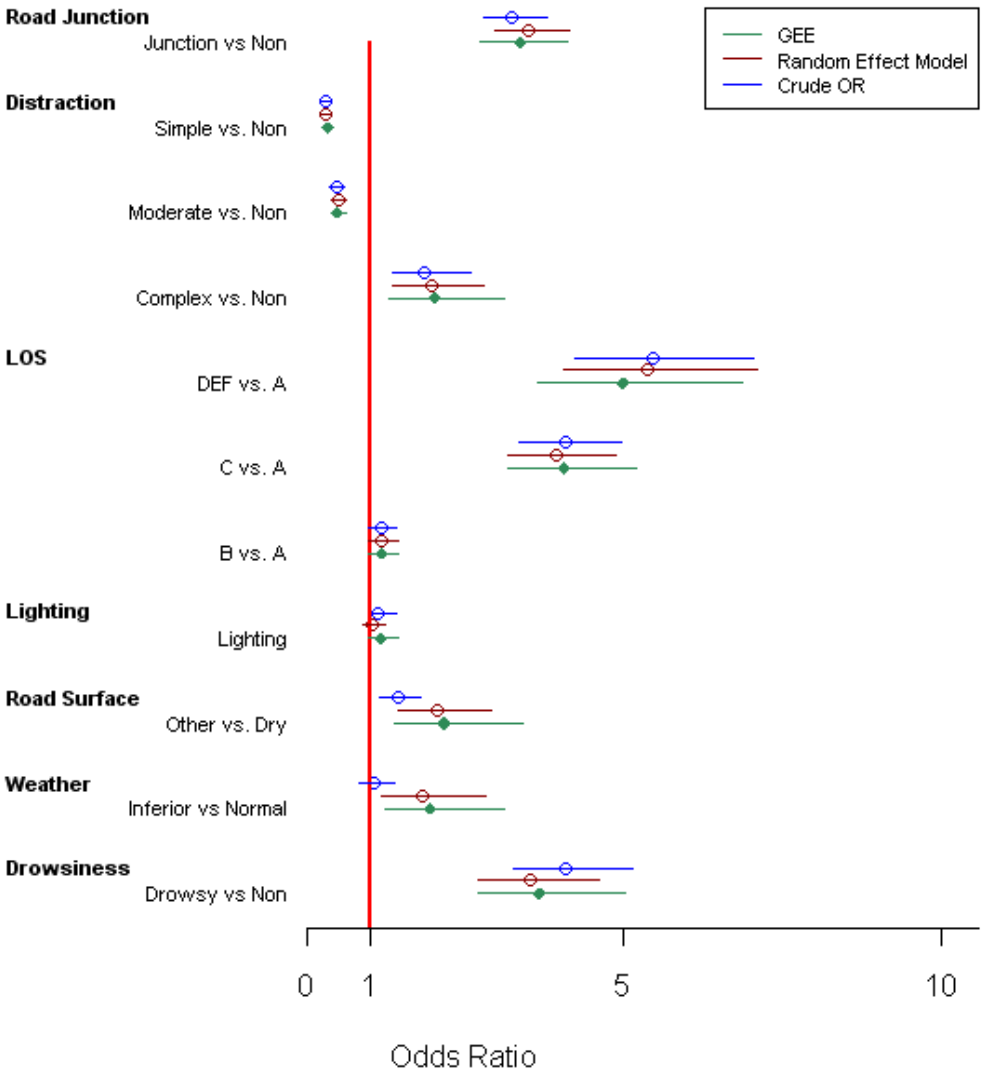


Figure 4. Graph. Near-crash odds ratios.

The main findings are summarized as follows:

- There are some discrepancies among results from the GEE, the mixed effect model, and the crude odds ratio estimation. The confidence intervals of the crude odds ratio are in general narrower than those from the two model-based approaches. However, this is considered as overly optimistic given that it ignores the driver-specific correlation and fails to adjust for potential confounding factors.
- The GEE and mixed effect model can be used to evaluate the level of correlations among observations from the same driver. The GEE analysis indicates that the marginal correlations among observations are weak. The mixed effect logistic regression model

shows moderate variations among drivers. This result is consistent with the fact that a small number of drivers contribute a large proportion of the safety events. The mixed effect model is preferred because of its connection with this individual variation.

- The odds ratios for crashes are consistently larger than for near-crashes. On the other side, the precision of the estimation for near-crashes, as measured by the length of the confidence interval, is substantially better than that for crashes. This result has significant implications for using near-crashes as a safety surrogate for crashes.
- Drowsiness will increase the risk of both crashes and near-crashes substantially (sixfold increase for crash and threefold increase for near-crash).
- Inferior weather conditions will significantly increase the risk of near-crashes and also show a considerable impact on crashes.
- Traffic densities show distinct patterns for crashes and near-crashes. For crash risks, the Levels of Service [LOS] B and C, which represent a moderate level of interactions among vehicles, are not necessarily more dangerous than free flow condition (LOS A). This could be attributed to the increased driver vigilance. However, LOS B and LOS C are associated with a high risk of near-crash. For both crashes and near-crashes, high traffic density (LOS DEF) will lead to higher risks.
- Complex secondary tasks will increase the risk of crashes by more than three times and the risk of near-crashes by two times. The simple and moderate secondary tasks, on the other hand, show some level of protective effects for both crashes and near-crashes.
- The highway junction is substantially more dangerous than the non-junction highway segment with a sixfold increase in crash risk and a threefold increase for near-crash.

This study focuses on analysis methodology issues for naturalistic driving study. The framework developed provides a solid theoretical justification for the case-based study method in naturalistic driving studies. It addressed critical issues on how to measure risk, how to conduct data reduction, and how to model the reduced data statistically. The framework can be directly applied to evaluate time-variant risk factors such as driver behavior and driving environmental factors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

CI	Confidence Interval
DAS	Data Acquisition System
EOR	Exposure Odds Ratio
ESC	Electronic Stability Control
GEE	Generalized Estimation Equation
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Effect Model
GPS	Global Positioning System
LOS	Level of Service
PDA	Personal Digital Assistant
ROR	Risk Odds Ratio
RR	Risk Ratio
RRR	Risk Rate Ratio
VTTI	Virginia Tech Transportation Institute

CHAPTER 1. INTRODUCTION

Naturalistic driving study is an innovative way of investigating traffic safety and driving behaviors. [1] The method is characterized by instrumenting participant vehicles with data acquisition systems (DAS) that include cameras and various sensors to continuously monitor the driving process. This type of study can record detailed vehicle kinematic information and traffic conditions with advanced instruments such as radar. The rich information collected by naturalistic driving study provides numerous advantages over the traditional accident-database-based analyses or driving-simulator-based studies. However, the complicated data collection process also demands novel approaches for data analyses and modeling. This study developed an integrated framework for modeling the safety outcomes of naturalistic driving studies and addressed several critical methodological issues. Specifically, the following research questions were addressed: 1) how to extract exposure information for safety events and baselines (the study design), 2) how to measure and interpret safety risks, and 3) how to statistically model safety risks.

Highway crashes are one of the leading causes of death in the United States; there are more than 40,000 deaths and approximately 2.5 million injuries annually that result from highway crashes.[2] As a result, safety has been a focus of transportation research for the last decade. Before the emergence of advanced data collection methods, accident databases and /police reports have been the main sources of traffic accident information. There have been numerous efforts to establish the relationship between accident frequency and potential risk factors such as highway geometric features and traffic characteristics. For this purpose, the accident data are commonly aggregated by intersection or highway segment. Comparable with aggregated accident data, counting data models, such as Poisson and negative binomial regression models, have been the mainstream modeling techniques.[3,4] Recently, more sophisticated models were developed incorporating spatial and temporal correlation and using full Bayesian framework.[5,6]

One inherent drawback for aggregated analysis is that a large proportion of information for individual crashes was lost during aggregation. Only those characteristics shared by all crashes within an aggregation stratum can be kept. For example, in intersection safety analysis, the response is the number of crashes for each intersection. Only those properties shared by all crashes at a given intersection (such as intersection design and traffic characteristics) can be incorporated into analyses. The risk factors for each crash, such as driver age, gender, vehicle type, etc., are different in most cases and thus cannot be considered as the attributes of aggregated crash counts.

Only limited studies have considered traffic safety at the discrete/individual crash level.[7] One critical issue in individual crash-based analysis is to find a proper control group and compare safety events with the control group. In evaluating the effectiveness of electronic stability control (ESC), Dang[7] used the crashes that were not directly related to ESC as the control group; e.g., crashes involving a parked vehicle, a backing up vehicle, vehicles entering/leaving a parking lot, and vehicles with a speed lower than 10 mi/h. The quantitative comparison of ESC frequencies in ESC-related crashes and non-ESC-related crashes was then used to evaluate the effect of ESC. The individual crash-based analysis can incorporate more information than the aggregated method. For this type of analysis, the appropriateness of the control group directly determines validity of the study and sometimes can be difficult to define.

The details and quality of data that naturalistic driving studies have provided are unprecedented. The video image and kinematic measures can provide not only the exact driving behavior, vehicle kinematic, and driving environmental information, but also the sequence and precise time stamp for each sub-event. This high resolution information is not readily available from the accident database. The traditional accident reconstruction techniques can recover some vehicle kinematic information but are often fragmentary and lack the exact time stamp for the sequence of events. To make maximum use of the rich information collected, individual safety-event-based analysis is preferred to the aggregated method.

Drivers' behavior is the main contributor to traffic safety events. However, accurately retrieving driver behaviors from post-accident reconstruction is challenging if not impossible. The records from accident databases are primarily based on the statements from driver(s)/witness(es) and that information is often fragmentary and based on witness' perception and memory.[8] Accident reconstruction may suggest the driver's behaviors before/during the crash but the results often tend to be speculative. For this reason, many driver behavior studies are conducted in a controlled experimental environment or on a simulator. However, the driver's behavior in a simulator and in a controlled environment may substantially differ from behavior during natural driving conditions. Therefore, simulator/controlled experiments cannot completely replace the field data collection. The naturalistic driving study can overcome these challenges and provides an opportunity to quantitatively evaluate the safety impact of drivers' behavior under natural driving conditions.

Driver behavior, along with many other factors such as weather and traffic conditions, is time-variant in that its status constantly changes over time. Models based on aggregated data are difficult to be implemented for those time-variant exposures because the aggregation requires accurate exposure duration information for each factor, e.g., duration of each period when the driver is drowsy. To extract this information is cost-prohibitive using the current data reduction method for naturalistic driving. The problem is further complicated when multiple factors are considered. The lack of exact exposure duration information could be an obstacle for Poisson and negative binomial models, which are based on aggregated data and require exposure duration. Therefore, the individual crash/discrete-based analysis method is preferred for analyzing naturalistic driving data.

In this study, an integrated analysis framework was developed for modeling the safety outcomes of a naturalistic driving study based on individual safety events. The focus is on time-variant risk factors. The main components of this framework are addressed in the subsequent chapters. The overall structure of the report is as follows.

- Chapter 2: the study design and some typical study design methods are introduced. The merits of each design and their relationship with naturalistic driving study are discussed. In addition, various measures of risk and their relationship with the study design are discussed in detail.

- Chapter 3: a random sampling scheme is introduced and implemented for the 100-Car Naturalistic Driving Study (100-Car Study).

- Chapter 4: several alternative models are introduced for the analysis of reduced data. The method was applied to the reduced data from the 100-Car Study.

- Chapter 5: summary and discussion.

CHAPTER 2. STUDY DESIGN AND BASELINE SAMPLING SCHEME

One primary goal of naturalistic driving study is to identify and evaluate factors with significant impacts on traffic safety, which is typically measured by number of crashes or crash surrogates. The study design is a critical component of naturalistic driving study. It guides the overall data collection and analyses and essentially determines the validity of a study. For a naturalistic driving study design, three major questions shall be addressed: 1) what is characteristic of the overall study design, 2) how to measure a risk, and 3) how the baseline information should be extracted. In this chapter, several typical epidemiology study design methods were introduced and their relationship with naturalistic driving study was investigated.

STUDY DESIGN

Naturalistic driving study investigates the factors that affect traffic safety. This is a direct analogy to epidemiology study whose focus is to evaluate factors affecting public health. Therefore, the naturalistic driving study design is similar to epidemiology design. The framework developed in this research is built upon epidemiology methods. The study design determines how the exposure information and health/safety events should be collected/extracted. To a large extent, it also determines how the data should be analyzed. The naturalistic study, by definition, does not involve direct intervention in the driving process; thus, it belongs to the observational study category. Unlike experimental studies in which exposure/treatment can be controlled by researchers, the participants in an observational study decide their own exposure status. Similarly, drivers in a naturalistic driving study determine their own driving behaviors. For example, a driver can make his/her decision on whether to use a cell phone during driving. Besides driving behaviors, the exposure status of potential risk factors such as environmental and traffic conditions cannot be controlled by researchers. Therefore, the study framework was developed based on observational study methods.

Because of the inability to control exposure status, observational studies are more prone to bias than experimental studies. In an experimental study, interaction and confounding factors can be addressed through randomization or appropriate assignment of exposure status. In observational studies, however, there is no guarantee that the effects of particular risk factors will be isolated from other factors. For example, texting might always be associated with eyes-off-road and, in this case, the effects of texting cannot be separated from the effect of eyes-off-road. Appropriate study design and data analysis methods can address those issues and are critical components of an observational study.

In a naturalistic driving study, the participants drive vehicles in a non-obstructive driving environment and their driving behaviors, environmental factors, vehicle kinematic information, and traffic conditions are continuously recorded through multiple video cameras and various instruments. Safety events (such as crashes, near-crashes, and critical incidents) are identified through kinematic signatures of the vehicle and confirmed through visual inspection for video recordings. The main objective of studying these safety events is to identify factors that have a significant impact on traffic safety. This is done by comparing the exposure status of risk factors that are present before/during safety events and during normal driving conditions. Following the convention of epidemiology research, the safety outcomes, i.e., crash and near-crash, are used interchangeably with cases; and the factors that might contribute to safety are used

interchangeably with exposures. Besides cases, the exposure status under normal driving conditions is also required, which is called baselines/controls. The quantitative evaluation of risk will be conducted through the comparison of the exposure status between cases and baselines.

The study design concerns how cases and baselines are selected and how exposure information is extracted. Depending on the order in which exposure and cases are identified and the timeline of a study, there are three basic types of studies: the cohort study, the case-control study, and the cross-sectional study. The basic setup for each study design and their characteristics are introduced as following.

The Cohort Study

In a cohort study, exposure information is identified first, and safety/disease outcomes (either case or non-case) are identified subsequently. A cohort is a group of individuals with similar exposure status. For example, in traffic safety studies, there could be a teenage driver cohort and an adult driver cohort. These two cohorts will be followed through the study period and the safety outcomes for each cohort (i.e., crash or no-crash) are identified through the course of study. For time-variant exposures such as weather and traffic conditions, the membership of cohort will change over time. A *dynamic cohort* is used to refer to a group of individuals with the same exposure status at a given time point/period. As will become apparent later, the majority of the risk factors in naturalistic driving study rely on the dynamic cohort concept. A schematic plot of cohort design is shown in figure 5.

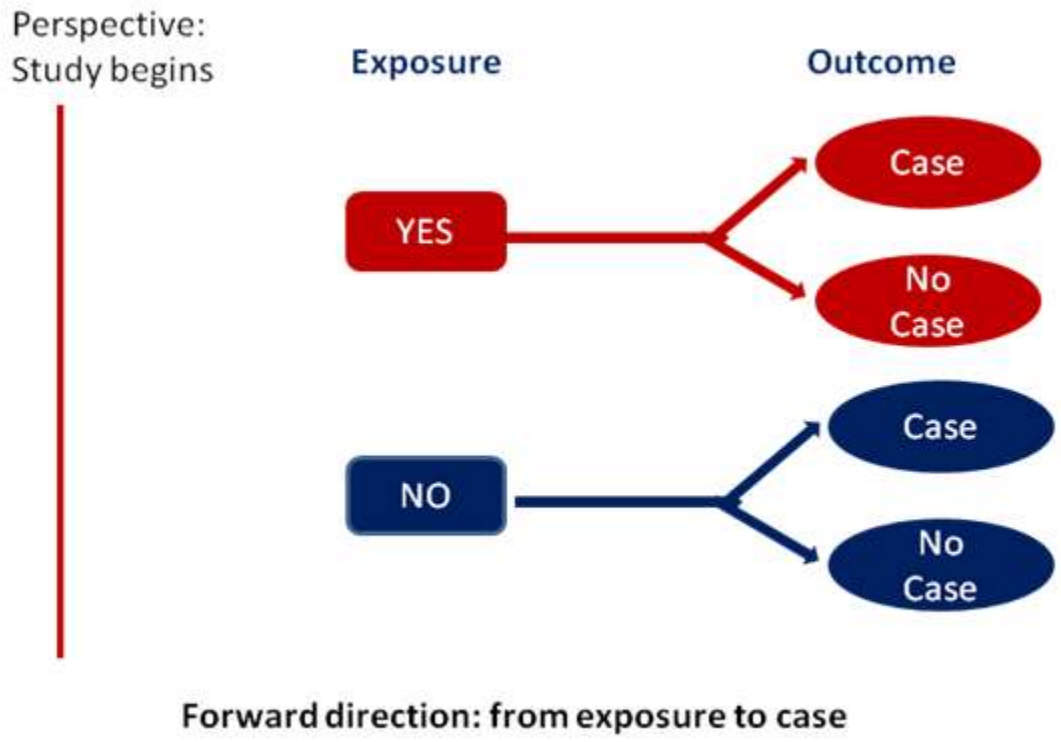


Figure 5. Diagram. Cohort study.

Compared to other types of observational studies, the cohort study is the least prone to bias. The direction of the study, i.e., from exposure to health outcomes, allows the risk to be evaluated directly. However, the cohort study usually requires a lengthy data collection period and the cost is high. For example, the Framingham Heart Study has lasted for decades.[9] When historical exposure data are available, the cohort study can be relatively time- and cost-efficient and is commonly used in occupational disease studies.

The Case-control Study

In a case-control study, cases and controls are identified first and their corresponding exposure status is subsequently extracted. A group of observations, i.e., controls, are selected to represent the general exposure status of the study population. In traffic safety studies, the cases will be the drivers who had experienced a safety event or the events themselves. The controls are the drivers without safety events or segments of normal driving process. The risk factors are evaluated by comparing the exposure frequencies for cases and for controls. A schematic plot for case-control study is shown in figure 6.

Depending on how cases are defined, the controls can either be drivers without any safety events or a short period of normal driving process. The selection of controls will determine the validity of a case-control study to a large extent. The general principle for control selection is that controls should reflect the characteristics of the source population from which cases are derived. How to implement this principle in practice, however, is context-dependent. Note that in a cohort study the cases are always derived from source population and thus are less prone to bias than a case-control study. Improper control selection can lead to invalid conclusions in a case-control study. A thorough consideration for baseline selection scheme is critical for the success of a study.

Another disadvantage for case-control study is that the design does not allow direct evaluation of risk. This is a serious weakness but can be addressed by using appropriate control selection scheme and risk measures. This research uses a combination of baseline sampling method and statistical analyses to address the risk measurement issue. The details will be discussed in a later part of this report.

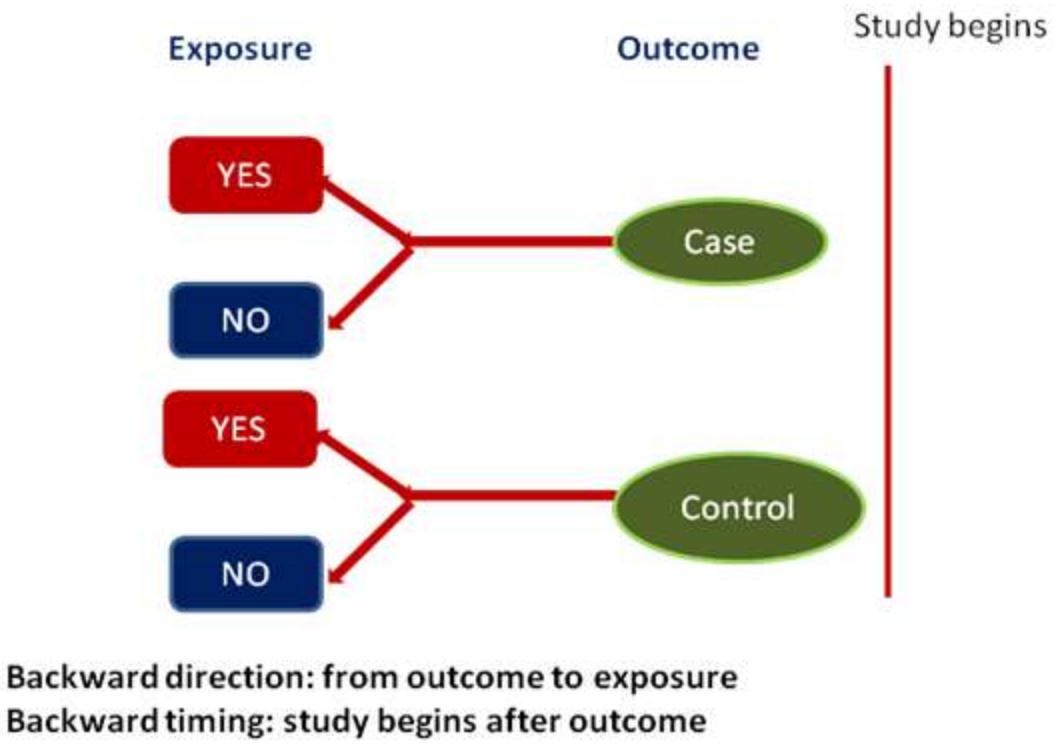


Figure 6. Diagram. Case-control study.

The Cross-sectional Study

In a cross-sectional study, the cases and baselines as well as the corresponding exposure information are collected at a particular time point (or time period). Many traffic safety studies belong to this category. For example, the crash happened during a specific time period and the corresponding traffic conditions and infrastructure characteristics in the same period can be collected in a cross-sectional study. A regression-based analysis is commonly used to connect safety outcomes with the exposures status. The cross-sectional method works best for those factors that do not change for a long period of time. In traffic safety studies, the road geometric design and traffic demand characteristics are examples of those measures. Due to the relatively small number of crashes that happen at each location, the duration of the cross-sectional time window is usually several years long. A schematic plot for the cross-sectional study is shown in figure 7.

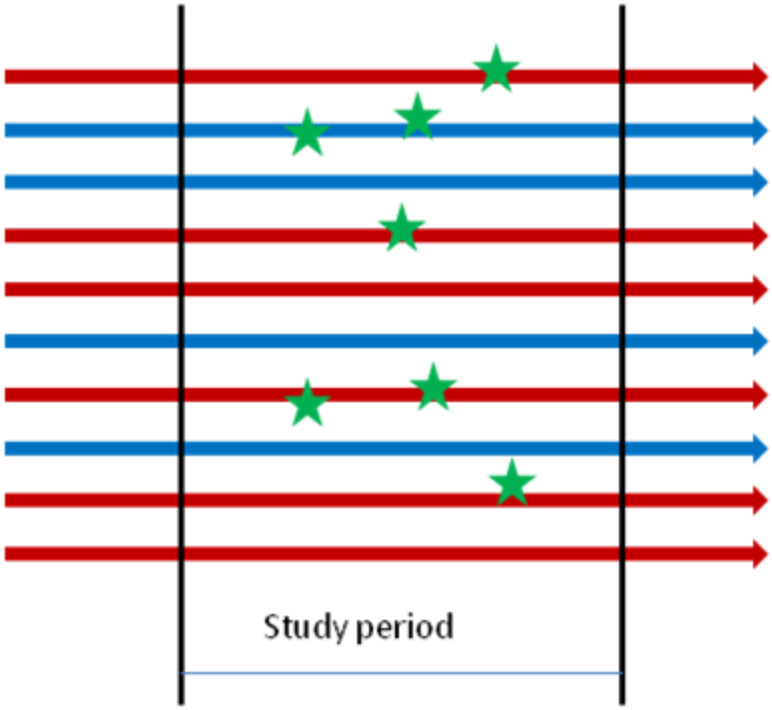


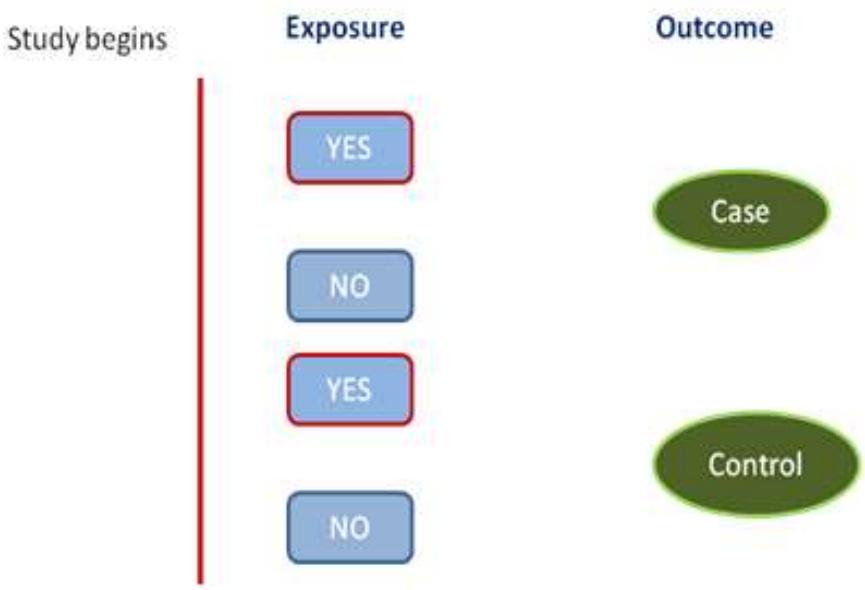
Figure 7. Diagram. Cross-sectional study.

The Case-cohort Study

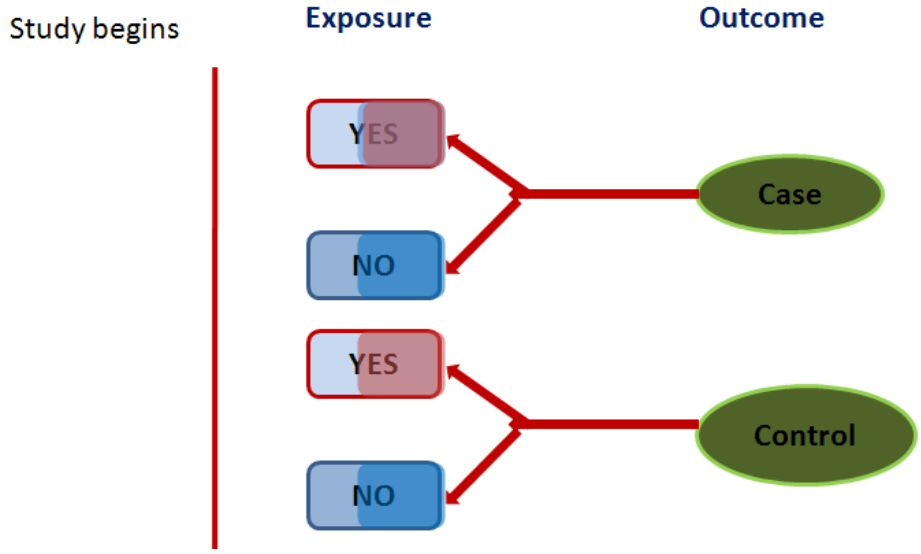
The cohort, case-control, and cross-sectional studies are the three primary observational study designs. In addition to those three methods, several hybrid design methods have been proposed to mitigate the drawback of the individual study design. The case-cohort is a hybrid design that combines the characteristics of both cohort and case-control study. In a case-cohort study, the data collection follows the procedure of a cohort study. However, the exposure status, or the original cohort, is not extracted at the beginning of the study as is the case for typical cohort studies. Instead, the information is "saved" for future analyses. There could be a number of reasons for this approach, e.g., the cost of identifying exposure status is too high, it is technically not practical to identify exposure status for a large number of samples, or the research questions are not full determined at the time of data collection. After the data collection process, the analysis, however, follows that of a case-control study in which the cases/safety outcomes are identified retrospectively from the original cohort. Instead of finding exposure status for all samples, only a subset (i.e., the controls) will be selected from the saved information. By doing this, only a subset of saved data needs to go through the exposure status identification process, thus significantly reducing the corresponding data reduction cost.

The case-cohort method combines the advantages of both cohort and case-control studies. The case-cohort study guarantees the cases are from the study population, thus reducing bias associated with control selection as in a case-control study. At the same time, since only a small proportion of the study population needs to be examined for exposure status, the associated cost is much lower than that of the cohort study. However, the case-cohort cannot totally eliminate the weakness of cohort and control studies, e.g., duration of the data collection cannot be reduced. At the same time, caution is still needed in selecting the control from the study

population; that is, the controls should still represent the characteristics of the general population. As will be discussed later, the selection of control also depends on the risk measures used and the modeling approach. The two-step procedure of case-cohort is illustrated in figure 8.



(a) Case-cohort study data collection step 1: exposure information not extracted



(b) Case-cohort study step 2: extract exposure information for case and control

Figure 8. Diagram. Case-cohort study.

Case-crossover Design

Case-crossover sampling is a matched sampling scheme for which a given number of baseline samples are selected for each case by matching certain conditions. The case-crossover method

requires that baseline samples have the same potential confounding/interaction factors as the case, such as driver, location, time of day, and weather conditions. In a case-crossover, the matching factors for a case are extracted first. The controls are then identified by matching those factors with the case. This procedure guarantees that the case and control have the same exposure for matched factors. Thus, the confounding/interaction factors are controlled through sampling. The case-crossover design is suitable for short exposures with transient effects such as drowsiness and inattention.

There are some disadvantages for case-crossover design. The matching process can be technically difficult. There are situations where no sufficient qualified baselines can be identified. Due to the matching scheme, the observations for each matched set shall be considered as not independent. Therefore, the analysis requires more sophisticated models. Furthermore, the baseline samples from case-crossover can only serve the specific analyses and are difficult to use in other studies. A schematic plot of the case-crossover design is shown in Figure 9.

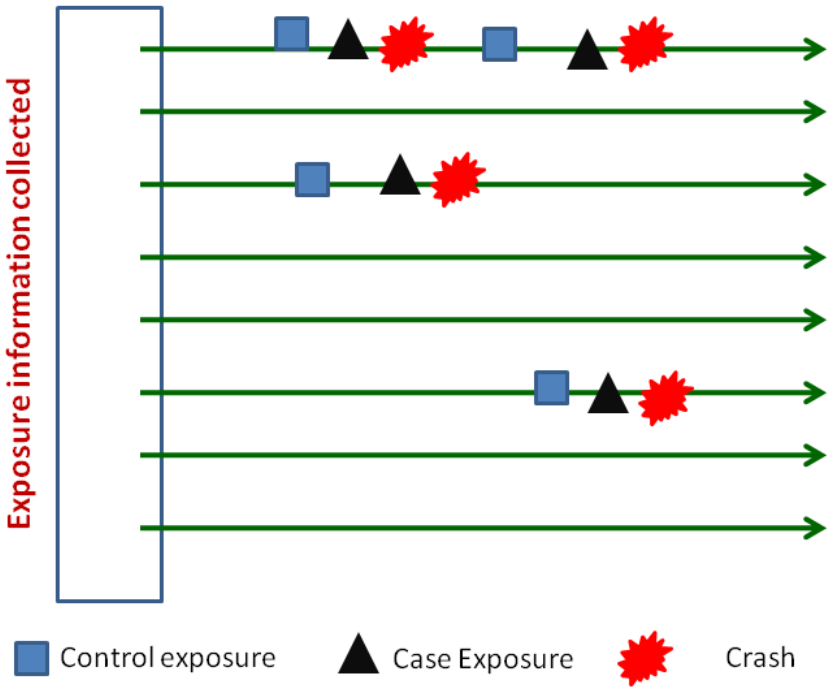


Figure 9. Diagram. Case crossover study.

NATURALISTIC DRIVING STUDY AND ITS DESIGN CHARACTERISTICS

Naturalistic driving study is characterized by its minimized interference with the driving process and the massive amount of information collection. Consequently, the study is an observational-type study. The data collection process can be considered as a plain recording of reality for the next step analysis. The rich information collected by naturalistic driving study allows various research questions to be answered. In term of study design, the naturalistic driving study data collection is perspective-type study but in itself does not constitute one specific study design. The reason is that a study design usually targets specific research questions, which is generally not fully determined at the beginning of data collection. There could be several alternative study designs for a given research question. The thorough information collected through naturalistic driving study provides great flexibility in study design.

There are some distinct characteristics for naturalistic driving studies. The data are collected prospectively; that is, all relevant exposure information will be recorded regardless of future safety outcomes. From this viewpoint, it has the characteristics of a cohort study. The participants for a study are usually fixed and it is a fixed population study. The cohort refers to a group of individuals with identical exposure status. This is easy to define for time-invariant exposures such as driver genders and vehicle type; e.g., a cohort of male drivers and a cohort of sedans. The cohort is difficult to be defined for time-variant exposures such as driver behavior, weather conditions, and traffic conditions. Those exposures will change over time and one participant driver might be in a drowsiness cohort during one period of time and in a non-drowsiness cohort during another time period. The concept of dynamic cohort is used to refer to a cohort of participants who belong to a certain exposure group but might change their membership. A dynamic cohort consists of participants at a certain exposure status for a given time point.

An analysis based on dynamic cohort is commonly measured by risk rate, which requires the knowledge about the duration for each exposure level. While the advances in technology may allow those durations to be calculated automatically in the future, this is obviously cost-prohibitive given that current data reduction relies primarily on manual data reduction by trained data reductionists. Therefore, a standard cohort-type analysis is not practical. Instead, case-based approaches are more appropriate.

The case-based approach follows the general framework of a case-control study design. For naturalistic driving data, the first step is to identify safety outcomes from the continuously recorded kinematic and image information. The current Virginia Tech Transportation Institute (VTTI) practice is to use kinematic triggers (e.g., abnormally high deceleration rate or yaw rate, etc.) to identify segments potentially being related to safety events. A visual inspection is then followed to each trigger to confirm the safety outcome.[10] The safety outcomes are analogous to cases for a case-control study. Similarly, a set of controls needs to be selected. In the context of a naturalistic driving study, the controls are segments of driving records for normal driving conditions. The exposure statuses for both cases and controls are extracted. The comparison of exposure statuses for cases and controls allows quantitative measure of the risk associated with each exposure.

The case-based approach for analyses of risk factors from naturalistic driving study is fundamentally different from a standard cohort study. Although the exposure data were collected in the data collection process, they were not extracted immediately. Instead, the raw data were simply stored and the exact exposure information is extracted later through the data reduction process. This is more analogous to the case-cohort type study. There are several alternative design methods for the case-based approach, including random baseline sampling and the case-crossover design method. The random baseline sampling method was adopted in this report and will be discussed in detail in Chapter 3.

MEASURE THE RISK OF EXPOSURE FACTORS

The output from an observational study can be conveniently arranged in a contingency table form although the interpretation can be dramatically different. For a simple dichotomous exposure factor, the output contingency table has the general form as shown in table 3.

Table 3. Output of observational study.

	E+	**E-**	**Total**
Crash	A	B	A+B
No crash	C	D	C+D
Total	A+C	B+D	A+B+C+D

In the above table, the exposure is assumed to have two levels: E+ and E-. For example, E+ could represent young drivers or inferior weather conditions and E- could represent adult drivers or normal weather conditions. In the first case, the value "A" is the number of young drivers having a crash during the study period and B is the number of adult drivers having a crash. The value C is the number of young drivers without a crash and D is the number of adult drivers without a crash.

The main difference among study designs is how the marginal sample size is determined. For cohort study, the sample size for each cohort is predetermined. That is, the numbers A+C and B+D are predetermined before data collection. For case-control study, the number of controls (i.e., the row marginal) is predetermined; e.g., A+B (the number of cases) is observed from the study and C+D (the number of controls) is predetermined. For cross-sectional study, the total sample size, A+B+C+D, is predetermined. This difference has a significant implication on what types of risk measures can be calculated. The details are discussed in the following section on risk measures.

The ultimate goal of the study is to establish the relationship between exposure and the safety outcome, commonly measured by crashes. At the same time, it is desirable to measure the magnitude of the impacts of risk factors to safety. Thus, a quantitative measure of risk is desired. Depending on the study design and nature of safety outcomes, three different measures can be used: the risk, the odds, and the rate.

Risk

Risk is the probability of crash for a specific factor (or combination of factors) over a specific period of time. As a probability measure, the risk is always between zero (i.e., no risk) and 1.

Risk is commonly evaluated by the relative frequency of cases in each cohort and can be directly calculated from the cohort study. Following table 3, the risk for the E+ cohort is A/(A+C), which is the relative frequency of crash for the E+ group; similarly, the risk for the E- cohort is B/(B+D). The risk has the direct interpretation of the probability of crash given exposure status; e.g., the risk for teen drivers is the probability of teen drivers having a crash.

Note that in a case-control study, the risk cannot be directly calculated since the total number of observations for the exposure group (A+C and B+D) is not predetermined. Instead, the numbers of cases and controls are observed/determined in advance; i.e., A+B and C+D. Therefore, A/(A+C) does not represent the risk for E+ group. The meaningful measure is the relative frequency A/(A+B), which is corresponding to the probability of exposure given that there was a crash. This measure is not as attractive as the risk of crash for a given exposure level. For example, the value A/(A+C) represents the probability that the driver is a teenager given that there is an accident. This is, of course, of less interest for the researcher and the general public than is the risk measure from cohort studies.

The risk is associated with the duration of the study period. For example, the risk of having a crash in 10 years will be much higher than the risk of having a crash in 1 year. For this reason, risk requires accurate information about the time at risk, i.e., the driving time or mileage. This can be challenging for time-variant factors.

Comparison of the relative risk of two exposure levels can be done through the comparison of the risk. Some commonly used measures include:

1. Risk ratio (RR) for cohort study

$$RR = \frac{Risk|E+}{Risk|E-} = \frac{A/(A+C)}{B/(B+D)}$$

The neutral value is 1, which indicates that there is no difference in the risk of the exposure and non-exposure groups. An RR greater than 1 indicates elevated risk and a RR of less than 1 indicates a protective effect or a lessening of risk. Note that this value differs from that based on the case-control study.

2. Risk difference (population attributable risk)

$$\text{Risk difference} = Risk|(E+) - Risk|(E-) = \frac{A}{A+c} - \frac{B}{B+D}$$

A zero value risk difference implies there is no difference for the two exposure levels.

Odds

Odds are another measure of uncertainty. The odds are defined as the ratio of the probability that an event will occur to the probability that an event will not occur; i.e.,

$$odds = \frac{p}{1-p} = \frac{p(Event\ will\ occur)}{p(Event\ will\ NOT\ occur)}$$

For a cohort study with the notation in table 3, the odds for young drivers to have a crash is

$$Odds|E+= \frac{P(crash|E+)}{p(no-crash|E+)} = \frac{\frac{A}{A+C}}{\frac{C}{A+C}} = \frac{A}{C}$$

Similarly, the odds for adult drivers are B/D.

The two exposure levels can be compared using an odds ratio. The odds ratio can be calculated as

$$ROR = \frac{Odds|E+}{Odds|E-} = \frac{A/C}{B/D} = \frac{AD}{BC}$$

Note that the equation above is based on risk measure from cohort study, thus commonly known as the *risk odds ratio (ROR)*.

The odds ratio for case-control study has a different interpretation from the cohort study. Because the total numbers of cases and controls are fixed, the odds are based on the probability of exposure conditioning on case or control; that is, the probability of exposure given a crash/control has happened. The probability of exposure is apparently less attractive than the risk probability in a cohort study. The odds ratio calculation for case-control study is shown below.

$$Odds|case = \frac{P(E+|crash)}{p(E-|crash)} = \frac{\frac{A}{A+B}}{\frac{B}{A+B}} = \frac{A}{B}$$

$$Odds|control = \frac{P(E+|control)}{p(E-|control)} = \frac{\frac{C}{C+D}}{\frac{D}{C+D}} = \frac{C}{D}$$

$$EOR = \frac{Odds|Case}{Odds|Control} = \frac{A/B}{C/D} = \frac{AD}{BC}$$

The odds ratio for case-control is based on the probability of exposure; thus, it is named the *exposure odds ratio (EOR)*. Although there are fundamental differences between risk odds ratio and exposure odds ratio, the formulas are identical. Under appropriate conditions, the exposure odds ratio can be used to approximate the risk ratio.

Risk Rate

The risk and odds ratios are based on probability measures. As discussed previously, the probability has to be considered for a specific period. To compare two exposure levels, the exposure duration for those two levels should be equal or comparable. For example, to compare young drivers and adult drivers, the driving time for each driver should be comparable. This is not necessarily true for naturalistic driving study where different drivers drive different amounts and distances. Furthermore, the driving behaviors and driving environments are constantly

changing over time. For the dynamic cohort defined from time-variant exposures, it is challenging to conduct the comparison based on probability measure. The rate and rate ratio are more appropriate in this context. The risk rate and risk rate ratio (RRR) is defined as

$$Rate|E+= \frac{number\ of\ event\ under\ E+}{Miles\ (time) traveled\ under\ E+}$$

$$Rate|E-= \frac{number\ of\ event\ under\ E-}{Miles\ (time) traveled\ under\ E-}$$

$$Risk\ Rate\ Ratio = \frac{Rate|E+}{Rate|E-}$$

Higher risk rate is associated with increased risks. Similar to risk and odds, the RRR can be used to evaluate a particular factor. For example, the RRR for driver distraction versus no distraction can be used to evaluate the safety impact of driver distraction. Under certain conditions the RRR can be approximated by the EOR, which will be discussed later.

Baseline Exposure Information

The typical first step in analyzing naturalistic driving data is to identify safety events; i.e., crash, near-crash, or critical incident. After the safety events have been identified, data reduction is then conducted to extract information on driver behaviors and driving environments before and during the events. However, as illustrated below, the exposure information for events alone is not sufficient for quantitatively evaluating the safety impact of a risk factor. This is partly due to the stochastic nature of a safety event.

The occurrence of safety outcomes is random: even driving under the influence does not necessarily lead to an a crash every time. It is compelling to use the frequency of an exposure factor before crashes to evaluate its safety impact, in which case a higher exposure frequency would indicate elevated risk for that factor. This seemingly reasonable approach actually does not reflect the true impact of a risk factor due to the lack of baseline driving conditions. The following hypothetical example illustrates this idea. Assume that 100 crashes were identified in a study and it was found that in 95 of them the drivers were listening to the radio and in 5 out of the 100 crashes the drivers were in a severely drowsy condition. In this very-likely-to-happen scenario, one could incorrectly conclude that listening to the radio is more dangerous than severe drowsiness. The reason for this counterintuitive result is that an observed high exposure frequency during crash could be due to its high frequency during normal driving conditions. For example, if it was found that under normal driving conditions drivers will listen to the radio 95% of the time, then the fact that 95 out of l00 people involved in a crash were listening to the radio could be purely due to randomness. On the other hand, if virtually no severe drowsiness occurred under normal driving conditions, the five drowsiness cases would indicate a strong association between crashes and drowsiness. To evaluate the safety impact of a risk factor, the exposure status under normal driving conditions is also necessary. Therefore, there is a need to extract exposure information under normal/non-crash driving conditions. This is done through baseline sampling. The appropriateness of the baseline sampling scheme has a critical impact on

the validity of the analysis. In this report, a random sampling scheme stratified by participant drivers is used. Its theoretical foundation and implications will be discussed in Chapter 3.

CHAPTER 3. A RANDOM SAMPLING SCHEME FOR BASELINE REDUCTION

Chapter 2 concluded that the overall analysis framework for naturalistic driving study is analogous to a case-cohort type epidemiology study. The data collection follows that of a cohort study while the analysis is based on case-control study. In a case-based approach, the safety events are identified after the data collection has finished. The controls, which represent the non-event, normal driving conditions, are selected subsequently. The selection of baseline is critical to the validity of the study. It is argued that the baseline sampling scheme should be considered in conjunction with appropriate risk measures and corresponding statistical models. In this study, it is proposed that the appropriate measure of risk for naturalistic driving study is the RRR. By using a random baseline sampling scheme the odds ratio can be used to approximate the RRR. The details of the development are discussed in this chapter.

The advantage of naturalistic driving study lies in the number of variables that can be collected. In particular, the video recordings can be used to assess driver behaviors which are difficult to retrieve from accident databases. The driver behaviors, however, change constantly over the driving process. Due to this time-variant property, the exposure status for a safety event is typically identified a short moment before the event. For example, the driver's behavior and environmental factors were identified within 6 s before the onset of a safety event.[(1)] To assess the exposure status of controls (which represent the exposure status under normal driving conditions), the critical question is where those controls should be located. This problem should be considered in conjunction with the risk measure adopted.

As discussed in the measure of the risk, it is difficult to assign probabilistic risk measures for time-variant exposures such as distraction and traffic conditions. A proper measure is the RR under each exposure level. An example using drowsiness is shown in figure 10. The exposure status is categorized into two levels: drowsiness and non-drowsiness. When the length of a segment is sufficiently small, the exposure status in this segment can be considered as homogeneous and can be categorized into either *drowsy period* or *non-drowsy period*. Conceptually, all the drowsy driving periods can be pulled together and all the non-drowsy driving periods can be pulled together. Thus the whole driving period can be divided into two exposure levels: the drowsy period and the non-drowsy period. As illustrated in figure 10

figure 10, the length of the drowsy period is represented by the red bar on the left and the non-drowsy period is represented by the blue bar on the right. The red stars represent crashes. In this setup, the crash rates for drowsy and non-drowsy exposure are:

$$Rate|Drowsy = \frac{\#\ of\ crashes\ happened\ during\ drowsy\ period}{Total\ duration\ of\ drowsy\ period}$$

$$Rate|Non-Drowsy = \frac{\#\ of\ crashes\ happened\ during\ nondrowsy\ period}{Total\ duration\ of\ nondrowsy\ period}$$

If $Rate|Drowsy$ is significantly greater than $Rate|Non-Drowsy$ then we can conclude that drowsiness is a significant factor contributing to traffic safety. The RRR for drowsiness is:

$$RRR_{Drowsy} = \frac{Rate|Drowsy}{Rate|Non-Drowsy}$$

Appropriate statistical tests can be used to test if the rate ratio is greater than 1, which corresponds to elevated risk for drowsiness. This evaluation based on crash rate is more accurately defined for time-variant exposure than the probability measures. However, there is a challenge in using the above approach: the total duration of drowsy and non-drowsy periods cannot be measured exactly using current technology. Extracting driver behavior information still relies primarily on manual data reduction and it is not practical to manually check thousands of hours of video data. Therefore, an alternative method has to be used.

Figure 10. Diagram. Illustration of accident rate for time-variance exposure.

The general principle for selecting controls is that "the controls should represent the population from which the cases were derived." Based on this general principle, a number of alternative sampling methods can be used; e.g., random sampling, matched sampling, case-crossover sampling, etc. The baseline sampling method adopted in this study is a total random sampling scheme, which is based on the event rate measure.

Consider the data collection process as a cohort study and let PT+ and PT- represent the exposure duration for E+ and E-, respectively. The number of accidents and the corresponding exposure duration can be represented in the following table.

Table 4. Crash rate.

	E+	E-
Crash	A	C
Duration	PT+	PT-
Crash rate	A/PT+	C/PT-

The RRR will be

$$Risk\ Rate\ Ratio = \frac{\frac{A}{PT+}}{\frac{C}{PT-}}$$

In a case control study, the total exposure durations PT+ and PT- are unknown. Instead, a set of baseline controls with size M0 (M0=C+D) are selected and their exposure statuses are identified as shown in table 5.

Table 5. Case-control contingency table.

	E+	E-	Size
Crash	A	B	M1
Control	C	D	M0

Like most contingency tables, the odds ratio can be calculated for the above table.

$$Odds\ Ratio = \frac{AD}{BC} = \frac{A/C}{B/D}$$

When the following three conditions are satisfied, the EOR can be used to approximate the RRR:

1. M0 subjects are randomly selected via source population
2. Their exposure odds (B/D) are similar to that in source population (Time+/Time-).
3. Steady state

As a case-cohort study, the first condition is automatically satisfied. If the duration of each baseline is short enough, the state within this short period can be considered as steady. The key of the baseline sample scheme is to satisfy the second condition.

$$\frac{B}{D} \approx \frac{PT+}{PT-}$$

In which case

$$Odds\ Ratio = \frac{\frac{A}{C}}{\frac{B}{D}} \approx \frac{\frac{A}{C}}{\frac{PT+}{PT-}} = \frac{\frac{A}{PT+}}{\frac{C}{PT-}} = Risk\ Rate\ Ratio$$

Two sampling methods can satisfy this critical condition: the total random sampling and systematic sampling. The details for each scheme are discussed below.

Random sampling

For random sampling, samples are randomly selected for the baselines. Typically, the samples are stratified by drivers and the number of samples for each driver is proportional to the valid moving hours or miles traveled. Under this total random sampling scheme, the probability that a baseline is from the $PT+$ period is proportional to its relative duration, i.e.,

$$Prob(A\ baseline\ is\ in\ PT+) = \frac{PT+}{(PT+) + (PT-)}$$

when the total sample size, $N_{base} = B + D$, is large, the number of baseline falls in $PT+$ is

$$B = N_{drowsy} \approx N_{base} * \frac{PT+}{(PT+) + (PT-)}$$

Similarly, the number of baseline falls in $PT-$ is

$$N_{drowsy} = N_{base} * \frac{PT+}{(PT+) + (PT-)}$$

Thus, critical condition for odds ratio to RR approximation holds as shown below.

$$\frac{B}{D} \approx \frac{N_{base} * \frac{PT+}{(PT+)+(PT-)}}{N_{base} * \frac{PT-}{(PT+)+(PT-)}} = \frac{PT+}{PT-}$$

The random sampling method is illustrated in figure 11.

Figure 11. Diagram. Random sampling scheme.

There are several advantages of random sampling. It is relatively easy to implement and easy to find replacement for invalid baselines. The random samples represent the general baseline status so they can be used in studies focusing on different risk factors. Finally, the statistical analysis is relatively straightforward. For the above reasons, the random sampling scheme was adopted in this study.

Systematic sampling

In a systematic sampling scheme, baseline samples are selected with equal intervals (moving hours or miles traveled) from the driving data. Systematic sampling is based on the same principle as random sampling and it can be shown that the odds ratio to RR approximation will hold for the systematic sampling. However, the systematic system scheme does have one drawback: when a control is not valid it is difficult to find alternatives. The invalid controls (baselines) are quite common because of misaligned video cameras, etc. The statistical analysis for systematic sampling is identical to the random sampling method.

CHAPTER 4. STATISTICAL MODELING

The objective of the statistical analysis is to quantitatively evaluate the safety impacts of risk factors and to conduct inference to the source population. As discussed previously, the odds ratio, which is an approximation to RRR, will be the primary risk measure. Two aspects of odds ratio are of interest: the point estimate and precision of point estimate. The point estimate represents the magnitude of the impact of a factor; i.e., an odds ratio of 4 implies that one level of the factor is 4 times more dangerous than the reference level. Another aspect is the precision of the estimation, which can be measured by the variance of estimates or the length of the confidence interval. A risk factor is considered as significant only when the statistical test indicates that the estimated effect significantly differs from a null value, which is 1 for odds ratio. The significant test has a direct relationship with the confidence interval: the odds ratio is statistically significantly different from 1 if and only if its 95% confidence interval does not include the null value 1.

There are several challenges in the analysis of naturalistic driving data. The confounding/interaction effects and driver-specific correlations are two main obstacles addressed in this report. As an observational study, the safety impact of a factor of interest can be easily distorted by other factors. The confounding and interaction effects could distort the true relationship between the factors of interest and safety outcome and have to be addressed in order to get a valid conclusion. Secondly, there are multiple events/baselines for each driver and those events/baselines collected for the same driver should not be considered as independent. However, most simple statistical models assume independence among observations and a more sophisticated modeling method should be used. It should be noted that the appropriate statistical method is always coupled with study design and baseline reduction methods. For example, matched baseline sampling methods such as case-crossover will bring extra correlation in each matched set.

The total random baseline sampling approach adopted in this study does not induce extra correlations and imposes few constraints. Therefore, the analysis is quite flexible. At the same time, the method is more prone to confounding and interaction effects. In this report, several alternative analysis methods were discussed and compared, including the simple contingency table analysis, regular logistic regression modes, the Generalized Estimation Equation (GEE) models, and the mixed effect models.

Simple contingency table analysis

The contingency table analysis is the simplest method by which to calculate odds ratios. The odds ratio and corresponding variance can be easily calculated. However, this method only considers one factor at a time and cannot address interaction/confounding effects. Furthermore, the contingency table analysis assumes observations are independent of each other, which does not fit the naturalistic driving study because of the unavoidable driver-specific correlations. Therefore, this method is more appropriate for exploratory analysis and caution should be given when using its results to draw formal conclusions.

The calculation for contingency table analysis is straightforward. For a factor with two exposure levels, E+ and E-, the safety events and exposure data can be arranged in a 2×2 contingency table as shown below.

	E+	**E-**
Safety event	A	B
Control	C	D

The point estimation for odds ratio is AD/BC. Two popular types of methods can be used for statistical inference. When the same size is large, an asymptotic normal approximation-based approach can be used with the following standard error.

$$\hat{\sigma}(\log\hat{\theta}) = (\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D})^{1/2}$$

The corresponding Wald confidence interval is

$$\log\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}(\log\hat{\theta})$$

When the sample size is small, the Fisher's exact test can also be used to conduct statistical inference.

The analysis method above does not adjust for potential confounding and interaction factors. This is especially problematic since the random sampling scheme developed in this study does not control those factors during the sampling process. If a random sampling method and a simple contingency table analysis are used together, the confounding and interaction effects will be totally ignored. Not accounting for these effects can negate the validity of the conclusions.

One remedy for this problem is to use stratified analysis. In a stratified analysis, the data will be ground into strata for every level of a confounding/interaction factor. One contingency table will be constructed for each stratum and the conclusion will be based on the results from stratified contingency tables. The main drawback of the stratified analysis is that the sample size in each stratum quickly decreases with an increased number of factors. For a naturalistic driving study, there will usually be a number of potential confounding/interaction factors and the number of safety events in each stratum is typically insufficient for a meaningful statistical conclusion. Therefore, the stratified analysis is not an attractive alternative. Model-based approaches, such as logistic regression, can address those issues relatively easily.

Ordinary Logistic Regression Models

In a naturalistic driving study, safety outcomes are either safety events (e.g., crashes and near-crashes) or baselines. The logistic regression model can be used for this type of categorical outcomes. For example, to model crashes and baselines, the model assumes the outcomes are from a binary distribution with two possible values. There is a single model parameter which is the probability of crash. The crash probability is then connected with risk factors to be evaluated through a logit link function. The effect of risk factors can be evaluated by examining the regression coefficients. The logistic regression model can accommodate multiple risk

factors, which allows those factors to be evaluated simultaneously. The capability of multiple factor analysis provides a mechanism to address the confounding and interactive effects through modeling.

The general setup for a logistic regression is described as follows. Define a binary random variable Y_{ij} such that

$$Y_{ij} = \begin{cases} 1 & Crash/near-crash \\ 0 & Baseline \end{cases}, \quad i = 1, \dots, I;\ j = 1, \dots, J_i$$

where I is the number of drivers and J_i is the number of observations for driver i. Assume Y_{ij} follows a Bernoulli distribution, i.e.,

$$Y_{ij} = Bernoullie(p_{ij}) \qquad (1)$$

The model coefficient p_{ij} represents the probability of being a safety event for observation Y_{ij}. It is assumed that this probability will be influenced by factors such as driver behavior and driving environments, etc. This connection between the safety impacts of a set of factors and the crash probability p_{ij} is mathematically modeled through a logit link function with the following form

$$logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_K X_{Kij} \qquad (2)$$

where X_{kij} is the variable based on a risk factor k and β_k is the corresponding regression coefficient. With proper parameterization, it can be shown that the exponential of the regression coefficient β_k is corresponding to the odds ratio for the k^{th} factor. Another merit of the odds ratio estimated from the logistic regression is that the odds ratio for a specific factor can be considered as an averaged value over all the levels of other factors included in the same model. Thus the confounding effect can be effectively addressed by simply including multiple factors that might confound with each other simultaneously in the model.

The ordinary logistic regression discussed above assumes observations are independent of each other, which is obviously not appropriate as each participant might have multiple safety events and baselines. In this report, two alternative methods were used to address this issue: the GEE and the mixed effect model.

Generalized Estimation Equation

The ordinary logistic regression model assumes independent observations. For naturalistic driving data, it is argued that events/baselines for the same driver should be correlated instead of independent because a driver usually has some unique characteristics and those characteristics are shared by the events/baselines from this driver. The GEE model can be used to incorporate this correlation. Originally developed to model longitudinal data (i.e., measures from a same patient from different time points) by Liang and Zeger[11], the GEE model assumes that the observations are marginally correlated. The GEE model specifies the mean and covariance (first

two moments) structures of a distribution from the exponential distribution family. For example, in naturalistic driving studies, the crash and baseline are assumed from a Bernoulli distribution (Equation 1). Similar to the ordinary logistic regression, the GEE model also assumes the logit of event probability p_{ij} is associated with a set of risk factors through a logit link function as in Equation 2.

In a GEE model, observations Y_{ij} and $Y_{ij'}$ from the same driver i are correlated and the correlation between is non-zero, i.e., $Corr(Y_{ij}, Y_{ij'}) \neq 0$. This violates the independence property of a Bernoulli distribution. The GEE thus is not based on a proper distribution/likelihood function; instead it is a Quasi-Likelihood-based approach for which no proper probabilistic models exist. For that reason the GEE should be considered as an estimation method rather than a modeling approach. Also because of this, the GEE cannot be extended to Bayesian framework which depends on proper probabilistic models.

The correlation structure $Corr(Y_{ij}, Y_{ij'})$ for the observations from the same driver needs to be pre-specified for the GEE model. Denote the correlation matrix with the following general form

$$corr(\boldsymbol{Y_i}) = \begin{bmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1J} \\ \alpha_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \alpha_{J-1,J} \\ \alpha_{J1} & \cdots & \alpha_{J,J-1} & 1 \end{bmatrix} \quad (3)$$

where $\alpha_{jj'} = Corr(Y_{ij}, Y_{ij'})$, and the vector $\boldsymbol{Y_i}$ represents all events/baselines for driver i, i.e., $\boldsymbol{Y_i} = (Y_{i1}, Y_{i1}, \cdots Y_{i,J_i})$.

In a GEE, the same correlation structure is assumed for all drivers. Therefore, the matrix entries do not depend on the index for a particular driver i. In general, the number of observations for each driver is not exactly the same. The J in above correlation matrix represents the maximum number of observations per driver, i.e., $J = \max_i(J_i)$.

There are a number of alternative covariance structures. For an unstructured correlation matrix, each $\alpha_{jj'}$ in Equation 3 can take different values. The unstructured correlation put the least constraints to the correlation structure but contains substantially more parameters. This could be problematic when the sample size is small or when J is large. Another widely used structure is the exchangeable correlation matrix, in which all $\alpha_{jj'}$s are assumed be to equal, i.e.,

$$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1 & j = j' \\ \alpha & j \neq j' \end{cases} \quad (4)$$

The exchangeable correlation matrix includes only one parameter and is easy for fitting. However, the assumption that all observations are equally correlated does not fit a real situation well. The autoregressive (AR1) model is popular in time-series analysis. The AR1 model assumes

$$corr(Y_{ij}Y_{ij+1}) = \alpha$$

Thus, the correlation matrix has the following form. The AR1 model also contains one extra parameter and imposes a pretty strong assumption for the correlation relationship.

$$Corr(\boldsymbol{Y}_i) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^J \\ \alpha & 1 & \alpha & \ddots & \vdots \\ \alpha^2 & \alpha & 1 & \ddots & \alpha^2 \\ \vdots & \ddots & \ddots & 1 & \alpha \\ \alpha^J & \cdots & \alpha^2 & \alpha & 1 \end{bmatrix}$$

As many researchers indicated, the GEE model is not sensitive to the choice of correlation structures and its quasi-likelihood estimator for regression parameter β is consistent, which means it will converge to the true value for a large sample, even when the correlation function is incorrectly specified.

Mixed effect models

The correlated categorical observations can also be modeled by the generalized linear mixed effect model (GLMM). Similar to the GEE approach, observations from the same driver are also assumed to be correlated. However, instead of specifying the correlations marginally, the GLMM builds the correlation structure through a conditional specification. More specifically, the GLMM model assumes that there is a random effect associated with each individual driver. One particular driver can be more likely to be associated with higher risk (if the random effect is positive), or less risk (if the random effect is negative). This assumption fit the observation from the naturalistic driver results that a small number of drivers contribute a large proportion of safety events, thus GLMM is a rather attractive alternative model.

The probability distribution part of the GLMM is identical to the ordinary logistic regression model and the GEE model (Equation 1). The difference lies in the modeling structure for the Bernoulli parameter p_{ij}, which is specified through the conditional expectation of Y_{ij} given a random effect term; i.e., $p_{ij} = E\left[Y_{ij}\middle|\boldsymbol{u}_i\right]$, where $\boldsymbol{u_i}$ is the random effect. The p_{ij} is connected with a set of covariates with a logit link function; i.e.,

$$logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{X'_{ij}\beta} + \boldsymbol{z'_{ij}u_i} \qquad (5)$$

For convenience, matrix notation was used in the above formulation. Here $\boldsymbol{X_{ij}}$ is the vector of covariates for observation Y_{ij}, $\boldsymbol{X_{ij}} = (1, X_{1ij}, X_{2ij}, \dots, X_{Kij})'$; the $\boldsymbol{\beta}$ is the vector of regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_K)'$. The $\boldsymbol{u_i}$ is a vector of mixed effects and the $\boldsymbol{z_{ij}}$ is the corresponding design matrix. For simplification, consider the special case with univariate mixed effect and $z_{ij} = 1$. The u_i is a random variable and is typically assumed from a normal distribution $N(0, {\sigma_u}^2)$.

In the mixed effect logistic model, the observations from the same driver, i, share the same random term u_i, which induces the driver-specific correlations. In this setup, the univariate mixed effect adjusts the intercept of the linear regression part but does not modify the fixed effect $\boldsymbol{\beta}$. This model is also called random intercept model.

The above formula implies that there is a random effect u_i associated with driver i. This random effect will vary among drivers and follows a normal distribution. The value of u_i is directly

related to the probability of driver i being involved in a safety event. When u_i is large, the probability of this driver being involved in a safety event, i.e., p_{ij}, will be high and vice versa.

The GLMM is an extension of the basic logistic regression to allow correlations among data to be incorporated. Compared to the GEE model, the GLMM has several advantages. The GLMM has a clear interpretation for the driver-specific correlation and fits the observed "good driver, bad driver" in naturalistic driving study. Furthermore, the GLMM is based on a proper probabilistic model and can be relatively easily extended to the Bayesian framework. The Bayesian approach has several advantages over the classical statistical approach. For example, it can easily incorporate prior information into the estimation of risk, which is very useful for small scale study when the researchers have good *a priori* knowledge about the risks. The hierarchical Bayesian models can combine multiple studies together and are especially useful for multi-center studies. The GLMM model developed in this project can be readily extended to a Bayesian framework.

APPLICATION

The 100-Car Study was a large scale naturalistic driving study for which more than 100 participant drivers were recruited from the Northern Virginia/Washington, DC area.[1] Various types of instrumentation were installed on the participant vehicles, including: five-channel video cameras, front and rear radar sensors, accelerometers, and global positioning systems (GPS). In addition, the ability to obtain information from the vehicle network (e.g., speed) and track lanes using machine-vision was possible. The study lasted for over a year and collected approximately 2 million miles and 43,000 hours of driving data. Three different types of safety event were defined: crash, near-crash, and critical incident. Those safety events were identified retrospectively. Altogether, 69 crashes, 761 near-crashes, and 8,295 critical incidents were identified.

The main focus of event-based analysis is to identify factors having significant impacts on traffic safety. The continuously recorded naturalistic driving data, including the vehicle kinematic characteristics, the driving environments, and driver behavior, provide an unprecedented opportunity to evaluate the safety impacts of those factors. In this study, it is believed that the status of factors immediately before a safety event shall have a direct impact on safety. Therefore, the exposure status for a 6-second time period, 5 s before and 1 s after the onset of a crash or near-crash, were extracted from the video and instrument data. The kinematic features could be extracted automatically. However, the information for driver behavior and the environmental factors relied on examining the video files visually by trained data reductionists. A rigorous data reduction protocol was implemented throughout the data reduction process. More details about data reduction and quality control can be found in the report by Klauer et al.[10]

As discussed in Chapter 3, the exposure information from safety events needs to be compared with that from normal driving conditions in order to quantitatively evaluate risks. To make the exposure information comparable, the same time duration of 6 s was adopted for both event and baseline data reduction. The data reduction for baselines followed the exact same protocol as that for safety events to ensure comparability. Conceptually, the selection of baseline samples

followed a total random sampling scheme as discussed in the methodology section (Chapter 3). In practice, a two-stage sampling approach was implemented to utilize a previously reduced baseline data set and supplement it with additional samples as necessary.

Initially, 20,000 baseline samples were reduced according to a proportional sampling scheme. In the proportional sampling scheme, the number of baselines for a given driver is proportional to the number of safety events for that driver. Therefore, no baseline samples were reduced if the driver had no safety events. After the number of baselines for a driver was determined, a random sampling scheme was then used to randomly sample from the driver's trips. The proportional sampling scheme, of course, does not fit the total random sampling scheme developed for this study. However, due to the high cost of data reduction, there was motivation to maximize the use of existing data. The within driver random sampling of the proportional sampling scheme makes it possible to use part of the existing data for the total random sampling scheme. The details of this approach are discussed below.

The total random sampling scheme adopted in this study requires equal probability for each 6-second period for all recorded data. So for a baseline sample, the probability that it will be located for a specific driver is proportional to the driver's total driving time; i.e.,

$$\Pr(A\ baseline\ is\ for\ driver\ i) = \frac{Driving\ time\ for\ driver\ i}{total\ driving\ time\ for\ all\ drivers}$$

When the sample size is large, the number of baselines for a given driver i will be proportional to this probability

$$E[N_i] = \frac{Driving\ time\ for\ driver\ i}{total\ driving\ time\ for\ all\ drivers} * N$$

where $E[N_i]$ is the expected number of baselines for driver i and N is the total number of baselines.

The above calculation indicates that when the sample size is large, the desired number of baseline samples for a specific driver can be predetermined. The total random sampling condition can be satisfied if baseline samples were randomly selected from within each driver's data. In this study a stratified random sampling approach was adopted. The method consists of two steps: the first step is to predetermine the number of baselines for each driver and the second step is a total random sampling within each driver.

The sampling rate is approximately 0.5 baselines per subject-driving hours. Based on the total length of available video files, this translated into approximately 17,660 baseline samples. The driving time for each driver is extracted and the expected number of baselines for each driver is calculated accordingly. This expected number is then compared with that from the original proportional sampling data (existing baselines). For each driver, two possible actions were taken based on the comparison. Let N_i denote the expected number of baselines for driver i by the stratified random sampling scheme and N_i^0 represent the number of baselines for driver i from the existing proportional samples. Specifically,

1. If $N_i < N_i^0$, that is, the number of existing baseline is more than desired, N_i baselines will be randomly drawn from the existing N_i^0 baselines.
2. If $N_i > N_i^0$, a new supplemental reduction will be conducted with sample size $N_i - N_i^0$, randomly drawn from the driving data of driver i.

This procedure guarantees that each driver will have the expected number of baselines. At the same time, the randomization in every step also guarantees the randomness within each driver. Thus the properties of a total random sampling scheme were ensured and the method maximally utilized the existing baseline reduction results. The summary of the data reduction results is shown in table 6.

Table 6. Baseline sample size.

SAMPLING RATE (SAMPLE/SUB_MOV_HR)	TOTAL EXPECTED SAMPLES	RESAMPLE FROM EXISING BASELINE	NEW SAMPLES NEEDED
0.5	17660	14036	3624

Because of the missing/invalid video file, the final data reduction resulted in a total of 17,344 baseline samples.

STATISTICAL ANALYSIS

The final data set includes 69 crashes, 761 near-crashes, and 17,344 baseline samples. The factors of interest are mostly time-variant exposure factors, including: drowsiness, distraction, traffic density (level of service – LOS), lighting conditions, relationship to junction, road surface conditions, and weather conditions. Three alternative analyses methods were presented in the order of simple contingency table analysis, the GEE, and the mixed effect models. The outputs from various analyses were also compared.

Simple contingency table analysis

The risk associated with each factor was analyzed using the simple contingency table analysis. Note that this analysis approach does not adjust for the potential interaction and confounding effects; nor does it incorporate the correlations among observations from the same driver. The contingency table analysis shall thus be considered as an exploratory analysis tool. The odds ratio estimation results for the risk factors are presented below.

Drowsiness

The drowsiness was evaluated by visually inspecting the driver's behaviors and eye closure information. The drowsiness refers to a driver who is either moderately to severely drowsy, as defined by Wierwille and Ellsworth.[12] A driver who is moderately drowsy will exhibit slack musculature in the facial muscles and limited overall body movement as well as a noticeable reduction in eye scanning behaviors. A severely drowsy driver will exhibit all the above behaviors as well as extended eyelid closures and will have difficulties keeping his/her head in a lifted position. The status was classified as either drowsy or non-drowsy. The contingency table for drowsiness is shown in table 7 and the odds ratio estimation is presented in

table 8. As can be seen, drowsiness significantly increases both crash risk and near-crash risk.

Table 7. Contingency table for drowsiness.

	Drowsy	Not Drowsy	Total
Crash	14	55	69
Near-crash	97	664	761
Baseline	599	16,745	17344

Table 8. Odds ratio estimation for drowsiness.

	Odds Ratio	*p*-value	95% Confidence Limits	
Crash	7.12	<0.001	3.94	12.87
Near-crash	4.08	<0.001	3.25	5.13

Traffic flow

The traffic density is evaluated by the LOS, which includes six levels as shown in table 9. The definitions of the six LOS levels are as follows:

- LOS A: Free flow
- LOS B: Flow with some restrictions
- LOS C: Stable flow, maneuverability and speed are more restricted
- LOS D: Unstable flow, temporary restrictions substantially slow driver
- LOS E: Flow is unstable, vehicles are unable to pass, temporary stoppages, etc.
- LOS F: Forced traffic flow condition, with low speeds and traffic volumes below capacity

Table 9. Contingency table for LOS.

Severity(Severity)	LOSs						
Frequency	LOS A	LOS B	LOS C	LOS D	LOS E	LOS F	Total
Crash	41	14	7	4	2	1	69
Near-crash	244	233	191	64	26	2	760
Baseline	8370	6789	1606	322	160	96	17343

Due to the small sample size for LOS E and LOS F, some LOS categories were aggregated. For crashes, LOS D, LOS E, and LOS F were aggregated and for near-crashes, LOS E and LOS F were aggregated.

Table 10. Odds ratio estimation for LOS.

	Odds Ratio	*p*-value	95% Confidence Limits	
Crash				
LOS B versus A	0.42	<0.01	0.23	0.77
LOS C versus A	0.89	0.78	0.40	1.99
LOS DEF versus A	2.47	0.03	1.10	5.54
Near-crash				
LOS B versus A	1.18	0.08	0.98	1.41
LOS C versus A	4.08	<0.0001	3.35	4.97
LOS D versus A	6.82	<0.0001	5.07	9.18
LOS EF versus A	3.75	<0.0001	2.49	5.66

There are some interesting patterns that can be seen from table 10. For crash, the odds ratio of LOS B versus LOS A is significantly lower than 1, which indicated some protective effect. On the other side, high traffic densities, i.e., LOS DEF, are associated with elevated risk compared to LOS A.

A quite different pattern exists for near-crash. LOS A is associated with the lowest risk. There is no significant difference between LOS B and LOS A. The LOS D is associated with the highest risk.

The above results have some interesting implications for the relationship between safety events and traffic density. Some level of interaction between vehicles (such as for LOS B and LOS C) will not necessarily increase the crash risk. However, the chance of near-crash will increase monotonically with the increase in traffic density.

Distraction

Distraction is commonly presented during driving. The level of distraction is associated with the complexity of non-driving-related tasks. Three levels of manual/visual complexity (complex secondary tasks, moderate secondary tasks, and simple secondary tasks) were defined as shown in table 11. The complexity levels are based on whether the task requires multi-step, multiple eye glances away from the forward roadway, and/or multiple button presses.[13] Moderate tasks are those that require at most two glances away from the roadway and/or at most two button presses, while simple tasks are those that require no or one button press(es) and/or one glance away from the forward roadway. The contingency table for distraction is shown in table 12 and the odds ratio estimations are shown in
table 13.

Table 11. Three levels of manual/visual complexity.

Simple Secondary Tasks	Moderate Secondary Tasks	Complex Secondary Tasks
1. Adjusting radio	1. Talking/listening to handheld device	1. Dialing a handheld device
2. Adjusting other devices integral to the vehicle	2. Handheld device-other	2. Locating/reaching/ answering handheld device
3. Talking to passenger in adjacent seat	3. Inserting/retrieving CD	3. Operating a personal digital assistant (PDA)
4. Talking/Singing: no passenger present	4. Inserting/retrieving cassette	4. Viewing a PDA
5. Drinking	5. Reaching for object (not handheld device)	5. Reading
6. Smoking	6. Combing or fixing hair	6. Animal/object in vehicle
7. Lost in thought	7. Other personal hygiene	7. Reaching for a moving object
8. Other simple tasks	8. Eating	8. Insect in vehicle
	9. Looking at external object	9. Applying makeup

Table 12. Contingency table for distraction.

Frequency	Complex	Moderate	Simple	No Distraction	Total
Crash	6	9	11	43	69
Near-crash	43	83	85	550	761
Baseline	388	3001	4759	9196	17344

Table 13. Odds ratio estimation for distraction.

	Odds Ratio	*p*-value	95% Confidence Limits	
Crash				
Simple versus Non	0.49	0.037	0.25	0.96
Moderate versus Non	0.64	0.226	0.31	1.32
Complex versus Non	3.31	0.006	1.40	7.82
Near-crash				
Simple versus Non	0.30	<0.001	0.24	0.38
Moderate versus Non	0.46	<0.001	0.36	0.58
Complex versus Non	1.85	0.0002	1.34	2.57

As shown in
table 13, the complex secondary task significantly increases the risk of crash and near-crash. It is also interesting to observe that simple and moderate secondary tasks actually show protective effect (the odds ratio is smaller than 1). This protective effect may be due to drivers selecting a relatively safe point to engage in secondary tasks whereas the complex task may require enough resources that it increases risk regardless of when the task is performed.

Weather

The contingency table for weather conditions is presented in table 14. Due to small samples size, the weather conditions were aggregated into two categories: the normal weather conditions and the inferior weather conditions. The normal conditions include clear and cloudy. The inferior weather conditions include fog, mist, raining, sleeting, and snowing. The aggregated table is shown in table 15 and the odds ratio estimations are shown in table 16. As can be seen, there are no statistically significant results for both crash and near-crash, though the odds ratio for crash shows a moderately elevated risk at 1.8.

Table 14. Contingency table for weather conditions.

Frequency	Clear	Cloudy	Fog	Mist	Raining	Sleeting	Snowing	Other	Total
Crash	54	6	0	0	8	0	1	0	69
Near-crash	599	99	1	1	57	0	3	1	761
Baseline	15436	562	29	20	1235	9	42	11	17344

Table 15. Aggregated contingency table for weather conditions.

Frequency	Inferior weather	Normal Weather	Total
Crash	9	60	69
Near-crash	62	699	761
Baseline	1335	16009	17344

Table 16. Odds ratio estimation for weather conditions.

	Odds Ratio	*p*-value	95% Confidence Limits	
Crash	1.80	0.10	0.89	3.63
Near-crash	1.06	0.65	0.82	1.39

Lighting conditions

The lighting conditions for event and baseline data are shown in table 17. Due to the small sample size for crashes, the data were aggregated into two categories: daylight condition and other lighting condition, as shown in
table 18.

Table 17. Contingency table for lighting conditions.

Frequency	Darkness lighted	Darkness not lighted	Dawn	Daylight	Dusk	Total
Crash	17	5	1	43	3	69
Near -crash	126	54	14	502	65	761
Baseline	2600	1633	75	12126	910	17344

Table 18. Aggregated contingency table for lighting conditions.

Frequency	Other lighting	Daylight	Total
Crash	26	43	69
Near-crash	259	502	761
Baseline	5218	12126	17344

The odds ratio estimations for lighting conditions are shown in table 19. As can be seen, the other lighting condition is associated with slightly increased risk with odds ratios of 1.41 and 1.12 for crash and near-crash, respectively. However, the odds ratio for crash is not significantly different from the neutral value of 1.

Table 19. Odds ratio estimation for lighting conditions.

	Odds Ratio	*p*-value	95% Confidence Limits	
Crash	1.41	0.17	0.86	2.29
Near-crash	1.12	0.02	1.03	1.40

Road surface condition

The contingency table for surface conditions is shown in table 20. Due to the small number of observations in each category, the data were aggregated into two categories: the dry road surface and the other road surface, as shown in
table 21.

Table 20. Contingency table for road surface conditions.

Frequency	Dry	Icy	Muddy	Snowy	Wet	Other	Total
Crash	51	1	0	4	13	0	69
Near-crash	654	4	0	4	98	1	761
Baseline	15573	9	1	127	1630	4	17344

Table 21. Contingency table for road surface conditions.

Frequency	Others	Dry	Total
Crash	18	51	69
Near-crash	107	654	761
Baseline	1771	15573	17344

As can be seen from
table 22, the non-dry surface condition has a significant association with safety events. In particular, the non-dry surface conditions are 3 times more dangerous compared to the dry surface condition.

Table 22. Odds ratio estimation for road surface condition.

	Odds Ratio	*p*-value	95% Confidence Limits	
Crash	3.10	0.0002	1.81	5.32
Near-crash	1.43	0.0007	1.15	1.78

Relationship to junctions

Table 23 is the contingency table for relationship to junction. Again, the data were aggregated into junction and non-junction categories, as shown in
table 24.

Table 23. Contingency table for relationship to junction.

Frequency	Crash	Critical Incident	Near-crash	Baseline
Driveway alley access, etc.	2	138	8	53
Entrance/exit ramp	6	311	40	396
Interchange area	0	71	16	255
Intersection	17	858	149	1065
Intersection-related	11	1742	76	926
Non-junction	26	5065	456	14194
Parking lot	6	90	14	408
Rail grade crossing	0	4	0	4
Other/No data	1	16	2	43
Total	69	8295	761	17344

Table 24. Aggregated contingency table for relationship to junction.

Frequency	Junction	Non-junction	Total
Crash	34	35	69
Critical Incident	2986	5309	8295
Near-crash	281	480	761
Baseline	2646	14645	17291

The odds ratio estimation for relationship to junction is shown in table 25. As can be seen, junction is much more dangerous than non-junction with an odds ratio of 5.38.

Table 25. Odds ratio estimation for relationship to junction.

	Odds Ratio	*p*-value	95% Confidence Limits	
Crash	5.38	<0.0001	3.35	8.64
Near-crash	3.24	<0.0001	2.78	3.78

GEE and Mixed Effect Model Fitting

The GEE model was implemented to the total random samples from this project. The results for crash and near-crash are shown in table 26 and table 27, respectively. The GEE model fitting used an exchangeable working correlation function. The estimations for the correlation are small (0.003 for crashes and 0.035 for near-crashes). These small values indicate a rather weak marginal correlation among the observations.

The mixed effect model fitting results for crashes are shown in table 28. The estimated variance for mixed effect is 0.713 with a standard deviation of 0.28. The mixed effect model fitting results for near-crashes are shown in table 29. The corresponding estimated variance for the random intercept is 0.888 with a standard deviation of 0.173. Compared to the rest of the parameter estimations, this does indicate that there are considerable individual variations among drivers. This result is consistent with the fact that a small number of drivers contribute a large proportion of the safety events.

Table 26. GEE model results for crash.

Label	Odds Ratio	Standard Error	95% Confidence Limits		p-value
Drowsy	**6.35**	2.04	3.38	11.91	<.0001
Weather: Inferior versus Normal	2.17	1.13	0.79	6.01	0.13
Road Surface: Dry versus Other	**4.81**	2.18	1.98	11.71	<0.001
Lighting: Day versus Other	1.04	0.31	0.58	1.86	0.89
LOS B versus A	**0.42**	0.13	0.23	0.76	<0.001
LOS C versus A	0.89	0.38	0.38	2.08	0.79
LOS DEF versus A	1.83	0.94	0.67	5.03	0.24
Distraction: Complex versus Non	**3.51**	1.95	1.18	10.41	0.02
Distraction: Moderate versus Non	0.65	0.31	0.25	1.64	0.36
Distraction: Simple versus Non	0.54	0.20	0.26	1.11	0.09
Junction versus Non-junction	**5.89**	1.55	3.51	9.86	<.0001

Table 27. GEE model results for near-crash.

Label	Odds Ratio	Standard Error	97% Confidence Limits		p-value
Drowsy	**3.67**	0.58	2.69	5.01	<.0001
Weather: Inferior versus Normal	**1.94**	0.46	1.22	3.10	0.01
Road Surface: Dry versus Other	**2.17**	0.50	1.38	3.40	<0.001
Lighting: Day versus Other	1.17	0.12	0.96	1.43	0.13
LOS B versus A	1.18	0.12	0.98	1.43	0.09
LOS C versus A	**4.06**	0.51	3.17	5.20	<.0001
LOS DEF versus A	**4.99**	0.81	3.63	6.87	<.0001
Distraction: Complex versus Non	**2.02**	0.45	1.30	3.12	<0.001
Distraction: Moderate versus Non	**0.48**	0.07	0.37	0.63	<.0001
Distraction: Simple versus Non	**0.33**	0.04	0.25	0.42	<.0001
Junction versus Non-Junction	**3.36**	0.35	2.74	4.12	<.0001

Table 28. Mixed-effect model results for crash.

Label	Estimate	Standard Error	Pr > |t|	Odds Ratio	95% CL Lower	95% CL Upper
Drowsy	1.8424	0.3310	**<.0001**	**6.3120**	3.2988	12.0774
Weather: Inferior versus Normal	0.7631	0.4915	0.1206	2.1449	0.8184	5.6214
Road Surface: Dry versus Other	1.5665	0.3812	**<.0001**	**4.7897**	2.2688	10.1114
Lighting: Day versus Other	-0.03134	0.2716	0.9081	0.9691	0.5691	1.6503
LOS B versus A	-0.8619	0.3170	**0.0066**	**0.4224**	0.2269	0.7863
LOS C versus A	-0.1207	0.4191	0.7733	0.8863	0.3897	2.0154
LOS DEF versus A	0.6836	0.4355	0.1165	1.9809	0.8436	4.6516
Distraction: Complex versus Non	1.2244	0.4676	**0.0088**	**3.4021**	1.3604	8.5075
Distraction: Moderate versus Non	-0.3861	0.3770	0.3058	0.6797	0.3246	1.4232
Distraction: Simple versus Non	-0.6699	0.3470	0.0536	0.5118	0.2592	1.0103
Junction versus Non-Junction	1.7997	0.2491	**<.0001**	**6.0477**	3.7116	9.8541

Table 29. Mixed-effect model results for near-crash.

Label	Estimate	Standard Error	Pr > |t|	Odds Ratio	95% CL Lower	95% CL Upper
Drowsy	1.2608	0.1358	**<.0001**	**3.5282**	2.7040	4.6038
Weather: Inferior versus Normal	0.5995	0.2218	0.0069	1.8213	1.1792	2.8129
Road Surface: Dry versus Other	0.7139	0.1797	**<.0001**	**2.0419**	1.4356	2.9042
Lighting: Day versus Other	0.03648	0.08884	0.6813	1.0372	0.8714	1.2344
LOS B versus A	0.1647	0.09829	0.0938	1.1790	0.9724	1.4295
LOS C versus A	1.3672	0.1093	**<.0001**	**3.9243**	3.1674	4.8620
LOS DEF versus A	1.6787	0.1440	**<.0001**	**5.3586**	4.0405	7.1068
Distraction: Complex versus Non	0.6685	0.1842	0.0003	1.9514	1.3601	2.7997
Distraction: Moderate versus Non	-0.7249	0.1295	**<.0001**	**0.4844**	0.3757	0.6244
Distraction: Simple versus Non	-1.2119	0.1245	**<.0001**	**0.2976**	0.2332	0.3799
Junction versus Non-Junction	1.2522	0.08561	**<.0001**	**3.4980**	2.9576	4.1371

Table 30. Modeling comparison for crashes.

	GEE Model			Random Effect Model			Contingency Table: Crude Odds Ratio		
Factors	**Odds Ratio**	**95% CI Low**	**95% CI High**	**Odds Ratio**	**95% CI Low**	**95% CI High**	**Odds Ratio**	**95% CI low**	**95% CI High**
Drowsy	**6.35**	3.38	11.91	**6.31**	3.30	12.08	**7.12**	3.94	12.87
Weather: Inferior versus Normal	2.17	0.79	6.01	2.14	0.82	5.62	1.80	0.89	3.63
Road Surface: Dry versus Other	**4.81**	1.98	11.71	**4.79**	2.27	10.11	**3.10**	1.81	5.32
Lighting: Day versus Other	1.04	0.58	1.86	0.97	0.57	1.65	1.41	0.86	2.29
LOS B versus A	**0.42**	0.23	0.76	**0.42**	0.23	0.79	**0.42**	0.23	0.77
LOS C versus A	0.89	0.38	2.08	0.89	0.39	2.02	0.89	0.40	1.99
LOS DEF versus A	1.83	0.67	5.03	1.98	0.84	4.65	**2.47**	1.10	5.54
Distraction: Complex versus Non	**3.51**	1.18	10.41	**3.40**	1.36	8.51	**3.31**	1.4	7.82
Distraction: Moderate versus Non	0.65	0.25	1.64	0.68	0.32	1.42	0.64	0.31	1.32
Distraction: Simple versus Non	0.54	0.26	1.11	0.51	0.26	1.01	**0.49**	0.25	0.96
Junction versus Non-Junction	**5.89**	3.51	9.86	**6.05**	3.71	9.85	**5.38**	3.35	8.64

To compare the contingency table, the GEE model, and mixed effect models, the estimations for odds ratios from the three methods are pooled into table 30 and table 31 and also illustrated in figure 12 and figure 13. There are some discrepancies among the three methods. For example, for LOS DEF versus LOS A, the crude odds ratio is significant greater than 1 but odds ratios from GEE and mixed effect models show non-significant results.

The drowsiness shows a large impact on the crash risk with an odds ratio of 6.31 to 7.12. The model-based results are lower than the crude odds ratio. The weather condition shows no significant effects on crash risks. The road surface, however, significantly impacts traffic safety. The estimated odds ratios for GEE (4.81) and for mixed effect model (4.79) are substantially higher than the crude odds ratio of 3.10. This could be caused by the interaction between weather conditions and road surface conditions.

Lighting conditions do not show a significant impact for traffic safety. The traffic density as measured by the LOS shows some interesting patterns. The LOS B (free flow with some

restrictions) appears safer than LOS A (the free flow). The safety for LOS C, (stable flow with more restrictions) statistically is similar to LOS A. Due to limited observations, LOS D, LOS E, and LOS F were aggregated for the analysis; the results for simple contingency table and model base approaches are different. The crude odds ratio for LOS DEF is 2.47, which significantly differs from neutral value 1. However, the odds ratios from GEE and mixed effect models are statistically non-significant. The results for traffic density indicate that some level of interaction among vehicles will not necessarily lead to increased risk. However, high traffic density could have some negative impact on traffic safety.

Distraction shows mixed messages for safety. The complex tasks as defined in table 11 have a definite impact on safety with an odds ratio of 3.5. The effect of moderate tasks is inconclusive (statistically not significant). The simple tasks, however, show a protective effect and reduce the crash risk by half (odds ratio is about 0.5). The small odds ratio indicated that the relative frequency of simple tasks during crash/near-crash is smaller than that during normal driving conditions. There are several possible causes for this protective effect. First, simple tasks might increase driver alertness without impairing driving capability. This would benefit safety. Another possible explanation is that during crash/near-crash events, the drivers might be involved in more hazardous situations, e.g., drowsiness or engaging in complex secondary tasks. As a result, the driver is less likely to engage in simple tasks. This can also explain the low relative frequency of simple tasks (the protective effects) during safety events. A detailed review of the interactions among simple tasks and other risk factors is needed for a better understanding of the role of simple tasks on safety risk.

Junctions are among the most dangerous locations on the highway. The analysis indicates that the crash risk at junctions is 6 times more than at non-junction segments.

Due to the limited number of crashes, the odds ratios as shown in table 30 have relatively large confidence intervals. For example, the odds ratio for inferior weather condition shows a considerably large point estimate of around 2. However, it is statistically non-significant partly due to the larger variation caused by the small number of observations. The near-crash represents a safety event that is not as severe as a crash but also contains important information on the risk associated with a factor. Table 31 provides the results from the simple contingency table, the GEE model, and the mixed effect model.

Table 31. Modeling comparison for near-crashes.

	GEE Model			Mixed Effect Model			Contingency Table: Crude Odds Ratio		
Factors	**Odds Ratio**	**95% CI low**	**95% CI High**	**Odds Ratio**	**95% CI low**	**95% CI High**	**Odds Ratio**	**95% CI low**	**95% CI High**
Drowsy	**3.67**	2.69	5.01	**3.5282**	2.7040	4.6038	**4.08**	3.25	5.13
Weather: Inferior versus Normal	**1.94**	1.22	3.10	**1.8213**	1.1792	2.8129	1.06	0.82	1.39
Road Surface: Dry versus Other	**2.17**	1.38	3.40	**2.0419**	1.4356	2.9042	**1.43**	1.15	1.78
Lighting: Day versus Other	1.17	0.96	1.43	1.0372	0.8714	1.2344	1.12	1.03	1.40
LOS B versus A	1.18	0.98	1.43	1.1790	0.9724	1.4295	1.18	0.98	1.41
LOS C versus A	**4.06**	3.17	5.20	**3.9243**	3.1674	4.8620	**4.07**	3.35	4.97
LOS DEF versus A	**4.99**	3.63	6.87	**5.3586**	4.0405	7.1068	**5.46**	4.23	7.04
Distraction: Complex versus Non	**2.02**	1.30	3.12	**1.9514**	1.3601	2.7997	**1.85**	1.34	2.57
Distraction: Moderate versus Non	**0.48**	0.37	0.63	**0.4844**	0.3757	0.6244	**0.46**	0.36	0.58
Distraction: Simple versus Non	**0.33**	0.25	0.42	**0.2976**	0.2332	0.3799	**0.30**	0.24	0.38
Junction versus Non-Junction	**3.36**	2.74	4.12	**3.4980**	2.9576	4.1371	**3.24**	2.78	3.78

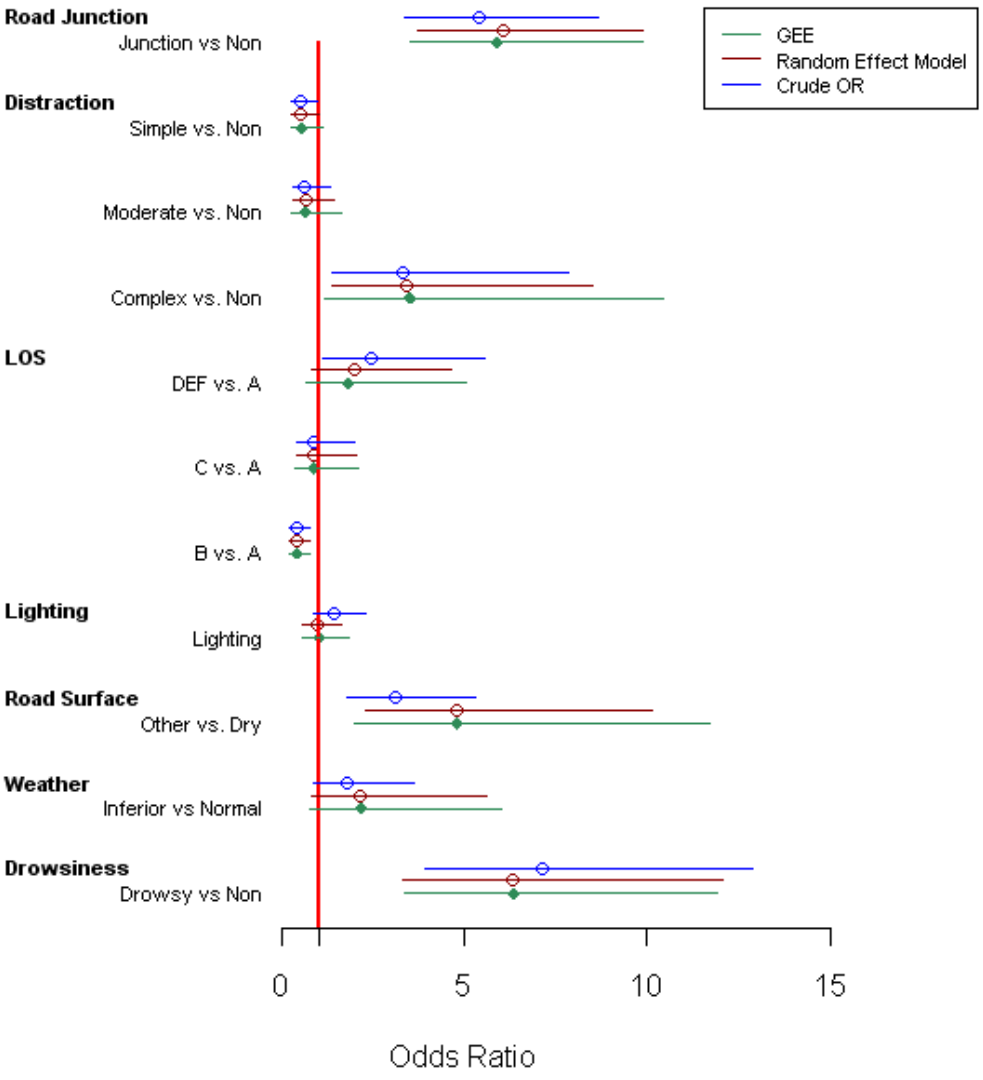


Figure 12. Graph. Crash odds ratios.

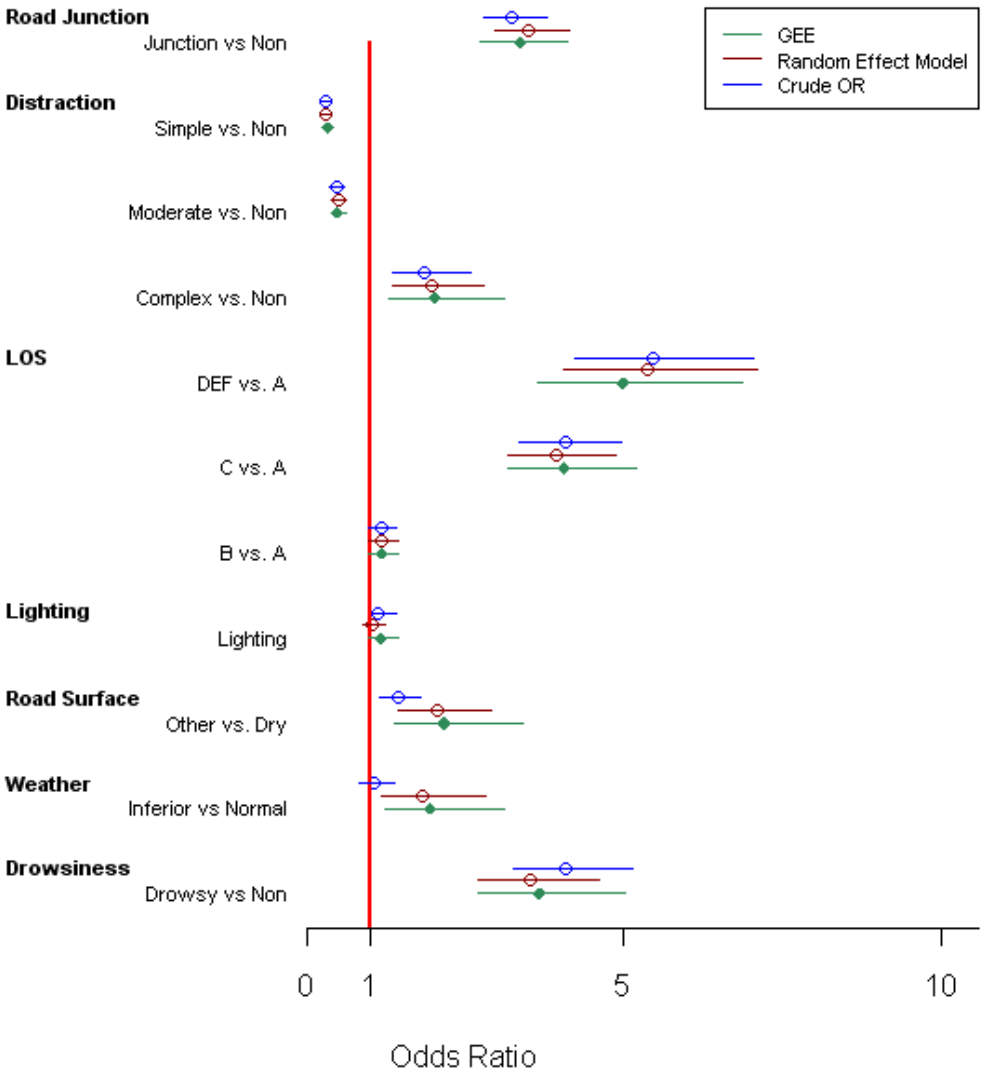


Figure 13. Graph. Near-crash odds ratios.

Compared to crash, the odds ratios for near-crash are smaller. For example, the odds ratio of drowsiness for near-crash is around 4 compared to 6-plus for crashes. At the same time, the precision of the estimation as indicated by the length of the 95% confidence interval is better; for example, the length of the CI of drowsiness odds ratio is 8.53 (11.91-3.38) for crash and 2.32 (5.01-2.69) for near-crash. The improved precision for estimation of the near-crash odds ratio is a direct consequence of the larger number of near-crashes.

The inferior weather conditions show a similar odds ratio as crash. However, results are statistically significant for near-crash based on GEE and random effect models. The inferior weather will significantly increase the risk of near-crash twofold.

Consistent with the results from crash, the lighting condition does not show a significant impact on the risk of near-crash.

The safety impacts of traffic density for near-crash show different patterns compared to crash. Higher traffic density shows consistent increase for the risk of near-crash. The odds ratios for LOS B, LOS C, and LOS DEF, contrasted with LOS A, increase monotonically. The LOS B shows no significant effect with a point estimation of 1.18. For LOS C, the risk of near-crash increases 4 times compared to LOS A, and the LOS DEF shows a fivefold increase for near-crash risk compared to LOS A. This result implies that with the increase in traffic density, there are increased interactions among vehicles and the possibilities of requiring evasive maneuvers will increase. However, for the alert driver the majority of those evasive maneuvers can be controlled so the risk of crash will not necessarily increase.

The effects of distraction for near-crash are similar to those for crash. Specifically, the complex secondary tasks show increased near-crash risk but the simple and moderate tasks show reduced near-crash risk.

The odds ratio for relationship to junction (3.36) indicates that near-crashes are more likely to happen at junctions.

CHAPTER 5. SUMMARY AND CONCLUSION

Naturalistic driving study is an innovative approach for studying traffic safety and driver behavior. The massive information collected provides an unprecedented opportunity for investigating research questions that cannot be addressed by accident databases or simulation studies. At the same time the naturalistic driving study approach also brings serious challenges for data analysis and modeling. This report focused on methodological issues for evaluating the risks using the safety outcomes of a naturalistic study. A comprehensive analysis framework was developed which consists of study design, measure of safety risk, and statistical models. The proposed framework was applied to the crash and near-crash safety events from the 100-Car Study.

The naturalistic study data collection shares the major characteristics of a perspective cohort study. However, the analysis of safety outcomes should follow a case-control design. Therefore, the study design is analogous to a case-cohort design in epidemiology study. The interpretation of risk and baseline sampling will follow the principles from the case-cohort study.

One major criticism for a case-based study design is that the corresponding risk estimation is based on exposure probability, which is undesirable for most researchers. This study addressed this issue by using an integrated baseline-sampling method and appropriate risk measures. It was argued that for most time-variant exposures, the risk rate as measured by number of safety events per unit of driving time/distance is the appropriate measure. Furthermore, it was shown that with a proper baseline sampling method, the odds ratio is an approximation for the RRR. This framework provides a solid theoretical foundation for safety-event-based risk analysis. It also provides a more intuitive interpretation of the main risk measure–the odds ratio–in the context of naturalistic driving study.

Another major concern for the analyses of naturalistic driving study is that there are multiple safety events and baselines for a single driver, thus the data are correlated instead of independent. Furthermore, the confounding and interaction among risk factors could impair the validity of the research. This study addressed those issues by proposing two logistic regression-based models, namely the GEE model and the mixed effect model. Although based on distinct statistical assumptions, both models can satisfactorily incorporate the within driver correlation. Furthermore, when multiple factors were input into a single model, the confounding and interaction among factors can be effectively adjusted. The validity of the results can be assured when the models were properly implemented.

The proposed framework was applied to the 100-Car Study. A random baseline sampling scheme stratified by the driving time of each driver was adopted. A total of 17,344 baseline samples were generated by re-sampling from an existing baseline set and a data reduction with more than 3,000 new baselines. Both crash and near-crash were modeled and the three analysis methods (the simple contingency table analysis, the GEE model, and the mixed effect logistic regression model) were applied to the reduced data. Following is a summary of the major conclusions from this analysis.

- There are some discrepancies among results from the GEE, the mixed effect model, and the crude odds ratio estimation. The confidence intervals of the crude odds ratio are in general narrower than those from two model-based approaches. However, this is considered overly optimistic given that it ignores the driver-specific correlation and fails to adjust for potential confounding factors.
- The GEE and mixed effect models can be used to evaluate the level of correlations among observations from the same driver. The GEE analysis indicates that the marginal correlations among observations are weak. The mixed effect logistic regression model shows moderate variations among drivers. This is consistent with the fact that a small number of drivers contribute a large proportion of the safety events.
- The odds ratio for crash is always larger than for near-crash. However, the precision of the estimation for near-crash, as measured by the length of the confidence interval, is substantially better than that for crashes. This result has significant implications for using near-crashes as a safety surrogate for crashes.
- The odds ratio results for crash and near-crash indicate that drowsiness will increase the risk of safety events substantially.
- The inferior weather condition will significantly increase the risk of near-crash and also show a considerable impact on crashes.
- Traffic condition shows complex effects on safety. Compared to free flow traffic condition (LOS A), high traffic density (as measured by LOS D, E, and F) is associated with higher risk for both crash and near-crash. Moderate levels of interactions among vehicles (as measured by LOS B and LOS C) provide a protective effect for crash, which could contribute to increased driver alert. However, LOS B and LOS C are associated with higher risk of near-crash.
- Complex secondary tasks will increase the risk of crash by more than 3 times. However, the simple and moderate secondary tasks show smaller exposure in crashes and near-crashes.
- The highway junction is much more dangerous than the non-junction highway segment.

In summary, the modeling results indicate that there are some discrepancies among model-based approaches (the GEE and random effect models) and the crude odds ratio. The model-based estimations considered both the among-driver correlations and the potential confounding effect among risk factors, thus they are considered to more accurately reflect the true underlying risk levels. The mixed effect model is considered as a preferred alternative due to two advantages. First, the mixed effect model is based on a proper distribution function and solid theoretical foundation. Second, the mixed effect model can directly reflect the variation of risk associated with drivers. This is consistent with observation that the number of safety events varies substantially among drivers.

The odds ratio results for crash and near-crash indicate that drowsiness will increase the risk of safety events substantially. The inferior weather conditions will significantly increase the risk of near-crash and also show a considerable impact on crashes. Certain levels of interactions among vehicles (LOS B and LOS C) do not provide a protective effect for the risk of crash, which could contribute to the increased driver alert. However, the LOS B and LOS C are associated with high risk of near-crash. For both crash and near-crash, a high level of traffic density (LOS DEF) is associated with higher risks.

Complex secondary tasks will increase the risk of crash by more than 3 times. However, the simple and moderate secondary tasks show smaller exposure in crashes and near-crashes. The highway junction is much more dangerous than the non-junction highway segment.

The framework developed in this study provides a theoretical justification for the case-based study method in naturalistic driving studies. The framework can be implemented on studying time-variance exposures such as distraction, drowsiness, and weather conditions.

There are several possible extensions for this study. The current analysis framework is based on data reduction in which all exposure factors were treated equally and independently. However, some crucial information was lost during the current data reduction and analysis method: 1) the sequence of the events happened before/during a crash and 2) the interaction between driver, vehicle, and driving environments. It is argued that the combination of factors, the sequence of events, and the chain of driver's reactions during a safety event contains far more information than each individual risk factor. For example, the AAA Foundation for Safety has listed the chain of events that lead to an accident and an accident can be avoided by breaking any of the links in the chain. An accident reconstruction and causal analysis can provide more insights into the true causal relationship between exposure and safety events. To assess and understand the effects of the combination and sequence of factors can shed light on the causal effects and help in developing safety countermeasures. This will bring more challenge into the analysis and will be worth further investigation. There is a need to develop a systematic approach to reconstruct the complete process of a crash and identify the corresponding critical factors, and this project will address these two issues.

There are two methods for accident reconstruction and critical risk set identification. Unlike the traditional accident reconstruction techniques that rely on post-accident evidence recovery, the naturalistic driving data not only have the true driver behavior and vehicle kinematic information but also the precise time stamps and order of events. Thus the proposed analysis will focus on the sequential relationships and interactions between events and the risk factors that happened before and during a crash. Three major components will be considered: the driver, the vehicle, and the outside driving environment. Various sequence diagram techniques will be explored for the reconstruction including the Event and Causal Factor Charting [(14)], Multiple Events Sequencing, and the Sequentially Timed Events Plotting Procedure.

Identification of critical risk set is based on the reconstruction results. A systematic approach is needed to minimize the impact of subjective judgment. Tree-based methods such as Fault Tree Analysis as well as other causal analysis methods, e.g., Root-Cause-Analysis and Barrier Analysis, will be considered in developing an appropriate analysis framework for naturalistic driving data.

The current analysis was conducted in a classical statistical framework and there are several benefits to extend this approach to Bayesian framework. Bayesian method has become popular in transportation safety study in recent years. Compared to the classical statistical method, Bayesian method has advantages of ease of interpretation, flexibility to accommodate spatial/temporal correlation, ability to incorporate prior information, and natural hierarchical structure in modeling multi-center/group studies. The most distinguishing characteristic of Bayesian method is its ability to incorporate *a priori* information. It is especially useful when

sample size in each individual study is small but there are multiple similarly structured studies available. With the popularity of naturalistic study, we expect there will be more naturalistic studies needing statistical analysis and appropriate Bayesian methodology will enable researchers to combine information from multiple sources to achieve more power in modeling.

REFERENCES

(1) 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., and Knipling, R.R. (2006). The 100-Car Naturalistic Driving Study: Phase II – Results of the 100-Car Field Experiment. (Interim Project Report for DTNH22-00-C-07007, Task Order 6; Report No. DOT HS 810 593). Washington, D.C.: National Highway Traffic Safety Administration.

(2) NHTSA, (2008) 2007 Traffic Safety Annual Assessment-Highlights; http://www-nrd.nhtsa.dot.gov/Pubs/811017.PDF

(3) Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of roadway geometric and environment factors on rural freeway accident frequencies. *Accident Analysis and Prevention* Vol. 27 pp. 371-89

(4) Abdel-Aty, M. and Radwan, E. (2000) Modeling Traffic Crash Occurrence and Involvement. *Accident Analysis & Prevention*, Vol. 32, pp. 633-642.

(5) Guo, F., Wang, X. and Abdel-Aty, M. A. (2009). "Corridor Level Signalized Intersection Safety Analysis Using Bayesian Spatial Models", *Proceedings of the Transportation Research Board 88th Annual Meeting*

(6) Miaou, S., Song, J. J. and Mallick, B. K. (2003). Roadway Traffic Crash Mapping: a Space–time Modeling Approach, *J. Transportation Stat.* Vol. 6, No. 1, pp. 33–57.

(7) Dang, J.N. (2007) Statistical analysis of the effectiveness of electronic stability control (ESC) systems – final report. Report no. DOT HS-810-794. Washington, DC: National Highway Traffic Safety Administration.

(8) Olson, P. L. & Farber, E**.** *(2003)* Forensic Aspects of Driver Perception & Response, *2nd ed.*, Tucson, AZ: Lawyers & Judges Publishing

(9) Jaquish, C. (2007). "The Framingham Heart Study, on its way to becoming the gold standard for Cardiovascular Genetic Epidemiology?" *BMC Medical Genetics*, 8(1), 63.

(10) Klauer, S.G., Dingus, T. A., Neale, V. L., Sudweeks, J.D., and Ramsey, D.J. (2006). The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data. (Report No. DOT HS 810 594). Washington, DC: National Highway Traffic Safety Administration.

(11) Liang, K.-Y., and Zeger, S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1), 13-22.

(12) Wierwille, W.W. and L.A. Ellsworth, Evaluation of driver drowsiness by trained raters, *Accid. Anal. Prevent.* **26** (1994) (5), pp. 571–581.

(13) Dingus, T. A., Antin, J. F., Hulse, M.C., & Wierwille, W. W. (1989). Attentional demand requirements of an automobile moving map navigation system. *Transportation Research, A23* (4), p. 301-315.

(14) Lambert, H.E. (1975) "Measures of Importance of Events and Cut Sets in Fault Tree" *Reliability and Fault Tree Analysis*, SIAM, 77-101,