CHAPTER

DATA ANALYTICS FOR SAFETY APPLICATIONS

9

Yuanchang Xie

University of Massachusetts Lowell, Lowell, MA, United States

9.1 INTRODUCTION

It is well-recognized that a majority of highway traffic collisions were caused by human errors. Such errors unfortunately are not well-captured with enough detail in police crash reports, which are often the main data source of many traffic safety studies. Currently, 90% of American adults [1] own a cell phone and 64% of them own a smartphone. The prevalence of mobile devices may make the human errors (e.g., distracted driving) problem even worse. Aside from causing distracted driving, the recent developments in wireless communications, mobile devices, and innovative mobile applications are generating a huge amount of rich data which could be useful for traffic safety studies. How to effectively utilize such rich information for safety applications is both challenging and rewarding. Besides human factors, roadway, weather, traffic, and vehicle conditions all have significant impacts on traffic safety and these impacts have been extensively studied. In this chapter, the existing highway traffic safety research topics are first summarized. The limitations of existing traffic safety studies due to lack of detailed data are then discussed. In addition, opportunities enabled by the recent developments in mobile devices and sensor technologies are presented.

9.2 OVERVIEW OF SAFETY RESEARCH

This section provides an overview of highway traffic safety research. It mainly focuses on safety data and data analytics related to human factors, crash count/frequency modeling, crash injury severity modeling, and safety of commercial vehicles. Additionally, some less common but very promising safety applications of data and data analytics are included at the end of this section.

9.2.1 HUMAN FACTORS

The main purpose of human factors research for traffic safety is to understand how drivers' performances are affected by various environmental, vehicle and roadway design, traffic, and psychological and physical factors. Given the fact that a majority of highway traffic collisions were caused by human errors, a lot of studies have been conducted to investigate the impacts of visibility and

Data Analytics for Intelligent Transportation Systems. DOI: http://dx.doi.org/10.1016/B978-0-12-809715-1.00009-2

lighting, human vehicle interface, driver assistance system design, and distracted driving on highway safety. Some of these research topics are closely related. For example, distracted driving has attracted particular attention in recent years, due to the increased usage of smartphones and other mobile and on-board electronic devices. Some of the on-board devices are designed to assist drivers with various driving tasks for improving safety and to connect with other drivers and the infrastructure. For these devices, properly designed human machine interfaces are very important for them to work effectively and safely as intended. These mobile and on-board devices have generated a tremendous demand for distracted driving research.

Human factors research typically requires collecting multiple drivers' behavioral data under a realistic and controlled setting. Such data collection efforts can be very expensive and time consuming. Therefore, most human factors studies were conducted in a laboratory environment that mimics practical driving scenarios, such as driving simulators ranging from a few thousand dollars to millions of dollars. Even with the most expensive driving simulator, drivers may still feel that they are participating in a controlled test and may not behave naturally in the same way as they drive on real-world roadways. To address such a problem, the Transportation Research Board (TRB) SHRP 2 Naturalistic Driving Study (NDS) [2] installed a variety of sensors in over 3400 participants' vehicles to collect their driving behavior data in a naturalistic setting continuously for more than 1 year. The NDS was designed primarily for better understanding of driver performance and behavior immediately prior to crash or near-crash events. The collected 2 million GB data (e.g., video, audio, speed, acceleration, and radar) also provides rich information for studying driver performance under various traffic, weather, and roadway conditions. It is anticipated that the TRB SHRP 2 NDS data can be useful to transportation and traffic engineers for the next two decades. A number of other countries and regions also initiated safety programs similar to the SHRP 2 NDS to collect data, including China, Europe, Australia, and Canada.

9.2.2 CRASH COUNT/FREQUENCY MODELING

Crash count/frequency modeling is another major component of traffic safety research. Numerous studies have been conducted to understand how different factors (e.g., lane width, pavement type, horizontal and vertical curves, and annual average daily traffic (AADT)) contribute to the occurrences of crashes. To perform crash count modeling, a road network is first divided into components such as intersections and road segments. Each of these components is further divided into homogenous units/sections for further analysis. For example, a 100-mile road is divided into 150 sections of varying lengths. For each section, the conditions (e.g., lane width, shoulder width, pavement type, and AADT) of the road should not change. The total number of crashes inside each section is considered the model output, while the variables associated with road conditions are the model inputs. This also explains why a road has to be divided into homogenous sections. For intersections, the total number of crashes at an intersection (or for a specific approach) can be considered as the model output, and intersection (or approach) related variables such as right-turn-on-red allowed, left-turn control type, turning traffic counts, lane width, and pedestrian volume are often used as the inputs.

Using crash count models, one can predict the number of expected crashes for a specific transportation facility under different conditions (i.e., by applying different safety countermeasures). In some cases, the observed and predicted crashes for a facility are very different. Such differences may be attributed to a number of reasons, such as the regression to the mean and omitted variable bias. If the observed numbers of crashes for a transportation facility over several years are consistently and significantly larger than the predicted numbers, this may suggest that there are other important explanatory variables not included in the model and this facility should be singled out and examined carefully. Therefore, the results of crash count modeling can be used for (i) hot spot analysis to identify high-risk intersections; (ii) identifying major crash contributors from those explanatory variables; (iii) conducting cost−benefit analysis to optimally allocate safety improvement funds; (iv) performing before and after studies; and (v) developing crash modification factors.

9.2.3 BEFORE AND AFTER STUDY

A before and after study is often used to evaluate the effectiveness of safety countermeasures. It typically applies four methods [3]: (i) naïve before and after study, (ii) Empirical Bayes (EB) method, (iii) before and after study based on a comparison group, and (iv) before and after study based on cross-sectional data. The naïve method simply compares the predicted crash count (using crash count models) without a safety treatment to the observed crash count with the safety treatment to find out whether this safety treatment is effective. A fundamental assumption behind this method is that the predicted crash counts without the safety treatment during the before and after periods are the same (or do not change significantly). Instead of the predicted crash counts, the EB method calculates an expected crash count, which is a weighted combination of predicted crashes and observed crashes. The expected crash count in the after period without a safety treatment is then compared to the observed crash count in the after period with the safety treatment.

For the comparison group method, a group of comparable sites is first identified. A safety treatment is applied to selected sites (called treated sites). During the after period, the observed crash counts for the treated and nontreated sites are compared to evaluate the effectiveness of the treatment. The cross-sectional method is very similar to the comparison group method. Strictly speaking, it is not a before and after method. The cross-sectional method first identifies a group of sites with and without a safety treatment. These sites should have very similar site characteristics such as traffic volume and road geometry. The only main difference is the safety treatment. Comparing the observed crash data from the two groups of sites may reveal the effects of the safety treatment. For both the comparison group and cross-sectional methods, it is critical to identify a large enough sample of sites of similar characteristics. This task in many cases is the main obstacle for conducting such before and after studies.

9.2.4 CRASH INJURY SEVERITY MODELING

Crash injury severity is used to understand the relationship between the injury outcomes of individual crashes (either for passengers or drivers) and explanatory variables. The results can be used to improve the design of vehicle, roadway, and traffic control devices and to provide justifications for new traffic laws and regulations. It is different from crash count modeling, which focuses on the total number of crashes of a transportation facility over a certain time period. Due to the different modeling outcomes, there are some differences in input variables between the two types of analysis. Injury severity modeling takes factors related to vehicle, driver, roadway, weather, traffic control, etc. into consideration, while crash count modeling mainly considers factors such as roadway, weather, traffic volume, traffic control, land-use, and sociodemographics.

The injury severity of a crash is normally labeled as “no injury,” “no injury but complaint of pain”, “non-incapacitating injury”, “incapacitating injury”, and “fatal” [4]. In the 2003 National Automotive Sampling System (NASS) General Estimates System Coding and Editing Manual [5], some general criteria are used to classify crash injury severity into those five categories. In the manual, fatal injury is interpreted as death from crash; incapacitated injury refers to severe injuries; nonincapacitated injury refers to other visible injuries such as bruise, abrasion, and swelling; possible injury means no visible injuries but subjects feel pain or faint; and no injury means property damage only. Given the categorical nature of the injury outcome, models such as ordered Probit model, ordered Logit model, multinomial Logit model (MNL), nested Logit model (NL), ordered mixed Logit model, heteroscedastic ordered Logit model, and Logistics regression have been commonly used. A variation of the injury severity analysis is to jointly model crash count and severity. A number of models have been developed for this purpose and will be detailed in the methodology section.

9.2.5 COMMERCIAL VEHICLE SAFETY

The research on commercial vehicle safety includes detecting fatigue driving, development and evaluation of Hours-of-Service (HOS) rules, policies and strategies for increasing seat belt usage, electronic logging devices, and nonintrusive truck and cargo inspection tools. Some of these research topics are interrelated. For example, the electronic logging devices are for better tracking commercial vehicle drivers’ activities such as on-duty, sleeper berth, and off-duty. In the United States, commercial vehicle drivers’ work activities are governed by the HOS rules set by the Federal Motor Carrier Safety Administration (FMCSA). Some important terms in the most recent HOS regulations (published in Federal Register on December 27, 2011) include the 11-hour driving limit, 14-hour limit, rest breaks, 60/70-hour on duty limit, 34-hour restart, and sleeper berth provision [6]. The main purpose of these HOS rules is to prevent fatigue driving from happening. Some studies focus directly on fatigue driving by developing advanced sensors and algorithms to detect driver fatigue based on drivers’ facial expressions, eye glances, etc. Other studies take an indirect approach and focus on the relationship between crash history and factors such as driving hours, trip start time, and rest breaks. By utilizing advanced statistical tools, these studies aim to associate the explanatory factors with different levels of crash risk, and the findings are used to further improve the existing HOS rules or to develop new commercial vehicle safety regulations. These studies described so far are for improving safety from the perspective of drivers. A number of studies have also been conducted to automate commercial vehicle inspection to ensure that vehicles are in healthy and safe conditions.

9.2.6 DATA DRIVEN HIGHWAY PATROL PLAN

The previously reviewed topics are mostly based on crashes that have already occurred. The purposes are to learn from historical events and to develop reactive solutions/strategies. Recently IBM [7] developed a proactive data-driven system to help the Tennessee Highway Patrol (THP) dynamically allocate their limited resources to combat traffic violations, mitigate crash risk, and better respond to crashes. Such a system is built upon the times and locations of previous crashes, DUI arrests, and public events. It also takes weather conditions and holidays into consideration. The system is able to predict the probabilities of future incidents (e.g., DUI) at various locations and times. The predicted results are provided in 4-hour increments. During a 6-month deployment of the system, the THP has seen a 6% decrease in fatal and seriously-injured crashes, 46% increase in seat belt citations, 34% increase in DUI arrests, and an 8.9% decrease in alcohol-impaired crashes.

This data driven safety predictive system can be considered as an extension of crash count modeling. The difference is that crash count modeling typically considers AADT as the input and cannot take the traffic volume variations into consideration. Also, it cannot correlate crashes with sporadic events such as holidays, public events, hurricanes, and snow storms. With the development of big data initiatives and the increased availability of data generated by mobile devices and other sensors, more detailed and real-time information can be fed into the safety predictive system for improved accuracy. Additionally, vehicle routing algorithms can be built into such a system to further improve highway patrol schedules and reduce costs. This system can generate data to optimally locate automated law enforcement devices such as radar speed sign.

9.2.7 DEEP LEARNING FROM BIG AND HETEROGENEOUS DATA FOR SAFETY

Although it is well recognized that a majority of highway traffic collisions were caused by human errors, there are many other factors (e.g., rain, fog, and stop-and-go traffic) that may contribute to human errors and subsequent crashes. Some of these factors are difficult to quantify, highly unstructured, and potentially correlated. Also, the corresponding data is in many different forms (e.g., text and video), may contain missing and erroneous information, and is being generated at various speeds. Developing a comprehensive model that can take all these factors into consideration for real-time crash risk prediction is both very challenging and desirable. Other than the above model developed by IBM for the THP, very few studies have been conducted along this direction. In a recent study, Chen et al. [8] proposed a deep model of Stack denoise Autoencoder to process big and heterogeneous data for predicting crash risk. They used GPS devices to anonymously track 1.6 million passengers' locations at fixed time intervals over a 7-month period and took the GPS activity records as the model input. The basic assumption behind this study was that areas with dense GPS records typically have high crash risk. Although the idea of using deep learning for modeling crash risk is plausible and the data size is big, the model input data considered in this study is oversimplified and does not take other important crash contributing factors into consideration.

9.2.8 REAL-TIME TRAFFIC OPERATION AND SAFETY MONITORING

Recently, Shi and Abdel-Aty [9] proposed a novel system based on real-time traffic sensor data to generate congestion and safety warnings. They used spot speed, volume, occupancy, and vehicle classification data from individual lanes as the system input. The data was collected using a 1-minute interval from three expressways in Florida through microwave traffic sensors spaced on average less than 1 mile from each other. In total, they obtained 1.5 million data readings. During the study period, 243 rear-end crashes occurred on these three highways. Up and downstream traffic data 5—10 minutes prior to these crashes were extracted. For comparison purpose, similar traffic data during noncrash periods for the same segments were also obtained. Specifically, the data considered in their study included traffic volume, truck percentage, average speed, standard deviation of speed, logarithm of the coefficient of variation of speed, speed difference between inner and outer lanes, number of lanes, posted speed limit, horizontal curvature, and existence of auxiliary lanes near ramps. Authors adopted the random forest method to identify important explanatory variables and applied a Bayesian logit model for predicting crash likelihood in real time. They also developed a First-Order Reliability Method (FORM) to determine the thresholds for triggering congestion and safety warnings that can be displayed on variable message signs. This pioneer study provides an excellent example of how increasingly available sensor data can be used to model crash risk at a much more detailed level. Compared to crash count modeling, the proposed modeling approach allows us to more closely examine the impact of variations in traffic flow parameters on crash risk. In future studies along this direction, traffic data provided by mobile apps (e.g., Waze) may be used instead of or in addition to microwave sensors as the model input.

9.2.9 CONNECTED VEHICLES AND TRAFFIC SAFETY

A main objective of connected vehicle technologies is to improve traffic safety. Connected vehicle research consists of two major components: One is focused on human factors and aims to facilitate the effective communications of critical information from on-board safety devices to drivers. The US Department of Transportation (USDOT) initiated several major safety programs [10] to demonstrate the benefits of vehicle-to-vehicle communications based connected vehicle technologies. One of them is the Safety Pilot Driver Clinics, which recruited volunteers to test driver vehicles equipped with on-board safety devices in a controlled environment. The purpose is to evaluate the effectiveness of various on-board device designs in communicating safety information. Another USDOT initiative is the Safety Pilot Model Deployment (SPMD), which installed vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) devices on approximately 3000 vehicles and 75 miles of roadways. These instrumented vehicles send out Basic Safety Messages (BSMs) at a 10 Hz frequency. The BSMs describe a vehicle's instantaneous position, speed, acceleration, vehicle status (e.g., lights, brakes, and wipers), and distances to surrounding objects. Instrumented vehicles can receive the BSMs sent by each other. If the received BSMs suggest that a crash is imminent if no proper actions are taken, an audio warning message will be triggered. The BSMs as well as drivers' responses to warning messages are recorded and archived for research purpose using a data acquisition system.

The USDOT SPMD project has generated a tremendous amount of rich information, providing great opportunities to closely examine crash risk and driver behavior at the microscopic level. Based on the SPMD data, Liu and Khattak [11] developed an algorithm to identify extreme events in terms of longitudinal and lateral accelerations. They utilized a one-day data set that consists of about 1 million BSMs from 155 trips made by 49 vehicles. The identified extreme events were later correlated with trip characteristics such as trip duration, number of turns, and average distance to the closest object. As a pioneer study utilizing the SPMD data, the findings and approaches adopted in this research are both encouraging and interesting. To further extend this study, the approach in crash count modeling can be considered to correlate the number of extreme events in each roadway segment with factors such as real-time traffic, roadway geometry, control, land use, and weather. In addition, the criteria of defining extreme events can be modified to include distance to the closest object. This modification will allow researchers to take near-crash events into consideration, which is an important aspect of traffic safety but cannot be properly examined based on traditional crash data.

9.3 SAFETY ANALYSIS METHODS

Over the past few decades, abundant statistical, artificial intelligence, and machine learning methods have been applied to model traffic safety data. These safety analysis methods are reviewed and summarized below. The purpose of this summary is not intended to provide a detailed description of these methods, which can be readily found in many textbooks and journal articles. Instead, this summary is to outline the innovative applications of various statistical, artificial intelligence, and machine learning methods and to comment on their pros and cons.

9.3.1 STATISTICAL METHODS

9.3.1.1 Count data modeling

Traffic safety research often involves count data, for example, number of crashes, number of traffic violations, and number of extreme events. Initially, multiple linear regression was widely used for modeling safety related count data. Multiple linear regression, when not corrected for unequal variance, assumes that the number of crashes follows a normal distribution. It is inadequate for analyzing discrete, non-negative, sporadic, and asymmetrically distributed random events. Generalized Linear Models (GLMs) later became very popular for modeling crash count data, including Poisson regression, Poisson-gamma or Negative Binomial (NB) regression, Gamma regression model, and other variations of the NB regression model [12]. It is generally agreed that when the sample variance is significantly greater than the sample mean, NB models should be used in lieu of Poisson regression models. On the other hand, if the sample variance is significantly smaller than the sample mean, which is defined as underdispersion, Gamma models are the models of choice. Since crashes are rare events, Zero-Inflated models (Poisson and NB) have been proposed for modeling crash count data with an apparent excess of zero observations. However, their applications have been discredited when the characteristics and the nature of the data do not warrant the application of such models [13].

The previously mentioned models consider only one set of count data as the dependent variable. In some cases, it is desirable to model several count variables simultaneously. For example, a safety analyst may want to jointly model the numbers of head-on and rear-end crashes. Another reason is that fitting a crash count model typically requires a large sample, particularly for rare events such as vehicle−bicycle crashes. These rare crash events may be correlated to more common crashes such as multiple vehicle crashes. By jointly modeling vehicle−bicycle and multiple vehicle crashes, a smaller sample size might be sufficient compared to the sample size requirement of modeling vehicle−bicycle crashes separately. For this purpose, Bayesian the multivariate generalized linear mixed model [14] and the multivariate Poisson regression model [15] have been proposed.

All regression-based models such as the ones described above share one common characteristic: they need a well-defined function relating the dependent variable (crash counts) to the independent (explanatory) variables. This function is often referred to as the “rate function,” “functional form,” or “safety performance function (SPF)” in the traffic safety literature. The specification of the functional form can significantly affect the goodness-of-fit of GLMs. The functional form is usually estimated via a trial and error process based on the safety analyst’s experience, and can seldom be completely optimized. Normally, the functional form depends on the nature of the data and its selection should be based on the combination of statistical and logical properties linking the crash count data to the covariates of the model [16].

To get around the problem of specifying the functional form, a number of studies adopted Generalized Additive Models (GAMs) for modeling count data. GAMs use smooth splines to replace the parametric terms in GLMs. Although the GAM results suggest that a nonlinear functional form is more appropriate and performs better than GLMs in many cases, some researchers argue that such nonlinear trends in the function form may be caused by the omitted variable bias, not by the true physical relationship between crash counts and covariates. For flexible models (e.g., GAMs) that can handle nonlinear relationships, the model estimation process may adopt nonlinearities to account for unobserved heterogeneity due to omitted important covariates. Also, Mannering and Bhat [17] argued that parsimonious models due to omitted variables can generate biased parameters. Such parsimonious models cannot produce accurate crash count predictions, since changes in many other significant crash predictors are not taken into account. Also, these models cannot be used to develop safety countermeasures because only limited covariates (e.g., traffic volume) are included. Due to the difficulty in collecting detailed crash-related data, sometimes there may be a need in practice to develop relatively simple models with limited covariates available. GAMs are nonparametric models. It is difficult to characterize their modeling results using mathematical formulas, which are sometimes needed by safety analysts. A feasible and promising solution is to first apply GAMs. Based on the shapes of the GAM splines, one can develop appropriate GLMs.

For a comprehensive list of statistical models for crash count data modeling, readers are referred to Mannering and Bhat [17]. They conducted a comprehensive review of existing crash frequency literatures. They discussed potential issues caused by insufficient data (e.g., parsimonious models, unobserved heterogeneity, and biased parameters) and provided suggestions for mitigating these problems, including finite-mixture/latent-class and random parameters models. It would be very interesting to conduct a thorough evaluation of such advanced modeling techniques in future studies. With these advanced models, the nonlinear relationships identified in GAMs studies may no longer exist. If this is the case, the nonlinear relationship may simply be caused by the unobserved heterogeneities in the data that were previously not captured by conventional modeling techniques.

Finally, as Miaou and Lord [16] noted, developing a model that fits a particular crash count data set well is no longer a major challenge. This can be accomplished using smoothing techniques (e.g., GAMs) or other universal approximators such as multilayer feedforward neural networks to be discussed later. In addition to model goodness-of-fit, criteria based on logic (e.g., reason, consistency, and coherency), flexibility, extensibility, and interpretability should be considered in determining the functional form.

9.3.1.2 Categorical data modeling

Many statistical methods have been applied to traffic crash injury severity modeling, including ordered Probit model, ordered logit model, MNL, nested logit model (NL), ordered mixed logit model, heteroscedastic ordered logit model, and logistics regression. A comprehensive review of crash injury severity models can be found in Ref. [18]. Among these models, the ordered logit/Probit models and the MNL are the most widely used ones. In the ordered logit/Probit models, each explanatory variable has one coefficient, which means that the effects of this particular variable on all injury outcomes are restricted to be the same. In the MNL model, each injury outcome has a separate severity function (i.e., utility function in discrete choice modeling literature) and two severity functions can include different sets of explanatory variables. This modeling structure is quite flexible and can readily handle the distinct effects of the same variable on different injury outcomes.

Although the MNL model has some advantages in terms of flexible model structure, it has certain limitations due to its independence from irrelevant alternatives (IIA) property, which originates from the independence and identical distribution (IID) assumption of the error terms in each severity function. This limitation of the MNL model is demonstrated in a previous study by Abdel-Aty [19]. In his research, Abdel-Aty compared ordered Probit, MNL, and nested logit models for injury severity analysis. His research finding suggested that the MNL model produced even worse fitting results than the ordered Probit model. Although the nested logit model generated slightly better fitting results than the ordered Probit model, the author still recommended the ordered Probit model for their study after considering the difficulty in specifying the nested structure.

Recently, the latent class logit (LCL) model and the random parameters logit model were introduced to model traffic crash injury severity data. The LCL model is based on the MNL model. Similar to the standard MNL model, the LCL model has a flexible structure that can readily take into account the different effects of the same variable on each injury outcome. The key benefit of the LCL over the MNL is that its special structure has the potential to overcome the problems associated with the IIA property. To better explain the proposed LCL model and also to make this chapter self-contained, a very brief description of the standard MNL model is provided here. More details about the MNL model and the fundamental theory behind it can be found in Ref. [20]. For each single-vehicle traffic crash, assume there are k possible injury outcomes for the driver. The MNL model first constructs a severity function for each injury outcome as shown in Eq. (9.1).

$$U_{ij} = V_{ij}(\beta) + \varepsilon_{ij} \tag{9.1}$$

where U_{ij} is the severity function for the j^{th} possible injury outcome of the i^{th} driver involved in a traffic crash, with $i = 1, \ldots, n$ and $j = 1, \ldots, k$; $V_{ij}(\beta)$ is a linear-in-parameters combination of explanatory variables and is the deterministic part of the severity; β is a coefficient vector; ε_{ij} is an independent and identically distributed random variable following Gumbel distribution. Given the estimated coefficient vector β, the probability that the j^{th} injury outcome may happen is:

$$\begin{aligned} Prob(j|\beta) &= Prob(V_{ij}(\beta) + \varepsilon_{ij} > V_{it}(\beta) + \varepsilon_{it}, \forall t \neq j|\beta) \\ &= \frac{\exp(V_{ij}(\beta))}{\sum_{m=1}^{k} \exp(V_{im}(\beta))} \end{aligned} \tag{9.2}$$

One important assumption of the MNL model is that the random terms, ε_{ij}, of each severity function are independent and identically distributed (IID). However, this often is not the case due to many possible reasons. For instance, traffic crash injury severity is affected by various contributing factors. Therefore, the deterministic parts of each severity function, $V_{ij}(\beta)$, should consist of many explanatory variables. In real-world applications, it is very difficult to identify and collect all the relevant input data and include them in the severity functions. If some important explanatory variables are not included, the unobserved random portions of these severity functions are likely to be correlated, which leads to the violation of the fundamental IID assumption. The violation of the IIA property or IID assumption may lead to biased parameter estimates. It can also generate systematic errors in the choice probabilities, and a typical example is the famous red-bus—blue-bus problem. When the violation of IID assumption happens, one can choose to use a different type of model that is able to handle the correlation among the random terms of different alternatives. Another option is to modify the deterministic portions of the severity functions to capture the unobserved correlation, so that the remaining random terms can become independent.

To address the potential IID assumption violation, the LCL model is proposed for modeling traffic crash injury severity. The LCL model can be considered as a special form of the mixed MNL model. For a typical mixed MNL model, the probability that injury outcome j will happen is described in Eq. (9.3).

$$Prob(j) = \int Prob(j|\beta) f(\beta) d\beta \tag{9.3}$$

A major difference between the standard MNL model and the mixed MNL model is the coefficient vector β. The standard MNL model assumes a constant β vector, while the mixed MNL model considers vector β as a mixture of random coefficients (φ) and constants (α). Thus, the initial severity function becomes:

$$U_{ij} = V_{ij}(\beta) + \varepsilon_{ij} = \alpha^T W_{ij} + \varphi^T X_{ij} + \varepsilon_{ij} \tag{9.4}$$

where X_{ij} is a set of explanatory variables with random parameters and W_{ij} represents the explanatory variables with fixed parameters. By including the random coefficients, different injury severity outcomes become correlated even though their error terms, ε_{ij}, are still assumed to be independent and identically distributed. This is because $cov(U_{ij}, U_{ik}) = E(\varphi^T X_{ij} + \varepsilon_{ij})(\varphi^T X_{ik} + \varepsilon_{ik}) = X_{ij}^T \Omega X_{ij}$ [20]. Such a correlation can be very useful in addressing the aforementioned IID/IIA problem.

The mixed MNL model was first introduced into transportation research in 1980 [21]. It has since been applied to a number of areas due to the wide availability of computer simulation. To apply the mixed MNL model, distributions of each random coefficient in vector β must be explicitly specified, which is not a trivial task. To get around this problem, the LCL model was proposed [22], which can be considered as a special form of the mixed MNL model. In the LCL model, β takes a finite set of values and the integral in Eq. (9.3) is replaced by a summation of weighted $Prob(j|\beta)$ over all β values. In this case, the probability for injury outcome j to happen is

$$Prob(j) = \sum_{m=1}^{M} Prob(class = m) \cdot Prob(j|\beta_m) \tag{9.5}$$

The LCL model assumes that the entire crash data set can be categorized into M different classes. Each crash event belongs to different classes with certain probabilities that are not revealed to the analyst. $Prob(j|\beta_m)$ in Eq. (9.4) can be determined similarly as $Prob(j|\beta)$ in Eq. (9.2) with β being replaced by β_m. $Prob(class = m)$ is the probability that a crash event belongs to class m and can be determined by Eq. (9.5).

$$Prob(class = m) = \frac{\exp(V_{im}(\theta))}{\sum_{c=1}^{M} \exp(V_{ic}(\theta))} \tag{9.6}$$

It can be seen that the class probability $Prob(class = m)$ is also determined based on the MNL framework. $V_{im}(\theta)$ in Eq. (9.5) can be a linear-in-parameters combination of a constant and several covariates. In case that no appropriate covariates can be identified to enter $V_{im}(\theta)$, only a constant needs to be chosen. Compared to the mixed MNL model, the LCL model takes only a finite set of parameters β. This can potentially save the computation time for model fitting. In addition, the LCL model can avoid the trouble of specifying the probability distributions of each random coefficient. For more details about LCL models, readers may refer to Ref. [22].

9.3.2 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Compared to statistical regression models, the application of neural network models for crash data modeling has received less attention. A primary reason is attributed to the complexity for estimating these models. Other criticisms that have impeded on their use include the following: (i) over-fitting when the sample size is small; and (ii) unlike regression models, neural network models essentially work as black-boxes and do not generate interpretable parameters for each explanatory variable. For the first criticism, it has been reported that similar to neural network models regression models may also suffer from the over-fitting problem [23]. To respond to the criticism that neural network models work as black-boxes, Fish and Blodgett [24] and Delen et al. [25] proposed a sensitivity analysis approach to quantify the effect of each input variable on the network output. Despite these disadvantages, neural network models have some significant advantages over statistical regression models. First, neural network models do not require the establishment of a functional form. Statistical regression models, on the other hand, have to specify an approximate functional form linking the dependent variable and independent variables (note: the perfect functional form is unknown). Second, research has shown that standard multilayer feed-forward neural network models can approximate any continuous function defined on a compact set with arbitrary accuracy given enough hidden neurons are used [26], though this strong ability may sometimes lead to over-fitting.

To avoid the over-fitting problem and improve the generalization ability of neural network models, a number of approaches have been proposed in the literature. One of these approaches includes adding a weight-decay or regularization term in the estimation process [23]. However, Marzban and Witt [23] discussed that this improvement of generalization of neural network models impedes on their nonlinear approximation ability.

$$E_r = \eta \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 + (1-\eta) \frac{1}{n_p} \sum_{j=1}^{n_p} (\psi_j)^2 \tag{9.7}$$

where,

y_i = observed crash count at site i;
$\hat{y}_i$ = predicted crash count at site i;
n_p = number of network parameters, including weights and bias;
ψ_j = the j^{th} element in the network parameter vector; and
η = performance ratio.

On the other hand, they noted that the Bayesian inference method can improve the neural networks generalization ability without compromising the nonlinearity properties. The development of Bayesian Neural Networks (BNN) was first initiated by Mackay [27] and further developed by Neal [28]. Based on the previous BNN models, Liang [29] introduced an improved BNN model by incorporating a prior on both the network connections and the weights. This modification gives the network more flexibility for choosing hidden neurons and input variables. In Liang's study, the proposed BNN model was trained using an Evolutionary Monte Carlo (EMC) algorithm and was compared to a number of popular models such as the BPNN and the Box–Jenkins model for nonlinear time series forecasting. The testing results showed that the proposed BNN model consistently outperformed other prediction methods. BNN models began to gain popularity in late 1990s and have been used even more since 2000.

Xie et al. [30] first introduced BNN into traffic safety and applied it to model highway crash counts. They considered road segment length, average daily traffic volume, right shoulder width, and lane width as the model input. The BNN model used in their study was initially proposed by Liang [29]. In the BNN model, Liang used a fully connected multilayer feed-forward network structure with one hidden layer. The simplified network structure is illustrated in Fig. 9.1.

The network structure of the BNN model is very similar to that of multilayer feedforward neural networks. They are different in the prediction mechanism and the training process. The following example is given to illustrate the differences in the prediction mechanism. Assume there are n sets of accident data $(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$, where x_i represents covariates and y_i is for observed crash counts. Let θ denote all the neural network parameters or weights, β_j, α_k, and γ_{jk}

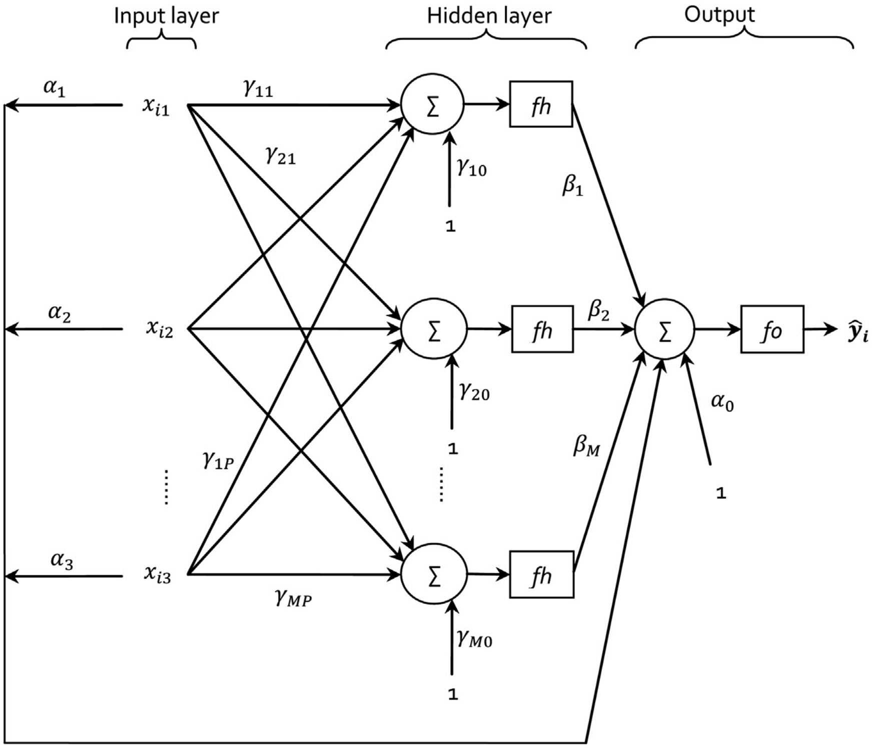

FIGURE 9.1

A fully connected multilayer feed-forward neural network.

($j = 1,...,M$; $k = 0,...,P$), in Fig. 9.1. The predicted number of crashes for site i using BNNs is given by Eq. (9.8) [28].

$$\hat{y}_i = \int f_B(x_i, \theta) \cdot P(\theta|(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n))d\theta \tag{9.8}$$

where $f_B(x_i, \theta)$ is defined as

$$f_B(x_i, \theta) = \alpha_0 + \sum_{k=1}^{P}(\alpha_k \cdot x_{ik}) + \sum_{j=1}^{M}\left\{\beta_j \cdot \tanh\left(\sum_{k=1}^{P}\gamma_{jk} \cdot x_{ik} + \gamma_{j0}\right)\right\} \tag{9.9}$$

$P(\theta|(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n))$ in Eq. (9.8) is the posterior distribution of θ given observed data $(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$. One can see that the main difference between BNNs and multilayer feedforward neural networks is that for multilayer feedforward neural networks the network parameter Ψ is fixed; while for BNNs the network parameter θ follows a certain probability distribution, and the prediction process for BNNs is to evaluate the integral of $f_B(x_i, \theta) \cdot P(\theta|(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n))$ over all possible values of θ as shown in Eq. (9.8). The actual BNN model is more complicated than the example given above. Readers are referred to Liang [29] for a more detailed description of the BNN model and its training algorithm.

Neural network models have been long criticized for not being able to generate interpretable parameters for each explanatory variable, and this is one of the major reasons that neural network models have not been widely used for modeling crash frequency. To minimize this problem, a method proposed by Fish and Blodgett [24] can be used to analyze the sensitivity of each explanatory variable. This method has also been used by Delen et al. [25] in their application of neural network models in accident injury severity study. The basic idea of this method is that for each explanatory variable, one keeps all other explanatory variables unchanged and perturbs the current variable's value within a reasonable interval. At the same time, the corresponding variation of network output is recorded, and from this variation, one can find the effect of changing single explanatory variable on the network output. The idea of this method is simple, but it can be useful to mitigate the black-box issue and help to illustrate the training result of neural networks. It should be pointed out that the explanatory variables may not be completely independent of each other. Due to the complicated relationship between crash frequency and all explanatory variables, if one changes the value of any of the remaining explanatory variables, the relationship between crash frequency and the current explanatory variable may change accordingly. Other than the neural networks, a number of machine learning methods have been adopted for safety analysis, including support vector machines, random forest, classification and regression trees, and clustering.

9.4 SAFETY DATA

Due to the difficulty involved in capturing the entire process of a crash event, most existing traffic safety studies were based on information extracted from police crash reports. Additional traffic, roadway geometry, and pavement information is obtained from various state and local agencies such as Departments of Transportation (DOTs) and Departments of Public Safety. This section discusses how necessary data for highway traffic safety analyses can be obtained from various sources.

9.4.1 CRASH DATA

All states in the United States have a well-developed accident report form to document crashes. Samples of these forms can be easily found on the internet. This form typically contains information related to driver, passenger, vehicle, weather, location, time, roadway geometry, traffic control, and cause(s) of the crash. Such data can be requested from State DOTs, Departments of Public Safety, or city police departments depending on where a crash occurred. For many years, data from this form has been the primary information source for traffic safety analyses such as crash count modeling and crash injury severity modeling. For crash injury severity analyses, such data in general is sufficient. While for crash count modeling, additional risk exposure data (e.g., traffic volume) is needed.

Although these accident reports provide critical and useful information for understanding how crashes occurred, there are still a number of issues with them: (i) many of these reports were filled out by hand. Turning them into digital forms for in-depth analysis is time consuming and may introduce errors; (ii) the quality of the reports varies. For example, some police officers may draw a clear diagram showing the precise location of a crash, while other officers may simply write down "near intersection"; (iii) it is not uncommon to see missing values and typos in these reports; (iv) some drivers involved in property-damage-only (PDO) crashes choose to resolve the issues among themselves and cause underreported PDO crashes; (v) near-crash events, which are also very important, are not captured by these reports; and (vi) such reports cannot reflect what exactly has happened prior to a crash. This probably is the most significant weakness of such a reporting system. Driver(s) involved in a crash will provide information regarding the causes(s), and the responding police officer will also exercise his/her judgment to determine who is at fault. However, such information is not detailed enough to reconstruct the entire crash events. Sometimes driver(s) cannot even recall what exactly has occurred and what could have been done to prevent it due to distraction or fatigue.

9.4.2 TRAFFIC DATA

Traffic data is an important index for measuring crash risk exposure. Intuitively, an intersection or road segment with heavy traffic is likely to experience more crashes than a similar transportation facility with light traffic. Therefore, traffic volume data is often needed for crash count modeling. For most existing crash count studies, AADT is used. As per the requirement of the FHWA Highway Performance Monitoring System (HPMS), all states are required to collect vehicular traffic volume data from selected roads and intersections. Typically, there are permanent counting stations mostly on interstate highways that collect traffic count data continuously, 365 days in a year. Therefore, the AADT data for them can be precisely calculated. For other roads and intersections, short-term traffic counts (a few hours to several days) are collected from them every 1−3 years to satisfy the needs of both HPMS and local transportation agencies. Clearly, not all intersections and road segments will be covered in each year's data collection. Different states have their own methods to estimate traffic volumes for these facilities. Data from the permanent counting stations is used to derive growth factors and seasonal factors to facilitate such estimations. In some cases, ad-hoc data collections are needed in order to further improve the AADT estimation accuracy.

To model vehicle—pedestrian and vehicle—bicycle crashes, additional pedestrian and bicycle volumes are needed. Unfortunately, the existing data collection efforts in many places are focused primarily on vehicular traffic. There are no well-developed official systems like the HPMS to coordinate the collection of pedestrian and bicycle volumes and to share data, and there certainly is a need for doing so. Based on some previous studies [31—33], intersection pedestrian volumes are affected by population density, median household income, and area type, although median household income was found to be an insignificant factor by other researchers [12]. In another study, Schneider et al. [34] identified four significant covariates for predicting weekly pedestrian volumes. These covariates are the total population within a 0.5-mile radius, number of jobs within a 0.25-mile radius, number of commercial retail properties within a 0.25-mile radius, and the presence of a regional transit station within a 0.1-mile radius of an intersection. Without pedestrian volume data available, safety analysts may consider adopting existing regression models. The required input data such as population, employment, and land-use data is usually available from regional transportation planning agencies. For safety analysts with short-term pedestrian volume data available, they may also need to estimate annual average daily pedestrian volumes. Recently, the National Bicycle and Pedestrian Documentation (NBPD) project developed a strategy to expand hourly pedestrian data into daily, weekly, monthly, and yearly pedestrian volumes [35]. This strategy has been applied to areas such as Denver, CO [36]. More information about this method can be found at the NBPD website http://bikepeddocumentation.org/.

AADT is so far the most widely used crash risk exposure for crash count modeling. However, it cannot describe the variations of traffic within a day. Crashes may occur during specific time periods with unique traffic flow characteristics (e.g., stop-and-go traffic). Such characteristics are unlikely to be revealed by studies using AADT data. Few people would argue against considering more detailed traffic flow data. The main issue is with how to collect, process, and archive such large data sets involving vehicles, pedestrians, and bicycles at a large scale. In the future, data generated by mobile devices may be used as a surrogate crash risk measure or to estimate the daily variations of traffic.

9.4.3 ROADWAY DATA

Data such as roadway geometry, pavement conditions, guardrail, and traffic signs is important for many traffic safety studies, including human factors, crash count modeling, and crash injury severity modeling. Roadway geometry (e.g., horizontal and vertical curves) data can typically be obtained from state DOTs. For the remaining data sets, most state DOTs have them available. A problem is that these data sets may not be updated as frequently as safety analysts would like due to high data collection costs. Some advanced mobile lidar sensors have been developed and successfully applied to collect 3D profile data at high speeds (e.g., 70 mph). From the 3D lidar data, detailed features such as existence of guardrail and slope can be derived. For pavement conditions such as the International Roughness Index (IRI), some more accurate laser sensors are often used that can generate vertical measurements with submillimetre accuracy. Although these laser sensors can be mounted on a vehicle that collects data at 70 mph, they typically can only collect pavement condition data for one lane during a single run. Both lidar and laser sensors are fairly expensive and cost from $200,000 to over $0.5 million. This limits the frequency for the pavement condition data to be updated.

In future studies, low-cost sensors and crowdsourcing strategies may be developed to address this issue. Recently, a mobile app has been developed for volunteer drivers to report potholes [37]. It utilizes accelerometers available on smartphones to detect potholes. This crowdsourcing strategy saved a tremendous amount of data collection trips. If the sensors for pavement condition inspection can be made more portable and less expensive, crowdsourcing may have great potential to reduce the cost for pavement inspection and to provide valuable data for safety research.

9.4.4 WEATHER DATA

For crash injury severity modeling, weather data is often obtained from individual police crash reports, which is very straightforward. Few crash count studies have taken weather into consideration. This is probably because of the difficulty in quantifying weather data over an extended time period (e.g., one year) for many geographic entities (e.g., road segments and intersections). With detailed weather data becoming increasingly available (e.g., National Weather Service provided by the national Oceanic and Atmospheric Administration), it is anticipated that weather data will be incorporated into more traffic safety studies for analyzing historical crash data and for providing real-time crash risk estimations.

9.4.5 VEHICLE AND DRIVER DATA

For crash injury severity analysis, some basic vehicle (e.g., defects) and driver (e.g., age and gender) data can be obtained from crash reports directly. For crash count modeling, additional vehicle and driver data can be obtained from census and Department of Motor Vehicles (DMV) in many cases. Some states do not require annual vehicle inspections. Thus, the corresponding updated vehicle information is not available from DMV. The above mentioned vehicle and driver information in general can meet the basic needs of crash count and injury severity modeling and is being considered in existing studies. It would be ideal that more detailed driver and vehicle data prior to a crash can be obtained, such as tire pressure, driver fatigue, blood pressure, and vehicle trajectory. Given the rapid developments in connected vehicles and wearable electronic devices, it might be possible that a data acquisition system can be developed that keeps track of the detailed vehicle and driver information. NDS described in the following section demonstrates that this idea is technically feasible. However, its implementation at a large scale is a complicated issue and is beyond the scope of this discussion.

9.4.6 NATURALISTIC DRIVING STUDY

A successful example of applying advanced sensor technologies to human factors research is the TRB SHRP 2 NDS. The NDS installed radar, GPS, and video cameras in over 3400 participants' vehicles to collect data continuously for more than 1 year in a naturalistic setting. It was designed specifically to collect data to better understand driver performance and behavior immediately prior to crash or near-crash events. The collected data includes:

- Driver characteristics: vision test results, demographic information, and physical and psychological characteristics;

- Lighting, weather, roadway surface condition, traffic control, and driver eye glance;
- Video data showing forward and rear roadway views, driver views, snapshots of passenger seats;
- Vehicle characteristics (e.g., year, make, and model), vehicle lateral and longitudinal accelerations, gas pedal position, lane offset, turn signal use, brake application, distances to front vehicles, and distance changing rates; and
- Horizontal curvature (e.g., radius and length), grade and super elevation, lane width and type, shoulder type, intersection location and control, and locations of speed limit signs, median, and rumble strip.

This massive multiyear data collection project was completed in 2013 and generated 5.4 million trip files in six states (Indiana, Pennsylvania, Florida, New York, North Carolina, and Washington). The collected roadway data is being hosted at Iowa State University and the remaining data is being hosted at the Virginia Tech Transportation Institute (VTTI). Although the main objective of the NDS was to improve highway safety, the NDS data has also generated significant interest among transportation researchers in many other areas.

A number of studies have been or are being conducted utilizing the NDS data [38]. The Federal Highway Administration (FHWA)/AASHTO established an Implementation Assistance Program (IAP) and funded nine research projects in December 2015 to utilize the NDS data for addressing various safety issues. These projects are summarized in Table 9.1.

9.4.7 BIG DATA AND OPEN DATA INITIATIVES

Big data has attracted tremendous attention from researchers in many fields. The large data sets being generated continuously from various sources can potentially enable many innovative safety research and applications. An excellent example is the real-time traffic operation and safety monitoring study conducted by Shi and Abdel-Aty [9]. Their study utilized very detailed (archived in 1-minute intervals) real-time traffic flow data (5−10 minutes prior to a crash) from microwave traffic sensors to model crash risk at a much detailed level. Many states and cities are making efforts to publish transportation related data and to encourage third-parties to develop various applications. For example, through the MassBigData initiative MassDOT has made real-time travel times on I-93, I-90, and Route 3 available. They also publish scheduled roadway events and real-time 511 traffic camera data online. The camera data may be used to extract detailed traffic flow information as that used by Shi and Abdel-Aty [9].

Using the connected vehicles data published through the FHWA Research Data Exchange program (https://www.its-rde.net/home), Liu and Khattak [11] analyzed extreme events based on driver behavior data, which is different from the Shi and Abdel-Aty [9] study that depended on traffic flow data. Liu and Khattak [11] only utilized one day of data from the SPMD project. Currently, there are over 100 GB of such data awaiting further exploration. The SPMD data set is similar to the NDS data set in that it focuses on the behavioral data of drivers, contains detailed vehicle dynamics information, and includes distances between the subject vehicle and other objects. The SPMD data set was collected from Ann Arbor, Michigan. The version published through the FHWA Research Data Exchange program does not include forward and rear view videos, while the NDS data set was collected from six sites located throughout the United States and covers a wide range of traffic and infrastructure conditions.

Table 9.1 Projects Utilizing NDS Data for Addressing Safety Issues

Project Title	State	Topic Area	Study Objective
Understanding interactions between drivers and pedestrians at signalized intersections	Florida	Pedestrian safety	• Effectiveness of pedestrian features • Driver characteristics, compliance with pedestrian features, and interactions between drivers and pedestrians • Driver distraction by pedestrians and pedestrian features
Evaluating the causes of roadway departures	Iowa	Roadway departures	• Model roadway departure crashes based on crash and near crash events • Use lateral position as a crash surrogate • Consider roadway, driver, and environmental characteristics
Identifying the interrelationships among speed limits, geometry, and driver behavior	Michigan	Speeding	• Investigate driver speed selection related to speed limits for different facilities • Identify countermeasures to reduce speeding related crashes
Examining the influence of roadway design features on episodic speeding	Washington State	Speeding	• Identify roadway design and traffic control features that can effectively prevent speeding from happening
Evaluating work zone safety	Minnesota	Work zones	• Identify factors that will increase/decrease work zone crash odds • Identify driver reaction point to traffic signs and presence of queues • Driver speed prediction model
Evaluating the interaction of traffic on rural, two-lane roads	North Carolina	Horizontal and vertical curves	• Identify the most dangerous combinations of horizontal and vertical curves • Identify countermeasures (e.g., advance warning and in-lane rumble strips) that can improve driver performance in those dangerous segments
Assessing driver behavior near closely spaced interchange ramps	Utah	Interchange ramps	• Examine crash and near-crash events at freeway weaving areas (on ramp followed by an off ramp) and identify safety countermeasures
Investigating driver performance and behavior in adverse weather conditions	Wyoming	Adverse conditions	• Investigate driver behavior under heavy rain, fog, snow, and ice conditions
Assessing the impacts of roadway lighting on night-time crashes	Washington State	Roadway lighting	• Identify critical lighting values for developing lighting design standards

In the future, the two ideas based on the Florida microware sensor data and the SPMD data may potentially be combined. In this way, both driver behavior and traffic conditions can be considered. Other than the MassBigData initiative, many states and cities have either formally or informally established their big data programs. Also, the concept of connected vehicles is evolving quickly from research labs or test beds to real-world implementations as evident in several major deployment initiatives funded by the USDOT in New York, Florida and Wyoming. The developments in big data and connected vehicles will generate additional detailed and informative data for traffic safety research and applications in the near future.

9.4.8 OTHER DATA

In the future, the concept of connected vehicles can be further extended to connect vehicles with drivers' wearable electronic devices, mobile devices, vehicle On-Board Diagnostics (OBD) system, dashboard camera, and radar sensor to generate rich safety related information. Both the SPMD and NDS studies also depend on the OBD system for obtaining accurate speed and acceleration data. This ubiquitous connectivity will enable us to understand how drivers' physical and psychological conditions affect their safety performance. These data can be shared using mobile apps (e.g., Waze) in real time and be further correlated with roadway and traffic data. The information (especially those near-crash events) from multiple drivers can be used to dynamically identify safety hazards caused by traffic, roadway, traffic control, and driver characteristics, generating real-time warning messages to be sent back to drivers via the mobile app.

9.5 ISSUES AND FUTURE DIRECTIONS

9.5.1 ISSUES WITH EXISTING SAFETY RESEARCH

Currently, many human factors studies are based on driving simulators. Even for the most sophisticated driving simulator, drivers may still not behave in the same way as they drive on a real road. The recently completed NDS has partially addressed this issue. By installing a data acquisition system in participants' vehicles, their naturalistic driving behaviors have been accurately recorded. The NDS data has inspired many innovative new ideas for human factors research and some of them have been recently funded by the FHWA IAP. An unresolved issue is that the NDS data covers limited traffic, roadway, control, and weather scenarios. It cannot be used to study new scenarios (e.g., new traffic control signs and roadway geometry designs). Some existing scenarios may only show up in the NDS data set with very limited sample sizes. In this case, no valid conclusions can be developed. Also, the NDS did not install any on-board warning system or eye tracking system on those instrumented vehicles. Given the significance of distracted driving and connected vehicles, it is unfortunate that the NDS data cannot be used to study distracted/fatigue driving and how drivers react to warning messages.

Most existing crash count studies are based on police crash reports, AADT, and roadway data. There are a number of issues with these studies: (i) they cannot take near-crash events into consideration. Since crashes are rare events, the number of observed crashes for a segment is often zero. Excess zero observations may cause problems to regression models. Using near-crash events as a surrogate risk index can help to address this model fitting issue and generate interesting and informative results. Also, introducing near-crash events can be particularly useful for before and after studies; (ii) crash risk at a particular location may be time-dependent. It changes as the corresponding traffic and environment conditions vary. Such a time-dependent feature cannot be captured by existing crash count models based on AADT data; (iii) for fitting crash count models, roads have to be divided into homogeneous segments. Some crashes may occur near the boundary of two segments. Simply classifying them into one segment may affect the model fitting result and the final conclusions.

Previously, crash injury severity and crash count studies focused on analyzing historical crash data obtained from crash reports and aimed at improving vehicle design, roadway design, traffic control, etc. However, from the information in a crash report, it is difficult to tell what could have been done to prevent the crash from happening in the first place. With the NDS data, similar crash and near-crash events can be compared to identify key factors that led to different consequences (i.e., crash vs. near-crash). Such comparison and analysis results based on real-world data can be very useful, especially for designing the control algorithms for connected and autonomous vehicles (CAVs) in the future. The research findings can be used to determine the best time for the automated system to take over the control or for the on-board computer system to alert the driver.

9.5.2 FUTURE DIRECTIONS

The recent developments such as NDS, smartphones and mobile apps, traffic sensors, connected vehicles, and autonomous vehicles have generated diverse sets of data that can be used for traffic safety studies. The generated data has recently inspired some innovative ideas for improving highway safety. These ideas are far from exhaustive given the rich information included in these data sets. Compared to the data that has been used in traditional traffic safety research and applications, these new data sets have the following important features:

- Widely deployed traffic sensors provide real-time traffic flow data covering large areas. For instance, in the Florida study by Shi and Abdel-Aty [9], spot speed, volume, occupancy and vehicle classification data for individual lanes were collected at a 1-minute interval using 275 microwave sensors on three highways (75 miles in total). The average distance between adjacent sensors is less than 1 mile. With this data set, temporal and spatial variations in traffic flow are well captured.
- Detailed driver behavioral data (e.g., accelerator pedal, brake pedal, speed, and acceleration collected at 10 Hz) in naturalistic driving settings has been collected in both NDS and SPMD. In particular, the SPMD installed a simple alert system in each vehicle and recorded how drivers responded to alerts.
- Near-crash events/extreme driving behaviors are well captured in the NDS and SPMD data sets. The NDS data also includes over 1000 crashes and video data prior to those crashes. These crash and near-crash events are critical for understanding what caused crashes and maneuvers that may prevent crashes from happening.
- Mobile apps and crowdsourcing are making large-scale and real-time driver behavior data collection possible. The NDS and SPMD were very expensive and were limited to a few selected study areas with several thousand drivers. Mobile apps can be connected to wearable electronic devices, vehicle OBD system, and connected infrastructure. This will generate very detailed data that enables a wide variety of safety applications.

Given the unique features of the new data sets and the limitations of existing safety studies, the following research areas may generate substantial interest and attract particular attention in the near future.

- *Next generation highway safety applications based on (i) NDS and SPMD data, (ii) traffic sensors, and (iii) data generated by mobile devices, vehicle OBD system, and crowdsourcing*: The NDS and SPMD ideas can be combined with roadside traffic sensor, in-vehicle GPS, and vehicle OBD data to comprehensively characterize macroscopic traffic flow status, microscopic vehicle trajectories, and vehicle health conditions. The new safety applications will be able to (i) help *safety analysts* dynamically identify safety hazards and characterize their relationship with traffic, roadway geometry, pavement condition, traffic sign and control, etc.; (ii) provide advance warnings to *traffic engineers* and assist them with generating proactive and real-time traffic control strategies to prevent safety hazards and congestion; (iii) identify safety hazards caused by poor roadway maintenance and help *maintenance staff* prioritize projects in terms of safety benefits; (iv) generate recommendations (e.g., vehicle design and vehicle—driver interface) to *car manufacturers* based on the analysis results of crash and near-crash events; (v) provide feedback to *drivers, insurance companies, and fleet managers* regarding dangerous driving behaviors/maneuvers in the form of videos and vehicle trajectories. This information can be used to train new drivers; and (vi) identify aggressive driving behaviors of other drivers and notify *police officers* automatically. This will help police officers develop targeted patrol schedules.
- *Vehicle—driver interface design*: Many vehicle safety and driver assistance technologies have been developed and implemented, including blind spot detection and warning, cross-traffic alert, and lane assistance. It is important to communicate the detected information to drivers without distracting them. Also, for connected and autonomous cars at different levels of automation, it is important to decide (i) when the vehicle should take over the control from a human driver during a hazardous situation; and (ii) how to alert drivers properly and turn the control over to a human driver when there is a sensor malfunction or other emergencies. There may be many other similar questions to be answered. The SPMD data can partially satisfy the needs for such human factors research. Additional studies may be needed to collect field data using vehicles equipped with vehicle safety and driver assistance technologies.
- *Cybersecurity*: with the rapid developments in CAVs, future traffic safety problems may become cybersecurity problems.
- *New methods*: although many statistical, machine learning, and data mining methods have been developed and introduced for analyzing traffic safety data, these methods are inadequate to model big and heterogeneous safety data generated by advanced sensors and various field studies (e.g., NDS and SPMD). Such data is sometimes highly unstructured, in many different forms (e.g., text and video), contains missing and erroneous information, and is being generated at various speeds. This brings many challenges to data storage, management, and analysis. These challenges will require traffic safety practitioners and researcher and data scientists to work together and clearly understand each other's needs.

9.6 CHAPTER SUMMARY AND CONCLUSIONS

This chapter briefly summarizes some main topics in existing highway traffic safety research. It also reviews methods and data that have been considered in these previous studies. Many of these traffic safety studies rely on static data obtained from police crash reports, DOT road network database (e.g., AADT), limited field tests, etc. Advanced statistical, machine learning, and data mining methods have also been developed and introduced for analyzing traffic safety data, including BNN, GLMs, support vector machines, and random forest models. These studies are useful in identifying safety countermeasures to improve vehicle design, roadway geometry, and traffic control. However, the data used is unable to capture near-crash events and precisely describe what exactly happened immediately prior to a collision. Also, it is difficult to analyze the joint effects of several crash contributing factors and find out how a collision could have been avoided in different ways given the static data. Although driving simulators can partially address this problem, there are still doubts about the fidelity of the results from driving simulators.

The developments in sensor technologies have made traffic sensors much affordable and helped to establish large-scale traffic sensor networks that can collect very detailed traffic data (at 20−30 second intervals). These advanced sensors also made possible some exciting safety applications such as the NDS and the connected vehicle SPMD project. The detailed macroscopic traffic data from roadside sensors and time series data describing driver behavior and vehicle operating conditions is very important for in-depth traffic safety analysis. These new safety data sets have attracted much attention recently. However, much work is still needed to effectively manage and to fully utilize such big and heterogeneous safety data that is often highly unstructured and in different forms, contains missing and erroneous information, and is being generated continuously at various speeds.

With the big and heterogeneous safety data, future highway traffic safety research and applications will feature dynamic, proactive, and intelligent strategies instead of being reactive and retrospective. They will rely heavily on new technologies and ideas such as connected vehicles, vehicle OBD system, radar and lidar sensors, video systems, mobile devices, and crowdsourcing. These new and existing technologies will enable safety analysts to closely examine what exactly happened prior to a crash or near-crash event. The analysis of near-crash events will provide additional valuable insights on how to prevent a crash from occurring. The dynamically changing nature of crash risk will be recognized and characterized. Instead of using average crash risk exposure such as AADT, high-resolution traffic data will be used in safety analyses.

9.7 EXERCISE PROBLEMS AND QUESTIONS

9.1 What are the main components of human factor studies for traffic safety? What are the advantages of NDS data compared to data generated by driving simulators for human factor studies? Conduct a review of naturalistic driving studies worldwide and discuss how you think the existing NDS plans can be further improved.

9.2 Briefly discuss the explanatory variables that may be needed when developing crash count models for intersections and road segments, respectively. Why should roadway links be divided into homogenous segments? How can the crash count modeling results be used? What are the methods often used in crash count modeling? What could be reasons that the predicted and observed crash counts are significantly different?

9.3 Discuss the methods often used in before and after studies and their pros and cons.

9.4 What is crash injury modeling and how can the results be used? What are the explanatory variables typically needed for crash injury severity modeling?

9.5 How can detailed sensor data and time series driver behavior and vehicle operating condition data (e.g., those from NDS) enable in-depth traffic safety research? Try to come up with some research topics other than those discussed in this chapter.

9.6 Why multiple linear regression models are not suitable for crash count data modeling? How to handle crash count data with excessive zeros?

9.7 Compared to statistical models, what are the benefits and limitations of using neural networks for crash count modeling? How to address the limitations of neural networks?

9.8 Discuss the methods that are often used in crash injury modeling and pay particular attention to their pros and cons.

9.9 Discuss the different types of data that are often used in traffic safety studies and where to obtain them?

9.10 In your opinion, what are the limitations of using averaged data such as AADT for crash count modeling?

9.11 What are the potential opportunities and challenges brought by the big and heterogeneous safety data from traffic sensor networks and sensors of future CAVs?

REFERENCES

[1] Mobile Technology Fact Sheet. Available from <http://www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet/>, (accessed 1.10. 16).

[2] K.L. Campbell, The SHRP 2 Naturalistic Driving Study. Available from <https://insight.shrp2nds.us/documents/shrp2_background.pdf>, (accessed 5.10.16).

[3] http://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec6.cfm, (accessed 1.10.16).

[4] C.S. Duncan, A.J. Khattak, F.M. Council, Applying the ordered Probit model to injury severity in truck-passenger car rear-end collisions,, Transp. Res. Rec. 1635 (1998) 63−71.

[5] National Automotive Sampling System (NASS) General Estimates System (GES) 2010 Coding and Editing Manual. U.S. Department of Transportation, National Highway Traffic Safety Administration, and National Automotive Sampling System. 2011.

[6] Federal Motor Carrier Safety Administration [FMCSA]. Summary of Hours of Service Regulations <https://cms.fmcsa.dot.gov/regulations/hours-service/summary-hours-service-regulations>, (accessed 1.10.16).

[7] http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=AB&infotype=PM&htmlfid=GVC03014USEN&attachment=GVC03014USEN.PDF, (accessed 1.10.16).

[8] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, Learning deep representation from big and heterogeneous data for traffic accident inference, In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), February 12−17, 2016, Phoenix, Arizona, USA, 2016, pp. 338−344.

[9] Q. Shi, M. Abdel-Aty, Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways,, Transp. Res. Part C 58 (2015) 380−394.

[10] USDOT. Connected Vehicle Safety Pilot Program. <http://www.its.dot.gov/factsheets/pdf/JPO_SafetyPilot.pdf>, (accessed 1.10.16).

[11] J. Liu, A. Khattak, Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles, Transp. Res. Part C 68 (2016) 83−100.

[12] D. Lord, A. Manar, A. Vizioli,, Modeling crash-flow-density and crash-flow-V/C ratio for rural and urban freeway segments,, Accid. Anal. Prev. 37 (1) (2005) 185−199.

[13] D. Lord, S.P. Washington, J.N. Ivan, Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory, Accid. Anal. Prev. 37 (1) (2005) 35–46.

[14] J.J. Song, M. Gosh, S.-P. Miaou, B. Malik, Bayesian multivariate spatial models for roadway traffic crash mapping,, J. Multivar. Anal. 97 (1)) (2006) 246–273.

[15] E.S. Park and D. Lord, Multivariate Poisson-Lognormal models for jointly modelling crash frequency and severity. Paper presented at the 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2006.

[16] S.-P. Miaou, D. Lord,, Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes,, Transport. Res. Rec. 1840 (2003) 31–40.

[17] F.L. Mannering, C.R. Bhat, Analytic methods in accident research: Methodological frontier and future directions, Analytic Method. Accid. Res. 1 (2014) 1–22.

[18] P.T. Savolainen, F.L. Mannering, D. Lord, M.A. Quddus, The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives, Accid. Anal. Prev. 43 (3) (2011) 1666–1676.

[19] M. Abdel-Aty, Analysis of driver injury severity levels at multiple locations using ordered Probit models, J. Safety. Res. 34 (5) (2003) 597–603.

[20] K. Train, Discrete Choice Methods with Simulation, Cambridge University Press, New York, 2003.

[21] S. Cardell, F. Dunbar, Measuring the societal impacts of automobile downsizing, Transport. Res. Part A 14 (5-6) (1980) 423–434.

[22] J. Swait, A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data, J. Retail Consumer Serv. 1 (2) (1994) 77–89.

[23] C. Marzban, A. Witt, A Bayesian neural network for severe-hail size prediction, Weather Forecasting 16 (5) (2001) 600–610.

[24] K.E. Fish, J.G. Blodgett, A visual method for determining variable importance in an artificial neural network model: An empirical benchmark study, J. Targeting Measurement Anal. Market. 11 (3) (2003) 244–254.

[25] D. Delen, R. Sharda, M. Bessonov, Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks, Accid. Anal. Prevention 38 (3) (2006) 434–444.

[26] K. Hornik, M. Stinchcombe, H. White, Multilayer feed forward networks are universal approximators, Neural Netw 2 (5) (1989) 359–366.

[27] D.J.C. Mackay, Bayesian methods for adaptive models. Ph.D. Dissertation, California Institute of Technology, Pasadena, California, 1992.

[28] R.M. Neal, Bayesian learning for neural networks. Ph.D. Dissertation, University of Toronto, Toronto, Ontario, 1995.

[29] F. Liang,, Bayesian neural networks for nonlinear time series forecasting, Statistics Comput. 15 (1) (2005) 13–29.

[30] Y. Xie, D. Lord, Y. Zhang, Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis, Accid. Anal. Prev. 39 (5) (2007) 922–933.

[31] S. Handy, Critical assessment of the literature on the relationship among transportation, land use, and physical activity, Resource paper for TRB Special Report (2005) 282.

[32] J.N. Ivan, P.J. Ossenbruggen, X. Qin, J. Pendarkar, Rural Pedestrian Crash Rate: Alternative Measures of Exposure. Report No. UCNR 11-10, New England UTC, 2000.

[33] K. Shriver, Influence of environmental design on pedestrian travel behavior in four Austin neighborhoods, Transport. Res. Rec. 1578 (1997) 64–75.

[34] R.J. Schneider, L.S. Arnold, D.R. Ragland,, Pilot model for estimating pedestrian intersection crossing volumes, Transport. Res. Rec. 2140 (2009) 13–26.

[35] National bicycle and pedestrian documentation project (NBPD). <http://bikepeddocumentation.org/>, (accessed 1.10.16).

[36] Research Department, Downtown Denver Partnership, Inc. Downtown Denver Summer 2013 Pedestrian Count Report. <http://www.downtowndenver.com/wp-content/uploads/2013/10/DowntownDenverPedestrianCountReport2013.pdf>, (accessed 1.10.16).

[37] A. Mednis, G. Strazdins, R. Zviedris, G. Kanonirs, and L. Selavo, Real time pothole detection using android smartphones with accelerometers, in: 2011 International EEE Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS), pp. 1–6, 2011.

[38] SHRP2 Solutions Tools for the Road Ahead, Creating a safer transportation system: How the new SHRP2 safety databases can take us there, (accessed 1.05.16).