NCHRP 22-47

Incorporating Driver Behavior and Characteristics into Safety Prediction Models

Task 3: Identify Modeling Approach

Prepared for:

National Cooperative Highway Research Program

Transportation Research Board, the National Academies

Submitted by:

The University of North Carolina at Chapel Hill
Highway Safety Research Center (HSRC)

Vanasse Hangen Brustlin (VHB)

University of Wisconsin, Milwaukee

November 1, 2021

Task 3: Identify Modeling Approach

The objective of this task is to identify candidate modeling approaches that incorporate a wide variety of factors related to driver behavior and characteristics into crash prediction methods. Specifically, we want to approach this from four different aspects:

1. Modify the safety performance function (SPF) and its "base conditions" that reflect the overrepresentation of certain demographics of drivers (e.g., old drivers, young drivers).
2. Identify appropriate crash modification factors or functions (CMF) that account for the driver behavior differing between site conditions and base conditions. The adjustment will include, but are not limited to, driver behavior, awareness, performance, and actions; enforcement statistics of driver violations, errors, or lapses; enforcement strategies and traffic laws.
3. Develop predictive methods that specifically incorporate variables related to driver behavior and/or characteristics in a modeling structure.
4. Develop predictive methods that minimize the biases and impacts on crash prediction due to the omission of important driver behavior data.

The results will inform the development of the NCHRP 22-47 Phase 2 work plans and the procedural guidance. The results could also inform future research needs for considering specific driver characteristics and behavior measurements where data are available.

We evaluated six types of modeling approaches based on a) data availability; b) crash causal inference capability; c) statistical goodness of fit and prediction accuracy; d) interpretability and practicality; e) model parsimony and transferability. Following the rating criteria established in this task, we recommend methods of Multiple Risk Sources Model and Simultaneous Equations Model as primary approaches to explicitly incorporate driver behavior and/or characteristics and mitigate the impact by the omission of important variables.

The following sections present the details of modeling approaches, the rationale of rating, and the strengths and characteristics of the top candidate methods.

Opportunities to Incorporate Driver Characteristics/Behavior in Crash Modeling

We identified opportunities in each step of the two-step process for considering driver behavior and characteristics in a crash prediction model: Step one is to identify or create appropriate variables that best represent driver behavior and characteristics from available data sources; and Step two is to reduce the impact of data heterogeneity due to the omission of important driver behavior information through advanced data analytics.

The number of predicted crashes, $N_{predicted}$, can be generally expressed as the product of the function for traffic exposure and the function for crash risk, as formulated in Eq. 1

$$N_{predicted} = Exposure(function) \times Risk(function) \quad (1)$$

The traffic exposure function can measure not only the size but also the characteristics of road users such as age, race, gender, marital status, income, education, and employment. Considering explicit driver demographics in the exposure function helps to explain why large disparity of crash frequency may exist between sites with the same number of drivers, the same level of

AADT or VMT. Owing to the US census data, demographic variables are relatively easy to collect. Other exposure variables such as pedestrian or bicyclist count are more difficult. Alternatively, proxy variables may be considered. In the NCHRP 17-73 Systemic Pedestrian Safety Analysis, the density of light pole was found to be positively associated with midblock pedestrian crashes. It is possible that more pedestrian activities are present at locations with good lighting. Hence, the density of light pole could be treated as the proxy for pedestrian exposure rather than a predictive factor in the risk function.

In a risk function, certain traffic and geometric design variables can be used to characterize user behavior and performance based on human factor-related concepts. Macroscopic traffic flow variables (e.g., flow rate, average vehicle speed, speed variance, average density) are useful for predicting crashes or crash tendency as they measure driver interactions. Horizontal and vertical alignment variables such as degree of curve, curve length, radium, curvature change rate, gradient/grade are strong indicators of the level of design consistency along a roadway segment. It is known that geometric design inconsistencies can violate driver expectancy and increase driver workload, which in turn impact drivers' performance and ability to properly perceive and respond to an impending risk.

In the HSM predictive method, the exposure function can be viewed as the SPF, N_{spf} ; and the risk function can be viewed as the multiplication of CMFs; as specified in Eq. 2.

$$N_{predicted} = N_{spf} \times (CMF_1 \times \cdots \times CMF_n) \times C_{r/i} \quad (2)$$

SPFs are typically a function of only a few variables (e.g., AADT volumes) which can be modified to incorporate driver characteristics at a specific spatial scale (e.g., census tracts, TAZs, zip codes, counties). Existing traffic or geometric variables in CMFs can be improved to better capture driver expectation and performance. Moreover, new variables representing enforcement statistics and/or historical crashes related to risky driver behavior (e.g., traffic violations, driver errors or lapses) can be added as new CMFs. It is expected that the SPFs and CMFs will be represented by different sets of variables; some variables are more "causal" than "predictive".

Process of Incorporating Driver Characteristics and Behavior in Crash Modeling

SPFs and CMFs can be developed separately or together. If modeled separately, SPFs and CMFs are estimated and calibrated with different datasets; and their model forms are different. The SPF for base conditions can be modified through a (driver) population factor or through a function. An example in the Highway Capacity Manual (HCM, Chapter 26, Section 4) presents recommended values for Capacity adjustment factor (CAF) and speed adjustment factor (SAF) that describe the level of driver familiarity with traffic conditions: for all familiar drivers, regular commuters, both CAF and SAF are assigned to be 1; for mostly familiar drivers, CAF and SAF are 0.968 and 0.975, respectively; for balanced mix of familiar and unfamiliar drivers, CAF and SAF are 0.939 and 0.995, respectively; and for mostly unfamiliar drivers, CAF and SAF are down to 0.898 and 0.913, respectively. Alternatively, we can apply the quasi-induced exposure

method to estimate driver proportions by age, gender, or other demographic variables for the base SPF.[1]

Using CMFs to measure the safety impact of driver behavior can be designed from information in various sources. The current HSM 1st Edition already include some driver behavior related CMFs such as Number of Alcohol Sales Establishments near the Intersection (Table 12-30), Road-use Culture Network Consideration and Treatments (Table 17-4), Install automated speed enforcement (Table 17-5), Install changeable speed warning signs (Table 17-6). These CMFs need to be calibrated or converted to represent specific driver populations when new data are available. The CMF Clearinghouse provides a searchable database of CMFs for engineering treatments towards all drivers. To derive driver behavior related CMFs, the safety effect of engineering treatments for specific driver populations needs to be investigated and the relationships need to be calibrated. Another possible source for CMFs is a document called "Countermeasures that work" that is produced periodically by NHTSA. Moreover, CMFs can be developed for risky driving behaviors such as traffic violations (e.g., speeding, red-light running, reckless driving); driver lapse (e.g., distraction, inattentiveness, fatigue, drowsy); and driver errors (e.g., competency, experience, expectation, physical and mental limitations).

Along with CMFs for roadway conditions, we will now have CMFs for driver characteristics and behavior. So, there are potential issues due to the assumption of independence among CMFs. If that is the case, FHWA recommended procedures will be taken to address lack of independence[2]. When data permit, we may extend our investigation to complicated factors that can have either positive or negative safety impact. For example, driver's familiarity with roadway can be treated as a risk factor as familiar drivers tend to be risk prone due to distraction, inattention and over-confidence. Unfamiliarity can also be a risk factor due to the ignorance of the circumstances and negative interactions with other drivers[3]. A list of CMFs highly related driver behavior will be presented and ranked in Phase 2 based on the reliability and applicability.

Modeling SPF and CMF separately has the flexibility of associating each with a set of variables, and making use of existing studies (i.e., HSM and CMF Clearinghouse). This approach, however, is subject to available CMFs and the assumption of independence of CMFs. On the other hand, a more commonly used practice is modeling SPF and CMF together through a full model specification. Incorporating all variables in a single modeling framework presents opportunities for applying sophisticated analytical methods to mitigate data heterogeneity problem. Additionally, the methods for full model development should be able to incorporate different variables distinctively (i.e., exposure factors vs. crash risk factors, directly observed factors vs. derived or proxy factors). Some of the methodological advances include:

[1] Sharmin, S., Ivan, J. N., Zhao, S., Wang, K., Hossain, M. J., Ravishanker, N., & Jackson, E. (2020). Incorporating demographic proportions into crash count models by quasi-induced exposure method. Transportation research record, 2674(9), 548-560.

[2] Gross, F., & Hamidi, A. (2011). Investigation of existing and alternative methods for combining multiple CMFs. Highway Safety Improvement Program Technical Support. Task A, 9.

[3] Intini, P., Berloco, N., Colonna, P., Ranieri, V., & Ryeng, E. (2018). Exploring the relationships between drivers' familiarity and two-lane rural road accidents. A multi-level study. Accident Analysis & Prevention, 111, 280-296.

- *Random parameters models*: This type of model can potentially capture unobserved heterogeneity by allowing coefficients to vary across observations (such as a roadway segments or intersections) as opposed to a fixed value of a coefficient (meaning invariant influence). Notable random parameters count models include random parameters Poisson (RP-P), random parameters negative binomial (RP-NB), random parameters Poisson-lognormal (RP-PLN).
- *Finite-mixture/latent-class models*: This type of model addresses unobserved heterogeneity by identifying distinct subgroups with homogeneous attribute values by the data as opposed to arbitrarily grouping sites by some observed characteristics. Notable models include finite mixtures of Poisson (FM-P) or NB (FM-NB) regression models, and Markov switching negative binomial (MSNB) model.
- *Mixed models*: This type of model is used to incorporate heterogeneity into statistical analysis by adding a mixed distribution to account for extra variance in the crash data. Notable mixed NB models include the NB-Lindley (NB-L), NB-Generalized Exponential (NB-GE), and NB-Dirichlet process (NB-DP) generalized linear models (GLMs).
- *Multiple risk sources regression models*: This type of model can explicitly consider the impact of driver behavior and characteristics on crash occurrence. The rationale is that the crash risk factors stemming from a distinct risk source (e.g., highway and traffic, driver behavior and characteristics, vehicle performance, environment) affect crash occurrence in a similar manner, but vary between different sources. Notable models include multiple risk sources negative binomial model and multivariate multiple risk sources negative binomial model.

Two other approaches have unique structures to include driver behavior in the modeling process:

1) *Structural Equation Model*: This type of model tests and estimates the complicated relationships between variables (both endogenous and exogenous variables) through a combination of statistical methods and qualitative causal assumptions. Latent variables such as driver violation index, driver error index or driver lapse index can be estimated using path analysis through observable variables.
2) *Simultaneous Equations Model*: This type of model is employed to systematically model a number of dependent variables that are mutually interrelated such as traffic exposure and crash frequency. It addresses the simultaneity due to the possibility that unobserved factors affecting crash frequency may be correlated significantly with unobserved factors affecting traffic exposure.

Prioritization of Modeling Approaches

The methodological innovation is driven by sound theoretical reasoning of the crash generating process, statistical robustness and goodness-of-fit, and model parsimony. Although the simple form of a crash is appealing to practitioners, "*the real problem with parsimonious models is that practitioners, and even researchers, do not fully grasp, or often conveniently overlook, the limitations of these simplistic models*".[4] Therefore, a balance needs to be carefully evaluated

[4] Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research, 1, 1-22.

between the limitations of parsimonious models and the benefits of more sophisticated statistical approaches. Model performance depends on how the following questions are answered:

a) What are the modeling strategies that balance the data needs and prediction accuracy?

b) Can the mathematical relationships among variables be effectively modeled around crash causations?

c) Will the results generated from the model be easy to understand by practitioners?

d) How well can the model fit the data; and can a model reduce the biases or impact of data heterogeneity due to the omission of important variables or use of proxy variables?

e) What level of complexity is the model; and can the model be easily transferred for crash prediction in the future or to somewhere else?

Evaluation criteria are created in five areas to assess statistical models along with the safety data for their overall strengths in the crash causal theory, statistical rigor, and safety applications: a) data availability; b) crash causal inference capability; c) statistical goodness of fit and prediction accuracy; d) interpretability and practicality; e) model parsimony and transferability.

1) Data availability: the robustness of the model to handle data heterogeneity due to fewer or missing variables. (1 – poor; 2 – fair; 3 – good)
2) Crash causal inference capability: the capability of the model to identify causal effects. (1 – poor; 2 – fair; 3 – good)
3) Interpretation and practicality: the ease of the interpretation and implementation of the model (1 – poor; 2 – fair; 3 – good)
4) Statistical goodness of fit and prediction accuracy: the extent to which observed data match the values expected by the model; and strong correlation between the model prediction and the new observed data. (1 – poor; 2 – fair; 3 – good)
5) Model parsimony and transferability: the ease of use and the adaptability of the model to new data (1 – poor; 2 – fair; 3 – good)

The project team ranked the methods based on their total points by summing the points for each of the five criteria. Methods with more points are prioritized higher for than those with fewer points. The following describes each modeling approach with the five rating criteria.

Random Parameter (RP) Model: RP modeling addresses data heterogeneity by allowing some or all the model parameters to vary from observation to observation in a probabilistic distribution. This type of models can account for heterogeneity across observations that are mainly due to unobserved variables, especially driver behavior and characteristics related variables. Crash data studies that applied the RP models (i.e., RP-P, RP-NB, and RP-PLN) often report a significant improvement in the statistical model fit. However, this type of method may be susceptible to observations generated from different data sources. For example, although RP-NB outperforms NB when applied to within sample observations by making use of the observation-specific coefficients; but the predictive power of the RP-NB model is less accurate than NB when applied to out-of-sample observations (i.e., new data). This is particularly relevant when applying the predictive methods for new highway construction projects or on roadway sites

that were not used to estimate the models. In addition, the RP approach might not be practically useful for the development of national level SPFs which are estimated from a sample of roadway miles, if the models are intended to be used in somewhere else. Moreover, as the model assumes the parameters to follow a random distribution, model results may be difficult to interpret.

Therefore, for the RP model, we give a rating of 3 for data availability for its strength in handling data heterogeneity; 1 for its lack of crash causal inference; 1 for its difficulty in result interpretation; 2 for its superior performance in statistical goodness of fit but the out-of-sample issue; 1 for model parsimony and transferability because advanced statistical knowledge is required to develop and calibrate the model; and the total score is 8 out of 15.

Finite Mixture (FM) Model: FM models express the overall distribution of crash count data as a mixture of a finite number of component probability distributions which prescribe the observations from different subpopulations. A finite mixture model can handle various distributions for different sub-groups in the target dataset. Especially when crash data were collected from distinct sources or suspected to be heterogenous. Moreover, the FM-NB models can be enhanced by varying weight parameters to analyze the dispersed crash data. For example, the weight parameter of the FM models is assumed to be dependent upon the attributes of the sites (i.e., covariates) such as segment length, traffic flow. Opposite to random parameters models, finite mixture models consider unobserved heterogeneity by using a finite and specified number of mass points to identify homogeneous subgroups of data. The benefit by addressing unobserved heterogeneity in such way is unobserved states may exist due to different factors such as driver behavior and characteristics that is not available. The disadvantage is that it does not account for the possibility of within-group variation due its restrictive homogeneity assumption on characteristics of the within-group observations.

Therefore, we give a rating of 3 for data availability for its strength in handling data heterogeneity; 2 for crash causal inference and 2 for interpretation and practicality because of its capability of grouping observations into homogenous subpopulations to help understand the underlying crash generating mechanisms; 1 for statistical goodness of fit and for prediction accuracy because of mixed results; 1 for model parsimony and transferability because advanced statistical knowledge is required to develop and calibrate the model; and the total score is 9 out of 15.

Mixed Model: The mixed model is a well-known methodology used to incorporate heterogeneity into statistical analysis. The advantage of using a mixed model is that it adds a mixed distribution to account for extra variance in the crash data which is caused by preponderant zero crash responses and/or a heavy tail of crash counts. However, the mixed model does not resolve the issue of omitted variables that could affect the likelihood of crashes, which are basically the driver behavior and characteristics related variables. Though by incorporating both random parameters and mixed probabilistic distributions within a single model can be a viable alternative for handling crash data with unobserved heterogeneity, it will inevitably have the same limitations as the random parameter model (i.e., issues with out-of-sample data and crash causal inference capability).

Therefore, we give a rating of 2 for data availability for its ability of handling data heterogeneity; 1 for crash causal inference capability and 2 for interpretation and practicality because the model is more appropriate in fitting fat-tailed or long-tailed crash data; 2 for statistical goodness of fit prediction accuracy because the accuracy gain is often reported as marginal; 1 for model parsimony and transferability because advanced statistical knowledge is required; and the total score is 8 out of 15.

Multiple Risk Sources Regression Model: Observed crashes are not generated by a single crash occurrence process but instead arise from separate processes including observed network influences, unobserved spatial influences, and behavioral influences that appear random, which can be illustrated in Figure 1.

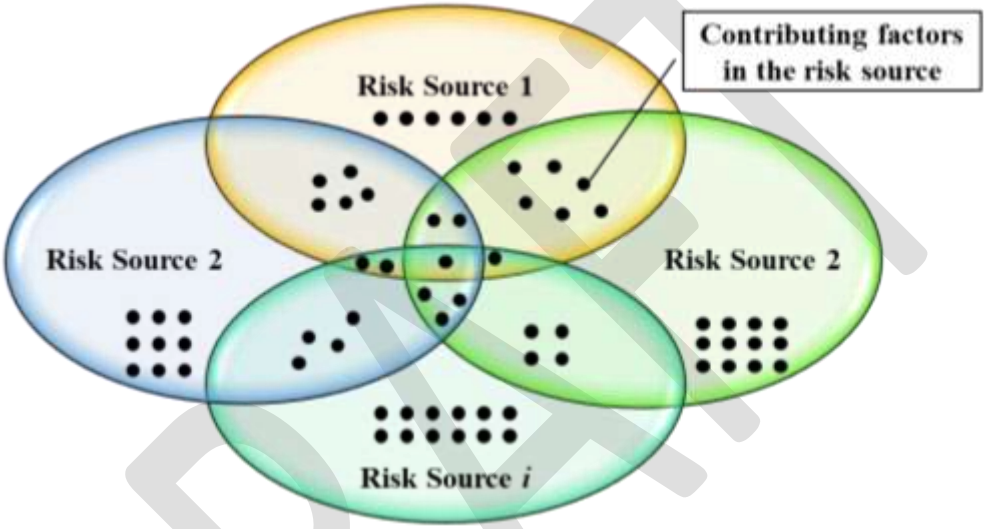

Figure 1 Example of Multiple Risk Source Model of Crashes

Ignoring the multiple sources in current modeling methodologies leads to model misspecification that associates crashes with incorrect sources of contributing factors (e.g., concluding a crash is predominately caused by a geometric feature when the cause is a behavioral issue). Modeling crashes that occur at locations as negative binomial (NB) distributed events that arise from a single crash generating process fails to capture the underlying complexity of motor vehicle crash causes, and thus hinders deeper understanding of crash causation, undermines site safety screening, and compromises the selection of appropriate treatments. A multiple risk source regression model is theoretical appealing and technically flexible for estimating crashes based on their originating risk sources and providing meaningful parameter estimates. With this novel method, the total crash count can be decomposed and modeled into multiple constituent components, representing complex mechanism of crash data generating processes such as engineering, unobserved spatial, and driver behavioral factors. Behavioral variables collected at larger geographic scale representing existing social norms that can influence driving behavior within a community. In association with engineering risk factors, these behavioral variables are considered to compose crash risk which is originated from a distinct source.

Therefore, we give a rating of 1 for data availability for its weakness to handle data heterogeneity; 3 for crash causal inference and 3 for interpretation and practicality because its model structure is designed for explicitly considering driver behavior as a risk source; 2 for statistical goodness of fit and prediction accuracy because of satisfactory results reported in recent studies; 2 for model parsimony and transferability because although advanced statistical knowledge is required, the model can be transferred to different sites; and the total score is 11 out of 15.

Structural Equation Model: The major strength of the structural equation model is its capability to verify the significance of hypothesized relationships between latent variable (unobserved) by means of observed variables. A structural equation model can effectively distinguish direct, indirect, and total effects among variables by means of three components as shown in Figure 2: 1) a measurement model for the observed endogenous variables (*y*); 2) a measurement model for the observed exogenous variable (*x*); and 3) a structural model[5]. Thus, the underlying relationships between factors can be captured more accurately. It would be very helpful to create new latent variables (e.g., driver behavior indexes) from observable variables (e.g., citations, liquor sales). However, this type of model may not be able to achieve an acceptable level of goodness of fit for multidimensional measures.

Therefore, we give a rating of 1 for data availability; 3 for crash causal inference capability; 3 for interpretation and practicality; 1 for statistical goodness of fit and prediction accuracy; 2 for Model parsimony and transferability; and the total score is 10 out of 15.

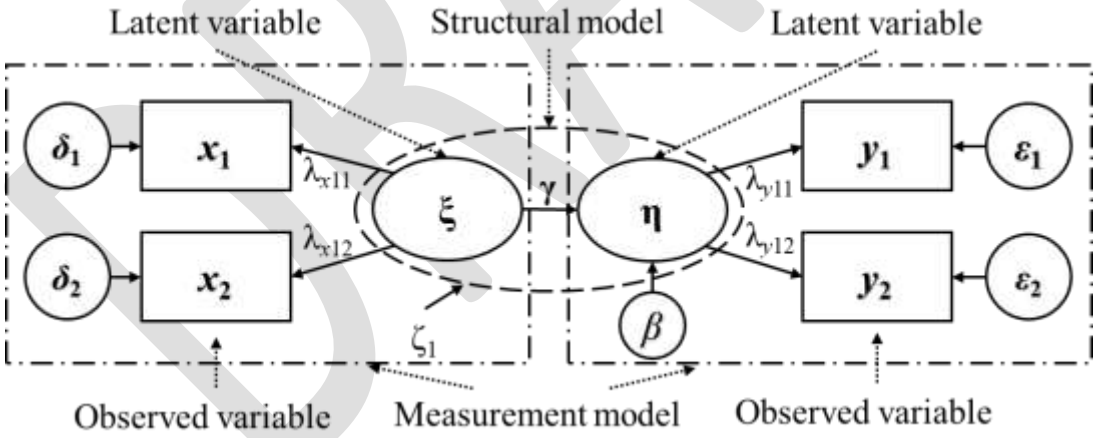

Figure 2 Example of A Structural Equation Model

Simultaneous Equations Model: One issue needs to be considered when developing the crash prediction model is the interdependency between exposure (e.g., AADT and VMT), driver behavior/characteristics and demographics, and crashes outcomes (e.g., frequency, type, and severity), as shown in Figure 3.

[5] Lee, J. Y., Chung, J. H., & Son, B. (2008). Analysis of traffic accident size for Korean highway using structural equation models. Accident Analysis & Prevention, 40(6), 1955-1963.

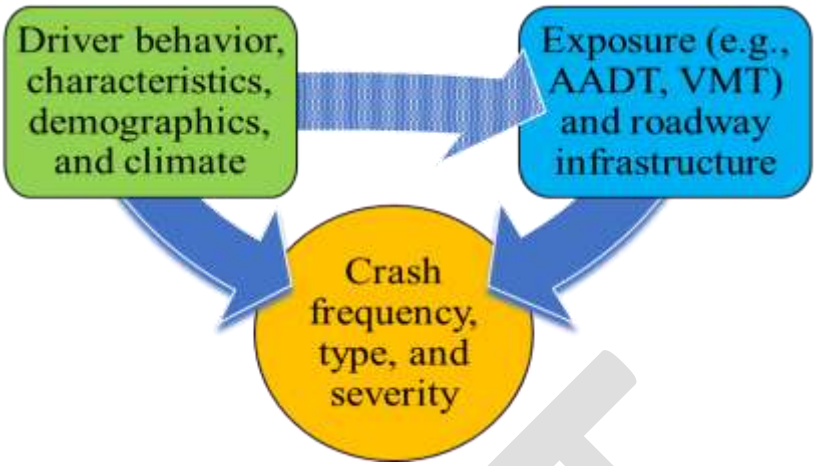

Figure 3 Relationship between Driver Behavior/Characteristics, Climate, Roadway Infrastructure, and Safety

One way to account for these interrelationships is to express these relationships in the form of multiple equations. For instance, the dependent variable will be AADT or VMT in one equation, and the independent variables will include data on the surrogates of driver behavior/characteristics, and climate. In the second equation, the dependent variable will be crash frequency, type, and severity, and the independent variables can include exposure, roadway infrastructure, and surrogates of driver behavior/characteristics. These multiple equations can be estimated in one modeling framework called simultaneous equation modeling.

Therefore, we give a rating of 2 for data availability because the joint equations can handle the confounding effect by the unobserved data; 2 for crash causal inference capability; 3 for interpretation and practicality; 1 for statistical goodness of fit and prediction accuracy because of the restrictive assumption of normal distribution; 2 for model parsimony and transferability; and the total score is 10 out of 15.

Table 1 shows the points assigned to each model type by criterion. In summary, the multiple risk sources model ranks 1st with the most total points (11 pts), followed by the simultaneous equations model (10 pts) and structural equation model (10 pts). Since the structural equation model is rarely used to make predictions, the multiple risk sources model and simultaneous equations model will be selected as the candidates to develop a full model to be used in Task 7: Develop Predictive Methodology, together with the CMF Clearinghouse/HSM method. A traditional NB model is also recommended to be developed as the baseline model.

Table 1 Comparison between Methods

Methods / Rating Criteria	CMF Clearing House / HSM	Random Parameter Model	latent-class model	Mixed Model	Multiple Risk Sources Model	Structural Equation Model	Simultaneous Equations Model
Data availability	1	3	3	2	1	1	2
Crash causal inference capability	3	1	2	1	3	3	2
Interpretation and practicality	3	1	2	2	3	3	3
Statistical goodness of fit and prediction accuracy	n.a.	2	1	2	2	1	1
Model parsimony and transferability	3	1	1	1	2	2	2
Total	X	8	9	8	11	10	10

Conclusions

Based on an assessment of potential and ongoing efforts, NCHRP 22-47 could benefit from testing the following modeling approaches:

1. Driver characteristics and behavior can be modeled differently in a predictive method, depending on their association with crashes, as exposure variables or risk predictors.
2. When considering driver characteristics in a SPF setting, the current base conditions for different highway facilities need to be modified to reflect the changes.
3. Crash modification factors or functions can be represented in the format of conditions deviated from the base; and/or effectiveness of safety treatments, following procedures in the HSM D.4.4 Application of CMFs to Estimate Crash Frequency (D-6).
4. Based on results of the rating criteria, methods of Multiple Risk Sources Model and Simultaneous Equations Model can be considered as the primary modeling approach. Possessing all the benefits of a traditional NB model, the Multiple Risk Sources Method can explicitly incorporate driver behavior or characteristics in its modeling process.