Mannering, F., Bhat, C., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research* 25, 100113.

Big Data, Traditional Data and the Tradeoffs between Prediction and Causality in Highway-Safety Analysis

Fred Mannering
Associate Dean for Research, College of Engineering
Professor of Civil and Environmental Engineering
University of South Florida
4202 E. Fowler Avenue, ENC 3506
Tampa, FL 33620
Email: flm@usf.edu

Chandra R. Bhat
Joe J. King Chair in Engineering and Distinguished Teaching Professor
Professor of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
301 E. Dean Keeton St. Stop C1761
Austin, Texas 78712
Email: bhat@mail.utexas.edu

Venky Shankar
Professor and Chair
Department of Civil, Environmental, and Construction Engineering
Texas Tech University
Lubbock, TX 79409
Email: venky.shankar@ttu.edu

Mohamed Abdel-Aty
Trustee Chair, Pegasus Professor and Chair
Civil, Environmental, and Construction Engineering
University of Central Florida
12800 Pegasus Dr. #211
Orlando, FL 33620
Email: M.Aty@ucf.edu

October 2019 (revised January 2020)

Abstract

The analysis of highway accident data is largely dominated by traditional statistical methods (standard regression-based approaches), advanced statistical methods (such as models that account for unobserved heterogeneity), and data-driven methods (artificial intelligence, neural networks, machine learning, and so on). These methods have been applied mostly using data from observed crashes, but this can create a problem in uncovering causality since individuals that are inherently riskier than the population as a whole may be over-represented in the data. In addition, when and where individuals choose to drive could affect data analyses that use real-time data since the population of observed drivers could change over time. This issue, the nature of the data, and the implementation target of the analysis imply that analysts must often tradeoff the predictive capability of the resulting analysis and its ability to uncover the underlying causal nature of crash-contributing factors. The selection of the data-analysis method is often made without full consideration of this tradeoff, even though there are potentially important implications for the development of safety countermeasures and policies. This paper provides a discussion of the issues involved in this tradeoff with regard to specific methodological alternatives and presents researchers with a better understanding of the trade-offs often being inherently made in their analysis.

Keywords:

Self-selectivity, endogeneity, highway safety, accident likelihood, accident severity

1. Introduction

The implicit assumption in traditional statistical analyses is that an appropriately estimated model will both uncover causal effects and have the highest possible prediction accuracy. But the recent development and application of data-driven methods, as well as issues of causality in traditional statistical modeling, suggest that safety analysts must often, even if not always, make a trade-off between prediction accuracy and uncovering underlying causality. That is, models that predict well may not be the best at uncovering causality, and models that are good at uncovering causality may not be the best for practical prediction purposes.

There are four general methodological approaches that are potentially suitable for the analysis of transportation safety data: traditional statistical models, endogeneity/heterogeneity models, data driven methods, and causal inference models.[1] Each of these models have an implicit trade-off between practical prediction accuracy and their ability to uncover underlying causality. Traditional statistical models, such as those in the Highway Safety Manual (AASHTO, 2010), use conventional statistical methods with limited data (data that is readily available to most safety practitioners) to predict the effect of various safety improvements on accident risk. The traditional literature (such as that supporting the Highway Safety Manual) claims predictive capabilities and causal explanations, but generally lacks fundamental support for these claims via assessments of parameter bias (for example, potential biases in parameter estimates and estimates of standard errors). Predictive capabilities of traditional highway-safety models are typically based on assessment of aggregate counts (total count of accidents for example), and there is scant support for true tests of predictability (such as tracking observational predictions against observed counts

[1] Causal inference models have become a key analytic approach in the economics field, and have been gaining in interest among transportation researchers. However, the complexities of applying the approach in the complex behavioral arena of transportation-related decision making are an ongoing concern (Brathwaite and Walker, 2018).

several years ahead of the estimated models). In fact, claims of predictive ability in many traditional models are limited in credibility, in large part due to temporal instability in parameters (Mannering 2018). Similarly, claims about causal ability in the traditional safety literature are limited because the true range of influential factors on accident likelihoods is unknown. Missing data problems, problems of consistency of measurement, and variation in unobserved effects due to economic, socio-demographic and vehicle characteristics amplify the potential bias in estimation.

To address some of the limitations above, endogeneity models (see Bhat et al., 2014) and heterogeneity models (see Mannering et al., 2016 for a thorough review) have been developed to extend traditional safety models by using advanced statistical and econometric methods. Endogeneity models account for the potential endogeneity of a safety-related variable when attempting to extract the “true” causal effect of the variable on a primary safety outcome variable of interest, after accommodating “spurious” associative effects or correlation effects between the variables. Unobserved heterogeneity models control for unobserved factors that may influence the likelihood and resulting injury severities in accidents. Endogeneity models and heterogeneity models are stylized, in that they are based on relatively limited datasets where the range of the potential endogenous and explanatory variables is much larger than widely available transportation highway data. A richer set of variables can potentially improve predictive capability and understanding of causality; however, the increased model complexity creates an additional burden on model transferability and predictive validation. Model complexity also poses challenges in estimation due to computational constraints. Estimation of highly complex endogeneity models and heterogeneity models involves simulation-based methods or analytic approximation methods due to the numerical integration needed to capture unobserved effects. While there has been

substantial progress in such methods in the recent past (see, for example, Bhat, 2018), the required estimation techniques can still present dimensionality challenges for large accident datasets.

Data-driven methods include a wide range of techniques including those relating to data mining, artificial intelligence, machine learning, neural networks, support vector machines, and others. Such methods have the potential to handle extremely large amounts of data and provide a high level of prediction accuracy. On the downside, such methods may not necessarily provide insights into underlying causality (truly understanding the causal effects of specific factors on accident likelihoods and their resulting injury probabilities).[2]

Finally, causal-inference models explicitly recognize that accidents are only observed for a portion of the driving population and that this can lead to erroneous interpretations of findings (more on this below). Causal-inference models have rarely been applied in the accident analysis literature, but such approaches in other fields base these models on time series data to identify causal effects. However, causal-inference models have weak predictive capabilities because, among other reasons, they typically are not based on individual-accident level data and thus address a limited number of explanatory variables. Besides, the time-series nature of these models, while supposedly providing more basis for inferring causality, raises additional issues about the possible presence of uncontrolled factors that change during the intervening periods of time thereby potentially tainting the presence and estimated extent of causation.

Figure 1 presents a graphic of the trade-offs associated with these methods regarding predictive capability, causal inference capability and big data suitability (the ability of the methods to address problems that involve large amounts of data.) The choice of one method over another often involves several important considerations that go beyond a simple tradeoff between

[2] Some insight into the influence of specific variables in data-driven methods can made through simulation and calculating factors such as Gini Index, but this may not necessarily provide insight into underlying causality.

prediction and causality. Each of these four methods (data-driven versus causal versus traditional versus endogeneity/heterogeneity models) involve different levels of data. In addition, the application of the model (modeling purpose) needs to be considered as well. For example, endogeneity models and heterogeneity models would seem to be superior to traditional models in both prediction and causality; however these models typically use highly detailed datasets, and the models are complex in their application. In contrast, traditional safety models have relatively modest data requirements that are easy to apply, but their utility comes at the expense of a loss of predictive capability and lack of insight into causal influences (with the added risk of biased inference).

With extremely large datasets (big data) such as those that might be available in naturalistic driving studies, traditional models, advanced endogeneity/heterogeneity models, and causal effect models can be challenging to estimate, often making data-driven methods the preferred approach. In fact, data-driven methods can cover a wide range of data sizes, but, with smaller data sizes, the advantages of other methods to uncover causality tend to be preferred among analysts. Also, data driven methods may not be adequately complemented with domain expertise, resulting in inference driven primarily by statistical reasoning. The advent of artificial intelligence (AI) methods and the explosive growth of AI potentially opens the door for introducing some level of “automated” domain expertise to fine-tune data driven models that are developed strictly by statistical reasoning. But, at the end, human judgement and domain expertise are still likely to be needed in some form, especially in the context of driving the formulation of models of causal inference, since directional relationships between variables are formulated based on apriori knowledge of influential factors. As an example, in big data problems studying the impact of factors affecting fatality likelihoods, sample size is a significant issue. Fatalities on average occur at the rate of

roughly 0.6 percent of all reported accidents. To extract meaningful policy, very large amounts of driving data are required to develop a sample size of non-traditional variables (for example, relating to impaired driving, access to taverns, breweries and pubs along commuter routes and proximity of these locations to drivers' residences). In a purely data-driven model, this insight will not be extracted because the database may not initially contain distances from breweries, taverns and pubs to commuter routes. If one were to estimate a model of fatality likelihoods, domain expertise helps fine-tune a data driven model to include distances and therefore measures of "access" to undesirable effects, since the likelihood of alcohol-impaired driving has a well-known causal effect on fatalities. There is also anecdotal and published evidence in the literature that correlates higher fatality rates with robust economic outlook. The contextual awareness value of domain expertise is therefore lacking in models that are developed on pure statistical reasoning. Therefore, it can be reasoned that big data models (and data driven models in general) could potentially suffer from a model-based data-definition disconnect which can cause issues relating to the identification of relevant variables and potential "missing data" issues. While some of this disconnect may be addressed by automated and trained AI systems, human involvement by way of domain expertise and judgment will still remain a requirement.

The discussion above raises an important issue. If the goal of big data modeling is to provide added insight, then the burden of proof lies in the quality of statistical information extracted from those models. In this sense, big-data modeling is not merely an exercise in techniques that accommodate large amounts of data or simply draw associations among variables, but the predominant burden of proof lies in the ability of these models to provide higher-quality inference ("big" inference). The example of drunk-driving fatalities described above is one example of big inference that can be limited without a basic understanding of the sources of

unobserved heterogeneity. Another example of limited inference from big data relates to not adequately making efforts to disentangle causation from correlation, leading to a comingling of the two that can lead to misinformed policy actions (more on this later). These issues can be described as limited big inference in the absence of model-based data definitions using domain expertise. On the other hand, big-data inference can bring in variables that can serve as a source of heterogeneity due to scale. For example, if one were to estimate driving risk models based on naturalistic driving datasets, several non-traditional fine-resolution variables can become available for modeling, such as lane offsetting variables or vehicle kinematic measures such as pitch, yaw and roll.

Figure 1 suggests that the future of big data applications in traffic safety modeling lies at the intersection of strong domain knowledge and quality of extraction of statistical information, and this intersection is heavily influenced by methods that attempt to uncover, to the extent possible, causal effects (after controlling for sources of correlation) and unobserved heterogeneity. Therefore, as a baseline for further evaluation of big data and data driven models, endogeneity models and heterogeneity models can potentially serve as useful tools for both model selection and model definition purposes.

Given the above discussion, with data size and application limitations, what are the potential consequences of trading off predictive capability to understand causality and what factors will compromise our understanding of causality to get better predictive capabilities? Various aspects of this tradeoff are discussed in the following sections, after first discussing causality considerations in safety modeling. In the rest of this paper, we do not discuss causal inference models because, as already indicated, these models have rarely been applied in the accident analysis literature and are not typically based on individual-accident level data.

2. Causality versus Other Explanations in Relationships

The difference between causality and other possible relationship structures involving variables will always be important from a policy action perspective and from the behavioral perspective of improving safety. This is an issue that has been long discussed and remains an important consideration as we enter a “big data” landscape. One possible reason for causality being incorrectly inferred may simply be the fact that the sample being used in safety analysis is itself not representative of the larger population, and thus a relationship estimated for a specific sample may not reflect “true” causal relationships in the larger population. For example, the use of observed accidents, and particularly data conditioned on an accident having occurred, can be potentially problematic for both accident occurrence likelihood and injury-severity statistical modeling because individuals involved in accidents may not be a random sample of the population. That is, the fact that less-safe drivers will be over-represented could potentially present a transference problem of the relationship to the population at large. Further, less-safe drivers may be particularly over-represented in specific types of accidents. To see the problem more clearly, consider a statistical model of the resulting accident-injury severities on a mountain pass. A study of this problem may conclude that high snow accumulations increase the resulting injury-severities in crashes. Injury severity will be known only after a crash has occurred, so it is conditional on a crash having occurred. However, due to the substantial increase in risk involved in driving in snow-related conditions, some drivers may choose to take other modes of travel or avoid traveling adverse weather. Thus, it is possible that the individuals who continue to drive over the mountain pass in adverse conditions are self-selected drivers with risk profiles significantly different from the driving population as a whole. This makes the interpretation of the high-snow-accumulation variable challenging. The variable’s estimated parameter could be picking up the actual effect of

the snow or merely picking up the unique risk characteristics of the drivers who continue to drive in snowy conditions. It is also possible this effect could be much more subtle than this extreme weather case. For example, safe drivers may avoid dangerous roadway sections or dangerous intersections with specific types of traffic controls by choosing alternate routes than drivers with less of a concern for safety (see, for example, Bhat et al., 2014). In such situations, estimating models on observed crash data will tend to overstate the risk of dangerous roadway segments and intersections because these roadways tend to have drivers with higher risk than the overall driving population. Some studies have considered only severe accidents (such as fatalities) thereby potentially compounding the problem because the sample is further restricted making less-safe drivers even more over-represented in the sample.

Another possible, and broader, reason why causality is co-mingled with other associative effects is that many of the explanatory variables used in accident-likelihood and injury-severity models could be viewed as endogenous, causing inconsistent parameter estimates and compromising the interpretation of the statistically estimated parameters (Washington et al., 2011, Abay et al., 2013). For example, seat belt use may be endogenous to injury severity. In other words, individuals who do not wear seat belts may be overrepresented in severe injuries (conditional on an accident), but this may be because those who do not wear seat belts are intrinsically aggressive drivers and this aggressive driving itself may contribute to severe injuries. Thus, one may have to consider seat belt use as an endogenous variable to determine the true causal engineering benefit of seat belt use in preventing serious injuries conditional on an accident. Importantly, such considerations are not merely esoteric scholarly pursuits, but are very germane to assessing the potential effectiveness of various countermeasures and selecting priority measures. In the next few sections, we discuss the ability to investigate causality effects from different types of data/methods.

3. Causality and the Nature of Traditional Accident Data

Of all the many safety-related studies that have been undertaken over the years, those that are based on police-reported accident data have formed the primary basis for developing statistical models to help guide specific safety-related highway and traffic-control improvements. Over the years, the analyses of these data have become increasingly sophisticated, evolving from simplistic regression analyses to highly sophisticated endogeneity/heterogeneity methods. Although the front-line statistical methods used to analyze these data are the mainstay of academic journals, from an application perspective, the culmination of this research is embodied in the Highway Safety Manual (AASHTO, 2010). The Highway Safety Manual approach is based on police-reported vehicle accident data, and has used that empirical basis to provide a practical and readily accessible way of quantifying the likelihood of safety-related impacts of specific highway improvements.

With regard to the likelihood of accidents, using police-reported accident data, studies commonly seek to model the number of accidents occurring on a highway entity, such as a segment of highway or intersection, over some specified time period using count-data or other statistical methods (Lord and Mannering, 2010). Explanatory variables may include roadway characteristics such as traffic volume, lane widths, pavement friction, highway grade and curvature, and so on. Regarding the injury severity of accidents (occupant injury levels such as no injury, possible injury, evident injury, disabling injury, and fatality), discrete-outcome statistical methods are typically applied (Savolainen et al., 2011). Information on injury severity is available only after an accident has occurred (thus conditioned on an accident having occurred). Using data conditioned on the fact that an accident has occurred, the explanatory variables can be expanded from the highway-

segment data used in the accident-likelihood models to include accident-specific variables such as seat-belt use, blood-alcohol level of drivers, weather conditions at the time of the accident, and so on.

As discussed earlier, the use of observed accidents, and particularly data conditioned on an accident having occurred, can be potentially problematic.[3] For sure, the possibility of such selectivity would make the interpretation of the parameters difficult, specifically for weather-related parameters and more so for some modes of highway travel (for example, motorcyclists are particularly likely to self-select in rain and snow as discussed in Mannering, 2018). More importantly, for forecasting with models estimated with traditional police data and even other real-time data, anything that would shift the self-selectivity of road users in adverse weather or on unsafe routes would result in inaccurate predictions. As examples of this self-selectivity shift, newer vehicles with advanced safety features may make drivers more confident in adverse weather conditions, thus changing the mix of drivers in such conditions. Regarding route choice, safe drivers may seek to avoid dangerous roadway segments and intersections, but as congestion increases, they may alter their travel routes as they trade off time and safety and this, in turn, could change the mix of drivers on specific roadway segments.

Methods to attempt to control for self-selectivity and related considerations are discussed in the next section. Data requirements and econometric complexities to implement these procedures for accident data analysis can be formidable obstacles. To circumvent data barriers, many economists have sought more simplistic causal-inference approaches to address

[3] The authors gratefully acknowledge Clifford Winston of the Brookings Institution for identifying this potential issue in traditional safety modeling, and subsequent discussions. There is also the issue of under-reporting of accidents, particularly less severe accidents. That is, minor accidents are less likely to be reported to police, which in turn affects what the analyst sees as observed accidents. This is a known issue that has been shown to create model estimation problems as discussed in Mannering and Bhat (2014).

identification issues and uncover causality, particularly with the application of ordinary least-squares regressions to choice applications (Dale and Krueger, 2002). This is generally done by using control variables such as indicator variables and fixed effects, with the intent of achieving the equivalent of a randomized trial where self-selectivity and endogeneity can be strictly eliminated (Angrist and Pischke, 2009; 2015; 2017). However, the generalizability of the fixed-effects results can be questionable, and even in a truly randomized trial likely temporal shifts in observed behavior can make prediction problematic with ongoing temporal variations inducing unknown errors in fixed effects (Mannering, 2018). In the relatively complex non-linear models of the likelihood and severity of highway crashes that include many explanatory variables relating to roadway characteristics, traffic conditions, weather conditions, and vehicle and driver characteristics, identification of control variables and their incorporation into the model is much more challenging than the more aggregated-data methods applied by economists to address this problem. In addition, predictive application can be quite limited because the variables used as controls may also be of interest for predictive purposes. It is important to note that even analyses that consider the likelihood of an accident, such as accident frequency models, that typically include roadway characteristics and do not include any driver characteristics, are still potentially affected by selectivity. For example, safe drivers may choose to avoid roads with certain characteristics so the observed accidents on specific roads may not be drawn for a random sample of the driving population. Thus, an estimated parameter for a dangerous curve could theoretically be over stated since high-risk (more accident-likely) drivers may be overrepresented on that curve.[4]

The potential bias that selectivity introduces and the effect it may have on prediction is not fully understood, though evidence of the potential biases due to ignoring self-selection has been

[4] While such road selectivity among safe drivers may exist, the authors are unaware of any studies that have quantified this effect.

presented in Shin and Shankar (2013) in an analysis of accident severity likelihoods. But, as pointed out in Mannering (2018), the issue is likely to be very context dependent. For instance, because everyone has a chance of being involved in an accident (even the safest drivers), it may be that an accident data sample collected just so happens to include the full spectrum of individuals from the safest to the least safe. In addition, when considering the injury severities in an accident, it is not clear whether the drivers observed in accidents will have more severe injuries, less severe injuries, or about the same injuries relative to drivers not observed in accidents. For example, drivers frequently appearing in accident data bases may get involved in more accidents of lower-injury severity than those less frequently involved in accidents. It is also important to note that in the cases just mentioned, the resulting injury severities are fundamentally different from traditional endogeneity applications that often have an outcome determined by a choice. In the case of vehicle accidents, once various driver actions are taken, the resulting injury severity is determined by physics where forces are transferred through the vehicle to its occupants (though even the physics involved in the crash are influenced by underlying risk profiles of the driver including vehicle choice and other factors). However, endogeneity of variables in accident data, where the self-selection is based on a choice (such as wearing seat belts or not, or whether a motorist decides to drive at all or not in severe weather, or where traffic engineers choose to place specific types of traffic control devices, or where engineers decide to place additional lighting), is likely to be a more serious issue, as has been demonstrated by Eluru and Bhat (2007), Oh and Shankar (2011), and Bhat et al. (2014).

What is clear, is that selectivity of any form (based on human choice or otherwise) should certainly be considered in the interpretation of any model results that use traditional accident data

(data that only includes accident-involved individuals), and even naturalistic driving and observed traffic data since selectivity on safe and less-safe routes could be a factor.

4. Endogeneity, Unobserved Heterogeneity and Causality

As just discussed, traditional statistical approaches to the analysis of highway safety data (based on observed accident data) have struggled with a variety of statistical issues, most notably endogeneity bias and omitted variables bias, because traditional statistical methods are often estimated with limited data for practical reasons. Despite these limitations, traditional models have the advantage of being accessible and easily applicable, and they have had a measurable real-world impact on highway safety practice. Nonetheless, traditional methods can be substantially enhanced in their value by recognizing elements of endogeneity and unobserved heterogeneity.

Endogeneity considerations (including those involving self-selectivity as discussed earlier) may be handled in one of two broad ways (for more details, please see Bhat and Eluru, 2009). One approach is based off Heckman's seminal work in the 1970s (Heckman, 1979), and has been extended to numerous transportation applications that have been undertaken over the years (for reviews see Mannering and Hensher, 1987; Washington et al., 2011). In particular, using variations of Heckman-style methods, transportation applications have considered a number of issues in this regard, such as selectivity bias corrections for vehicle usage models (Mannering and Winston, 1985; Mannering, 1986a, 1986b; Oh and Shankar 2011; Shin and Shankar 2013), which are needed because, for example, individuals that own newer vehicles (which are capable of being driven more with fewer repairs) are a non-random, self-selected sample of higher-use vehicle owners. There has also been work with selectivity-bias corrections for average speed by route (Mannering et al., 1990), with the idea being that drivers attracted to specific routes are a non-random sample

(for example, faster drivers may be more likely to take freeways and slower drivers may be more likely to take arterials). The basis of the Heckman-style approach is to start with a probabilistic model that captures the selectivity process and then to incorporate the probability of the outcomes under consideration to correct the bias in the model that estimates the magnitude of the outcome. In the case of safety research, this would presumably start with a model that considered individuals' overall probability of being accident involved (or, following the weather-related example earlier, the probability of a motorist driving in adverse weather), and then use this to correct statistical models to the overall frequency and severity of crashes as gathered from observed accidents. However, classic Heckman-style selectivity corrections are manageable because the equation being corrected is a simple linear model with a continuous variable (vehicle usage in miles driven per year, average speed in miles per hour, etc.). In the analysis of accident data, the likelihood of an accident and its resulting injury severities are typically modeled using non-linear count-data and discrete outcome models (Lord and Mannering, 2010; Savolainen et al., 2011), which makes a Heckman-style selectivity correction (using control function approaches) an econometric challenge, particularly if unobserved heterogeneity and other more advanced econometrics are involved in the model as well (Mannering and Bhat, 2014; Mannering et al., 2016).

Another approach to handling endogeneity is inspired by the work of Heckman (1974) and Lee (1983). Rather than use a two-step Heckman type approach, this second approach models the potential endogenous variable jointly with the outcome of interest. While this second approach has been used in a general transportation context for a long time (see Hamed and Mannering, 1993, Bhat, 1996, and Bhat, 1998), the approach has only been relatively more recently applied to models in the safety literature (Eluru and Bhat, 2007, Bhat and Eluru, 2009, Oh and Shankar 2011, Spissu et al., 2009, Pinjari et al., 2009, Abay et al., 2013, and Bhat et al., 2014). Thus, for example, by

modeling seat belt use as well as injury severity in a joint model system (allowing for correlation in the error terms of the underlying equations determining these discrete outcomes, say because of aggressive/risky driving behavior), one can estimate the remaining "true" causal effect of seat belt use on injury severity (addressing also the situation that aggressive drivers are likely to be over-represented in accident-only data). Importantly, through the use of copula methods employed in some of the more recent applications listed earlier of the joint approach, a variety of parametric distributions may be used to characterize the nature of the joint distribution of the errors in the joint system. While the Heckman-type control function approach is generally considered to be more robust to miss-specification of the error distributions, this issue is at least assuaged in the joint model system by testing different distributions forms through copulas and selecting the best fit copula (see Mannering and Bhat, 2014). Further, the joint model system is estimated in a "one-shot deal" and does not incorporate corrections for the standard errors as needed in the second step of Heckman-type methods. The joint model approach also technically does not need the a priori identification of an instrument variable that affects the selection equation (seat belt use in the example above) but not the outcome equation (injury severity) because identification is facilitated through the assumed parametric distribution of the error terms. However, for stability purposes, having at least one variable affecting the selection equation but not the outcome equation is helpful even in the joint model approach, and such exclusion restrictions can be determined through empirical estimations.[5]

While endogeneity models attempt to account for self-selectivity and related broader jointness issues, heterogeneity models (including random parameters models, latent class models

[5] Identification ensures the parameters of interest are uniquely estimable (see for example, Manski 1995; Manski 2009). Lavieri et al. (2016), based on the Generalized heterogeneous data model (GHDM) of Bhat (2015), extend this joint modeling approach by using a small set of common latent stochastic constructs affecting multiple outcomes to generate a parsimonious covariance matrix across the multiple outcomes.

and others) recognize the presence of countless factors that are unlikely to be observed by the data analyst (unobserved heterogeneity) and that influence accident likelihoods and resulting injury severities, despite the presence of a large number of potential explanatory variables. Because heterogeneity models have been the focus of an entire paper recently in the safety field (see Mannering et al., 2016), we do not expend too much space discussing the motivation and methods for such models here. But, using the random parameters application as an example, these heterogeneity models allow the effect of explanatory variables to vary from one accident to the next and from one roadway to the next in (or other units of observation for accident analysis, such as drivers, counties, vehicles, etc.). This can account for a vast variety of unobserved factors and can also potentially mitigate the selectivity issue (that riskier drivers will be over-represented) by giving different parameter values to different observations. However, restrictive distributional assumptions are often made, and prediction can be challenging due to the complexity of the models and the observation-specific estimated parameters. In the process of incorporating unobserved heterogeneity through random-parameter type specifications, it is important that observed heterogeneity not be given less attention. From a causality and policy insight perspective, it is critical that all sources of observed heterogeneity (through observed exogenous variables) be tested and specified first, and unobserved heterogeneity, as referred to in our label of "heterogeneity models", be included to recognize the inevitable presence of the moderating effect of unobserved factors ***after*** accommodating for the presence of observed heterogeneity, rather than ***in-lieu*** of observed heterogeneity.

5. Data Driven Methods, Big Data and Causality

Due to the structure of the models and estimation procedures, traditional statistical models and endogeneity/heterogeneity models have difficulties in processing very large amounts of data (big data). There are a number of data driven methods that have been applied to the analysis of accident data with the intent of uncovering correlations and developing accurate predictive models. Still, the field of accident analysis is ripe for additional applications of non-regression data-driven methods (which are often free from standard parametric assumptions used in traditional regressions). The class of non-regression methods is fairly broad, inclusive of: instance-based algorithms (such as K-nearest neighbor, or support vector machines, etc.); regularization algorithms (such as the least absolute shrinkage and selection operator); decision tree algorithms (such as classification and regression trees); Bayesian networks (such as naïve Bayes and Bayesian networks among others); clustering (K-means, expectations-maximization, etc.); association rule algorithms; artificial neural networks (such as back-propagation and stochastic gradient descent); deep learning algorithms (such as the convolutional neural network, deep belief network, etc.); dimensionality reduction algorithms (such as principal component analysis and its variants); ensembling algorithms (such as boosting and bagging, random forests); feature selection algorithms, reinforcement learners, natural language processing, and so on. In accident analysis, for example, outside of the numerous studies on regression applications, support vector machines have been employed (Li et al., 2008) for the estimation of both frequency and severity outcomes (Li et al, 2008; Li et al, 2011). In addition, artificial neural networks (Abdelwahab and Abdel-Aty, 2001; Abdel-Aty and Pande, 2005; Chang 2005; Delen et al., 2006; Abdel-Aty et al., 2008;), support vector machines (Li et al., 2008; Yu and Abdel-Aty, 2013), Bayesian networks (Hossain and Muromachi, 2012; Sun and Sun, 2015), classification and regression trees and hierarchical tree-based regression (Karlaftis and Goulias, 2002), Bayesian neural networks (Riviere et al.,

2006; Xie et al., 2007), deep belief networks (Pan et al., 2017) and classification trees (Pande and Abdel-Aty, 2006a,b) have been applied to evaluate real-time crash risk.

While the “universe” of data-driven methods is rich for application to accident analysis, several limitations exist relative to traditional econometric and statistical methods. Better prediction is a potential benefit; and the field of statistical reasoning has provided excellent tools for improved “curve fitting” of observations in a fundamental sense. However, questions still remain regarding the appropriate measures of the inferential quality of the data-driven algorithms. First and foremost, among the measures, is the measure of “why?” Are the variables extracted from the data-driven methods able to provide insight into cause and effect that are robust over time, and transferable to other domains? The current answer is that to date no data-driven method has been shown to provide true cause and effect and true cause-and-effect transportability to another domain of search. That is, for example, even if the training dataset was exhaustively analyzed to reveal purported cause and effect, the algorithm would more than likely fail in a different learning scenario with a very different set of hidden causal relationships. Transfer learning, domain adaptation and intelligent causal rule generation are still well beyond the reach of the big-data/AI claims that are published in the literature.

Most existing applications of data-driven methods in the accident-analysis field have also not really dealt with big data (where data-driven methods become the dominant approach), but instead have dealt with data sizes that place them in direct competition with other traditional statistical techniques. With traditional data sets (in terms of size), sophisticated forms of these data-driven methods have been shown to predict accident data with comparatively high accuracy (earning high predictability marks in Figure 1). However, the inability to uncover causality and provide substantive inferences has been a historical weakness of these approaches, often earning

them a “black-box” designation because of the difficulty of unraveling how specific elements might influence predictions with these approaches (giving it low marks for causality in Figure 1). While data-driven methods are likely to become increasingly popular with the emergence of truly high dimensional big-data in transportation safety (National Academies, 2013), the fundamental limitations relating to causality must still be given consideration in the interpretation and application of results. The bottom line is that, while data-driven methods may do well in capturing associations between one variable and another (that is, how variation in a variable influences another variable), they do not intrinsically study the issue of what exactly is the root cause of why variation in one variable influences another variable. While one could claim that this is the same even with traditional “structure-based” econometric analyses, there is some level of domain theory and knowledge that underlies structure-based analyses that facilitates drawing more causal inferences (especially when endogeneity and heterogeneity issues are recognized). In particular, traditional structure-based methods are driven by well-informed causal frameworks based on domain knowledge. While the relationships implicit in these frameworks may be characterized as assumptions by some, it is important to note that assumptions need to be made in all kinds of analyses, including data-driven analyses (for example, regardless of the methods used, one has to define what are the outcome variables and what are the explanatory variables, and not every variable can be associated with each other variable).

6. Discussion and Conclusions

Safety analysts often face challenges in trading off the predictive capability of the methodological approach with its ability to uncover underlying causality. The trade-offs must consider available data in terms of the number of variables and number of observations as well as the intended use of the results. In some practical applications, highway safety engineers may need

to know highly specific information. For example, what impact would increasing the shoulder width from 4 to 6 feet on a two-lane rural road with specific traffic characteristics and geographic location have on the likelihood and resulting injury severity of crashes. Getting to this level of detail necessitates specific data requirements and advanced methodologies, and likely some compromise between predictive accuracy and underlying causality. Those who strongly support causality as the only correct approach are often highly critical of methods that do not fully address causality, sometimes arguing that no prediction is better than a prediction based on a flawed causal model (although they rarely if ever provide empirical evidence to support this argument). But this argument does not fully appreciate the potential benefits of having some level of predictive capability. In contrast, those who consider purely data-driven analyses neglect potential insights into underlying causality. Without an understanding of underlying causality, changes in vehicle technology, roadway features, and human behavior may fundamentally shift model parameters that would ultimately impact predictions and safety-policies.

An ideal model would be one that uncovers causality, has excellent predictive capabilities, and is scalable to very large data. However, with currently available methods, safety analysts are often forced into a causality/prediction tradeoff that can entail serious compromises. Thus there is a clear need in the safety field to ground intrinsically predictive models within causal frameworks, while also taking insights from intrinsically predictive models (especially from big data) to improve upon causal structures through insights from associations involving variables not typically available in traditional safety data. One promising direction for future research would be a hybrid modeling approach of data-driven and statistical methods (with strong consideration to causal elements). Such a hybrid approach is likely to be perfected over time as integrative techniques are perfected and access to more and more big data becomes available. However, during this

development period it is important that strong domain knowledge remain at the front and center of all analytic approaches and their subsequent interpretations for predictions and policy actions.

Acknowledgements: The authors gratefully acknowledge support provided by the Center for Teaching Old Models New Tricks (TOMNET), a University Transportation Center sponsored by the US Department of Transportation through Grant No. 69A3551747116 and the Data Supported Transportation Operations and Planning Center, a University Transportation Center sponsored by the US Department of Transportation through Grant No. DTRT13GUTC58. The discussions with Clifford Winston with regard to causality models are also gratefully acknowledged.

References

AASHTO, 2010. Highway Safety Manual. American Association of State Highway and Transportation Officials, Washington, DC.

Abay, K.A., Paleti, R., Bhat, C.R., 2013. The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity, Transportation Research Part B 50, 74-89.

Abdel-Aty M., Pande A., Das A., Knibbe W., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transportation Research Record 2083, 153-161.

Abdel-Aty M., Pande A., 2005. Identifying crash propensity using specific traffic speed conditions. Journal of Safety Research 36(1), 97-108.

Abdelwahab, H., Abdel-Aty, M., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6-13.

Angrist, J., Pischke, J., 2009. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, Princeton, NJ.

Angrist, J., Pischke, J., 2015. Mastering metrics: The path from cause to effect. Princeton University Press, Princeton, NJ.

Angrist, J., Pischke, J., 2017. Undergraduate econometrics instruction: Throughout classes, darkly. Working paper 23144, National Bureau of Economic Research, Cambridge, MA.

Bhat, C.R., 1996. A generalized multiple durations proportional hazard model with an application to activity behavior during the work-to-home commute. Transportation Research Part B 30(6), 465-480.

Bhat, C.R., 1998. A model of post-home arrival activity participation behavior. Transportation Research Part B 32(6), 387-400.

Bhat, C.R., 2015. A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. Transportation Research Part B 79, 50-77.

Bhat, C.R., 2018. New matrix-based methods for the analytic evaluation of the multivariate cumulative normal distribution function. Transportation Research Part B 109, 238-256.

Bhat, C.R., Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. Transportation Research Part B 43, 749-765.

Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. Analytic Methods in Accident Research 1, 53-71.

Brathwaite, T., Walker, J., 2018. Causal inference in travel demand modeling (and the lack thereof). Journal of Choice Modelling 26, 1-18.

Chang, L.-Y., 2005. Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. Safety Science 43(8), 2005, 541-557.

Dale, S., Krueger, A., 2002. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. Quarterly Journal of Econometrics 117(4), 1491-1527.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention 38(3), 434-444.

Eluru, N., Bhat, C.R., 2007. A joint econometric analysis of seat belt use and crash-related injury severity. Accident Analysis and Prevention 39(5), 1037-1049.

Hamed, M., Mannering, F., 1993. Modeling travelers' postwork activity involvement: toward a new methodology. Transportation Science 27(4), 381-394.

Heckman, J., 1974. Shadow prices, market wages and labor supply. Econometrica 42(4), 679-694.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47(1), 153-161.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accident Analysis and Prevention 45, 373-381.

Karlaftis, M., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis and Prevention 34(3), 357-365.

Lavieri, P., Bhat, C.R., Pendyala, R.M., Garikapati, V.M., 2016. Introducing latent psychological constructs in injury severity modeling: multivehicle and multioccupant approach. Transportation Research Record 2601, 110-118.

Lee, L., 1983. Generalized econometric models with selectivity. Econometrica 51(2), 507-512.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis and Prevention 40(4), 1611-1618.

Li, Z., Liu, P., Wang, W., Xu, C., 2011. Using support vector machine models for crash injury severity analysis. Accident Analysis and Prevention 45, 478-486.

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

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A 44(5), 291-305.

Manski, C., 2009. Identification for Prediction and Decision. Harvard University Press, 368 pages.

Manski, C., 1995. Identification Problems in the Social Sciences, Harvard University Press, 194 pages.

Mannering, F., 1986a. A note on endogenous variables in household vehicle utilization equations. Transportation Research Part B 20(1), 1-6.

Mannering, F., 1986b. Selectivity bias in models of discrete/continuous choice: An empirical analysis. Transportation Research Record 1085, 58-62.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1-13.

Mannering, F., Abu-Eisheh, S., Arnadottir, A., 1990. Dynamic traffic equilibrium with discrete/continuous econometric models. Transportation Science 24(2), 105-116.

Mannering, F., Hensher, D., 1987. Discrete/continuous econometric models and their application to transport analysis. Transport Reviews 7(3), 227-244.

Mannering, F., Bhat, C., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1, 1-22.

Mannering, F., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16.

Mannering, F., Winston, C., 1985. A dynamic empirical analysis of household vehicle ownership and utilization. Rand Journal of Economics 16(2), 215-236.

National Academies, 2013. Frontiers in massive data analysis. The National Academies Press, Washington, DC.

Oh J., Shankar, V., 2011. Corridor safety modeling via segmentation: A selectivity bias perspective to crash count data. Lambert Academic Publishing, 84 pages.

Pan, G., Liping, F., Thakali, L., 2017. Development of a global road safety performance function using deep neural networks. International Journal of Transportation Science and Technology 6(3), 159-173.

Pande A., Abdel-Aty, M., 2006a. Assessment of freeway traffic parameters leading to lane-change related collisions. Accident Analysis and Prevention 38(5), 936-948.

Pande A., Abdel-Aty M., 2006b. A comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transportation Research Record 1953, 31-40.

Pinjari, A.R., Bhat, C.R., Hensher, D.A., 2009. Residential self-selection effects in an activity time-use behavior model. Transportation Research Part B 43(7), 729-748.

Riviere C., Lauret P., Manicom, J., Page Y., 2006. A Bayesian neural network approach to estimating the energy equivalent speed. Accident Analysis and Prevention 32(2), 248-259.

Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention 43(5), 1666-1676.

Shin, S., Shankar, V., 2013. Selection bias and heterogeneity in severity models: Some insights from an interstate analysis. Lambert Academic Publishing, 132 pages.

Spissu, E., Pinjari, A.R., Pendyala, R.M., Bhat, C.R., 2009. A copula-based joint multinomial discrete-continuous model of vehicle type choice and miles of travel. Transportation 36(4), 403-422.

Sun, J., Sun, J., 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. Transportation Research Part C 54, 176-186.

Washington, S., Karlaftis, M., Mannering, F., 2011. Statistical and econometric methods for transportation data analysis. Second Edition, Chapman and Hall/CRC, Boca Raton, FL.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural networks: An empirical analysis. Accident Analysis and Prevention 39(5), 922-933.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash evaluation. Accident Analysis and Prevention 51, 252-259.

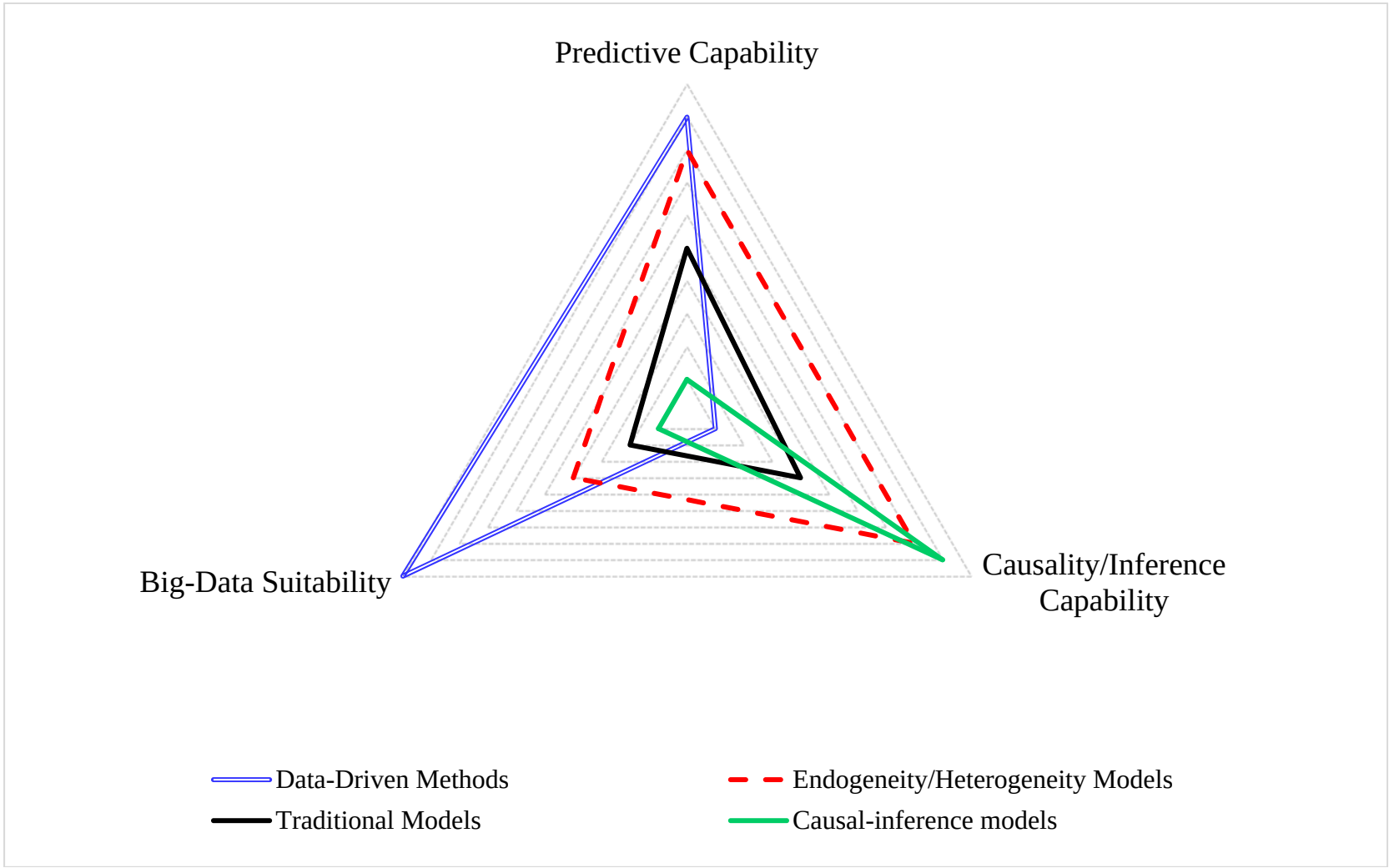

Figure 1. Current modeling trade-offs between relative big-data suitability, predictive capability and causality/inference capability.