

## CAUSAL INFERENCE IN TRANSPORTATION SAFETY STUDIES: COMPARISON OF POTENTIAL OUTCOMES AND CAUSAL DIAGRAMS

BY VISHESH KARWA, ALEKSANDRA B. SLAVKOVIĆ AND ERIC T. DONNELL

*Pennsylvania State University*

The research questions that motivate transportation safety studies are causal in nature. Safety researchers typically use observational data to answer such questions, but often without appropriate causal inference methodology. The field of causal inference presents several modeling frameworks for probing empirical data to assess causal relations. This paper focuses on exploring the applicability of two such modeling frameworks—Causal Diagrams and Potential Outcomes—for a specific transportation safety problem. The causal effects of pavement marking retroreflectivity on safety of a road segment were estimated. More specifically, the results based on three different implementations of these frameworks on a real data set were compared: Inverse Propensity Score Weighting with regression adjustment and Propensity Score Matching with regression adjustment versus Causal Bayesian Network. The effect of increased pavement marking retroreflectivity was generally found to reduce the probability of target nighttime crashes. However, we found that the magnitude of the causal effects estimated are sensitive to the method used and to the assumptions being violated.

**1. Introduction.** An estimated 2.2 million people suffered some kind of transportation-related injury in 2007. About 87 percent of these injuries resulted from highway crashes [Bureau of Transportation Statistics (2007)]. Transportation safety management aims at identifying causes of such crashes, developing countermeasures to mitigate crashes, and evaluating the effectiveness of a safety countermeasure. It is well known that causal propositions of this kind, and their effect sizes, are best estimated from randomized experiments.

The types of data available in transportation safety studies are primarily observational, which makes it difficult to consistently estimate causal effects of countermeasures. In this paper we evaluate and compare the application of two commonly used causal inference frameworks (one that is commonly applied in computer science and another that is commonly applied in statistics) to transportation safety. In particular, the aim of this paper is twofold:

- To introduce a unique transportation safety data set, created from multiple sources, and to highlight the problems associated with the data used in safety studies.

---

Received August 2009; revised October 2010.

*Key words and phrases.* Causal inference, potential outcomes, causal Bayesian networks, observational studies, transportation safety, nighttime crash data.

- To explore the application of causal inference methods in transportation safety studies and document the issues associated with the analyses. We do this by estimating the causal effect of pavement marking retroreflectivity (*PMR*) on target nighttime crashes using the causal modeling frameworks of the Potential Outcomes (PO) and Causal Diagrams (CD), and then compare the results.

Causal inference methods in transportation safety studies have received little attention. Davis (2000) provides a review and notes that the assignment mechanism must be included in statistical models to consistently estimate the effect of any countermeasure on crashes. Davis (2004) uses Pearl's causal Bayesian networks (CBN) for crash reconstruction and examines token causal claims to answer single-event causation questions; see Eells (1991) for a review of token and type causes.<sup>1</sup> In contrast to single-event causation, our work examines the application of causal inference methods to population level causal effects in transportation safety studies. Population causal claims are more applicable to transportation safety management since they reflect the effect of countermeasures in a population as opposed to singular causal claims which are geared more toward accident reconstruction and liability issues. We examine the hypothesis that low *PMR* levels are causative agents for an increase in the risk of nighttime crashes. Causal effects are estimated using the PO framework and CD framework, and the results are compared.

It has been shown that the PO framework and the CD framework are mathematically (theoretically) equivalent; see Pearl (2000), Chapter 7. However, there are different statistical implementations of these frameworks that offer different paths to estimate causal effects, and, in practice, the results may or may not be similar. For instance, the PO framework is commonly implemented using propensity score matching or inverse propensity score weighting, and the CD framework is commonly implemented using CBNs. Several assumptions relating to estimation of a causal effect from observational data are reviewed in Sections 4 and 5. Apart from these, differences in the estimates of causal effect could arise due to additional assumptions required by each framework, inherently tied to the aforementioned statistical implementations. For instance, the CD framework requires the use of a causal graph that represents the qualitative causal mechanism of the data generating process. This graph can be obtained from prior knowledge and/or data. The algorithms used to recover causal graphs from data require additional assumptions such as faithfulness and some require the data to be either discrete or Gaussian. The PO framework does not require these additional assumptions since it does not require such causal graphs.<sup>2</sup> Furthermore, while the CD framework

---

<sup>1</sup>Token causes are those associated with a single unit or event, for instance, "Since Jane was speeding, she ended up in an accident." On the other hand, type causal claims are associated with a population, for instance, "Speeding causes crashes."

<sup>2</sup>The PO framework requires the analyst to qualitatively model the treatment assignment mechanism (see Section 4.1). The use of graphs to represent the treatment assignment mechanism could make this easier to communicate; see the discussion section.

enables one to work with the complete data set, the PO framework could lead to elimination of a part of the data set if one uses matching. Sfer (2005) and Fienberg and Sfer (2006) show in a simple simulated logistic regression example that there is an implicit agreement between PO and CD frameworks. However, we are not aware of a study that explores the differences, compares the results of both modeling frameworks on real-life observational data, and examines the advantages and disadvantages associated with applying each method in a practical context.

A complete and rigorous comparison of both frameworks requires considerations of all possible implementations, which is beyond the scope of this paper. Here, we focus on two commonly used implementations in practice, both of which use the complete data set in order to allow for a better comparison of the results. We implement the PO framework using *inverse propensity score weighting (IPW)* and *regression adjustment*, and the CD framework using *discrete causal Bayesian Networks (CBN)*; the implementation details are provided in Sections 4 and 5. We also implemented the PO framework using *propensity score matching*<sup>3</sup> since it is a popular alternative to IPW. However, as previously noted, matching could lead to elimination of part of the data and the comparison of estimates with those from CBN may be biased due to differences in the data themselves. This issue is further discussed in Sections 4 and 7 of the paper.

The remainder of the paper is organized as follows: Section 2 introduces the problem statement; Section 3 introduces the data set and the design of the hypothetical experiment to estimate the causal effects; Sections 4 and 5 present the analyses and results of the PO framework and the CD framework; Section 6 compares the results; and, Section 7 concludes with a discussion.

**2. Problem description.** The traffic accident fatality rate increases by almost 75 percent during the period of time between 9 p.m. and 6 a.m. [National Highway Traffic Safety Administration (2007)]. This raises a question of what can be done about the fact that there is a greater rate of crashes at night than during the day? It is hypothesized that low pavement marking visibility may be one cause of the increased rate of nighttime crashes. One of the objectives of the paper is to examine this hypothesis.

Pavement markings delineate the limits of the traveled way and provide drivers navigation and control guidance. During the daytime, drivers are likely to use a combination of pavement markings, other traffic control devices (e.g., signs) and visual cues along the roadside (e.g., utility poles, vegetation, etc.) to navigate a roadway. Pavement markings have an important role at night. Apart from delineating the road, pavement markings reflect the light shone from a car's headlamps back to the driver, thus enabling the driver to see the limits of the traveled way.

---

<sup>3</sup>We report the estimates of causal effects from matching in the paper, and provide the implementation details in the supplementary material.

This is known as retroreflectivity and is measured in millicandelas per square meter per lux ( $\text{mcd}/\text{m}^2/\text{lux}$ ). Retroreflectivity in pavement markings is provided by glass spheres that are dropped-on or premixed with a wet pavement marking material. *PMR* degrades over time because of fatigue to the material and its bond strength with glass spheres or the pavement surface. State transportation agencies typically re-stripe pavement markings after the end of their useful service life, defined as the time when retroreflectivity falls below a minimum threshold level.

A considerable amount of research has been carried out regarding the safety benefits of *PMR*. To answer the question if improving *PMR* has any effect in reducing the number of traffic crashes, most of the published literature used regression models with observational data, ignoring the treatment assignment mechanisms; for a literature review, refer to Bahar et al. (2006) and Donnell, Karwa and Sathyanarayanan (2009). Also Donnell, Karwa and Sathyanarayanan (2009) point out that none of the studies explicitly relate the in-situ *PMR* levels to the crash event. This is due to the fact that *PMR* levels and crash data are obtained from separate sources and merging them is difficult. The problems associated with merging these databases have been described in Karwa (2009). Donnell, Karwa and Sathyanarayanan (2009) was the first study that explicitly combined the *PMR* data (which is representative of real life degradation patterns of *PMR*) with crash data to develop a comprehensive database. This work did show that there were statistical associations between *PMR* and nighttime crashes. We use this database to examine, for the first time, the nature of the effect of *PMR* on traffic safety (defined in Section 3) using the PO and the CD frameworks. The results are compared with a discussion of the application of the two methods, in an attempt to determine their possible broader application in transportation safety studies.

**3. Description of the data and design of the study.** The fundamental unit of operation in this paper is a homogeneous road segment; homogeneous refers to having uniform geometric characteristics such as number of lanes, lane width and shoulder width along a roadway segment. A segment within a fixed time period is considered to be different from the same segment at any other time period. A fixed time period of one month was selected to ensure homogeneity of *PMR* levels and other characteristics of a segment. For instance, the *PMR* level and monthly traffic volumes can be assumed to be reasonably uniform within this period.

Crash and *PMR* data were collected from three districts in North Carolina for a period of 2.5 years. As noted in Section 2, the data were obtained from two different sources. The *PMR* data were measured by a private contractor using a mobile retroreflectometer with a 30-meter geometry. These data were collected on two-lane and multi-lane highways in North Carolina, approximately every 6 months. All pavement markings were of thermoplastic material. Since retroreflectivity estimates were not measured at the exact time and place of occurrence of the crash, a neural network model was used to interpolate the values of retroreflectivity on the segments where crashes were observed; see Karwa and Donnell (2011).

TABLE 1  
*Roadways with pavement marking and crash data available for safety analysis*

County	Route	District	Begin MP	End MP	Number of lanes	Total length (miles)	Nighttime target crashes
Bertie	US13	1	0.00	11.07	4	11.07	8
Gates	US13	1	0.00	14.78	2	14.78	14
Northampton	US158	1	12.35	24.04	2	11.69	4
Washington	US64	1	10.54	19.67	2	9.13	2
Durham	I-85	5	7.88	14.19	4	6.31	5
Durham	US15	5	3.66	6.56	4	2.90	9
Durham	NC98	5	0.00	11.06	2	9.44 <sup>a</sup>	15
Durham	NC157	5	0.70	3.98	2	3.28	2
Granville	I-85	5	0.00	23.73	4	1.80 <sup>b</sup>	5
Person	US158	5	0.00	22.36	2	16.22	5
Vance	I-85	5	0.00	14.47	4	12.47 <sup>c</sup>	46
Vance	US158	5	0.00	8.96	2	5.94 <sup>d</sup>	1
Wake	I-40	5	6.47	20.19	4/6/8	13.72	67
Wake	NC98	5	0.00	4.55	2	4.55	1
Warren	I-85	5	0.00	9.88	4	9.88	39
Warren	US158	5	12.38	22.93	2	10.55	0
Catawba	I-40	12	13.13	19.67	4	6.54	10
Iredell	I-40	12	0.00	22.76	4	22.76	63
Iredell	I-77	12	14.75	23.75	4	9.00	17
Total						182.03	313

<sup>a</sup>Roadway inventory and crash data were not available between mileposts 0.17 & 1.79.

<sup>b</sup>Roadway inventory and crash data were not available between mileposts 0.76 & 22.69 and 22.73.

<sup>c</sup>Roadway inventory and crash data were not available between mileposts 3.96 and 5.96.

<sup>d</sup>Roadway inventory and crash data were not available between mileposts 3.77 and 6.79.

The roadway inventory and crash event data were obtained from the Highway Safety Information System (HSIS) data files, maintained by the Federal Highway Administration (FHWA). These data were collected for 19 roadway sections in North Carolina. There were 192 total segments (segments are a subset of sections) that corresponded to the 19 sections of roadway where PMR estimates were computed based on the degradation model. Table 1 shows the sections where roadway inventory, crash and PMR data could be linked. There are a total of 5,916 observations, based on 192 segments, 12 months of data per year for each segment and approximately 2.5 years of crash data per segment.<sup>4</sup>

<sup>4</sup>The assumption of independent segments over time is a common assumption in the safety prediction literature that also shows that weak temporal (or spatial) correlations result in a loss of estimation efficiency but not bias.

Crashes that satisfied the following criteria, referred to as *target crashes*, were used in the analysis: occurred during dusk, dawn or at night; dry roadway surface conditions; ran-off-the-road crashes; fixed object crashes (off-road); and opposite- or same-direction sideswipe crashes. Crashes that satisfied the following criteria were excluded from the analysis: work zone area; no alcohol-involvement; weather contributing circumstances; roadway contributing circumstances. It must also be noted that the data are sparse due to rarity of crashes. Only 6 percent of the total number of segments had more than one target crash during the study period.

*Safety* of a road segment is defined as a *Bernoulli* random variable, taking the value 1 if there was at least one target crash in the segment during the treatment (or control) application period, and 0 if there was no target crash in the segment. The safety of a segment is stochastic and each segment has a fixed probability  $p$  of at least one target crash occurring, which is assumed to be an inherent property of the road segment. This definition was chosen to ensure the absence of confounders between safety and *PMR*, based on the past *PMR* related safety literature. For instance, if it was clear according to the police crash report that a particular crash occurred due to driving under the influence of drugs or alcohol, such a crash would have been deemed to occur because of human error, and thus excluded from the current analyses. Similarly, crashes in which weather was a contributing factor (such as heavy snow or icy road conditions) were also excluded from the analyses. Weather conditions, human errors, etc. are stochastic factors that may cause crashes but not an inherent property of the segment; hence, any crash occurring due to such conditions would fall into the error term of the observed safety of a road segment.

*Treatment* variable on a segment is defined as the application of *PMR* with levels {*Low*, *Med*, *High*}; the exact range of *PMR* levels for each class is specified in Table 2. *Control* is defined as application of pavement markings at one of the two remaining levels of retroreflectivity. Out of the total sample size ( $N = 5,916$ ), about 36 percent of the segments had *Low* levels of *PMR*, about 46.5 percent had *Med* levels and the remaining segments had *High* levels of *PMR*. The assignment of *PMR* levels is clearly not random.

Apart from the data on *PMR* and the crash counts per month, data on 12 other covariates were collected. See Table 2 for definitions and summary statistics of random variables representing information on the attributes of a segment such as the shoulder width, number of lanes, presence of a median, traffic flow characteristics such as monthly traffic volumes (hereafter referred to as *ADT*), percentage of trucks, location related variables such as the geographic district in which the segment is located, the urban or rural setting of the segment location, and the terrain type. The data are very sparse, which is typical of safety data. For instance, in the five way cross-classification of the entire sample with respect to the discrete variables *District*, *Terrain*, *PMR*, *Multilane* and *Safety*, 58 percent of the cells have sampling zeros.

TABLE 2

*Definition of variables and their descriptive statistics. Mean (st. dev), [Min, Max] values are given for the continuous variables and number of observations (percentage) for each level of categorical variables. Total sample size,  $N = 5,916$*

Variable	Definition	Descriptive statistics
Right Shoulder	Outer shoulder width in feet on right side of roadway	9.39 (4.03) [0, 14]
	0 if shoulder width $\leq 7$ feet	1,238 (9.97%)
	1 otherwise	4,678 (90.03%)
ADT	Annual average daily traffic adjusted for a month, vehicles per day	30,383 (27,580) [1,615, 114,400]
	0 if ADT $\leq 30,000$ vehicles per day	2,957 (49.9)
	1 otherwise	2,959 (50.1)
Truck	Percentage of ADT that consists of heavy vehicles	14.8 (8.5) [0, 83]
	0 if Percentage of Trucks $\leq 18$	2,065 (35)
	1 otherwise	3,851 (65)
PMR	Mean PMR of all markings on a segment (mcd/m <sup>2</sup> /lux)	227 (65) [139, 447]
	Low if $139 < \text{retroreflectivity} \leq 200$	2,134 (36)
	Medium if $200 < \text{retroreflectivity} \leq 280$	2,748 (46.5)
	High if $280 < \text{retroreflectivity} \leq 447$	1,034 (17.5)
Age	Time (in months) elapsed since the application of markings on a segment	15.5 (8.63) [1, 30]
	0 if Age $\leq 1$	1,761 (30)
	1 if $10 \leq \text{Age} \leq 20$	1,980 (33.5)
	2 otherwise	2,175 (36.5)
Multilane	1 if there is more than 1 lane in each direction	3,868 (65.4)
	0 if there is 1 lane in each direction	2,048 (34.6)
Median	1 if the segment contains a median	4,018 (68)
	0 if the segment has no median	1,898 (32)
Safety	1 if at least 1 target crash occurred in the segment, during the month	376 (6.36)
	0 otherwise	5,540 (93.65)
Urban	1 if the segment is located in an urban area	2,680 (45.3)
	0 if the segment is located in a rural area	3,236 (54.7)
Terrain	1 if the segment is on flat terrain	1,320 (22.3)
	0 if the segment is on a rolling terrain	4,596 (77.7)
District	0 if the segment is located in District 1	1,290 (21.8)
	1 if the segment is located in District 5	3,286 (55.5)
	2 if the segment is located in District 12	1,340 (22.7)

As per Rubin (2008) and Maldonado and Greenland (2002), we conceptualize our problem as a hypothetical experiment to make the problem statement clear. Consider a population of homogeneous road segments. We wish to examine the effect of increased *PMR* on the safety of a road segment. Ideally, we would like to apply treatment (e.g., *PMR = Low*) and control (e.g., *PMR = High*) to the same population and observe the expected safety outcome to measure the causal effect.<sup>5</sup> The causal effect is defined as the risk ratio of expected safety outcome under the treatment and controls, for the same population. Since this is not possible in practice, we use analytical simulations of this process. Sections 4 and 5 describe the conceptualization of this hypothetical experiment under two different frameworks.

**4. Potential outcomes framework.** In this section we present the PO framework as applicable to the current study as well as the results of the analysis. Section 4.1 defines the causal estimands [the “science,” see Rubin (2005)] and explains the treatment assignment mechanism and the assumptions required to estimate causal effects from observational data. Section 4.2 provides details about the implementation of the PO framework, that is, the use of inverse propensity score weighting to achieve balance in the data and the use of regression adjustment to estimate the average causal effect (ACE), after balancing. Section 4.3 presents the results of the analysis.

*4.1. Treatment assignment, potential outcomes and assumptions.* Let the homogeneous segments be indexed by the letter  $i$ . We focus on one hypothetical experiment at a time, introduced in Section 3, and estimate the effect of a binary *PMR* treatment on safety from a sample of segments. Extension to the case of three levels of treatment of *PMR* is performed using the method proposed by Rubin (1998) which involves creating a separate propensity score model for each two-level treatment comparison, equivalent to conducting three hypothetical experiments. Thus, in the present case, three separate propensity score models are estimated. This method is followed since it is difficult to simultaneously balance all three treatment groups on all covariates.

The PO framework uses potential outcomes as the fundamental element to estimate the causal effects. We denote the treatment variable by  $T_i$ , where  $T_i = 0$  denotes no treatment or the baseline condition for unit  $i$ , and  $T_i = 1$  denotes the treatment condition. For instance, if we wish to estimate the effect of changing the *PMR* levels from *Med* to *Low*, the treatment would be application of *PMR = Low* and the control would be application of *PMR = Med*. Associated with each segment are two potential outcomes: *Safety*( $S$ ) of the segment at the end of a month after the treatment has been applied,  $S_i(T = 1)$ , and *Safety* of the same segment

---

<sup>5</sup>In practice, “treatment” would be application of Higher *PMR*, but here, we define it as application of Low *PMR* to obtain causal risk ratios greater than 1.

at the end of the month if there was no treatment, that is, the baseline condition was applied,  $S_i(T = 0)$ . Covariates that represent the attributes of a segment are denoted by the vector  $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  for unit  $i$ . The average causal effect (ACE) of the treatment relative to the baseline for segment  $i$  is then defined as a causal risk ratio

$$(4.1) \quad ACE_{0 \text{ to } 1} = \frac{E[S_i(1)]}{E[S_i(0)]},$$

where  $E[\cdot]$  denotes expectation and  $E[S_i(\cdot)] = E[E[S_i(\cdot)|P(S_i(\cdot) = 1)]]$ . We assume that  $E[S_i(0)] > 0$ , and drop the  $T$  in the notation for simplicity.

For any particular segment, only one of the two values of  $S(0)$  and  $S(1)$  can be observed. This has been termed the “fundamental problem of causal inference” [Rubin (1978); Holland (1986)], because of which unit level causal inferences are not possible.<sup>6</sup> However, given certain assumptions which are outlined below, the ACE of the treatment on a population can be estimated consistently.

*Stable unit treatment value assumption* (SUTVA) [Rubin (1990)]. This assumption states that the treatment applied to one unit does not affect the outcome of any other unit and that there are no hidden versions of the treatment (i.e., no matter what mechanism was used to apply the treatment to the unit, the outcome would be the same). The last part is sometimes referred to as the consistency assumption [Cole and Frangakis (2009)]. We make this assumption in the current study even though the treatment has been applied in groups (several segments along a particular route may have the same value of *PMR*). The following example illustrates a scenario where this assumption could be violated. Consider two consecutive segments on the same route. A vehicle traveling on this road could end up in a crash in segment 2 because of low visibility on segment 1. Such scenarios are not uncommon, but when crashes are reported, the reporting officer estimates the approximate segment location where the crash was initiated (after careful analysis of the evidence available at the crash site, such as skid marks, etc.) and the crash is attributed to that segment. Hong and Raudenbush (2005) extend SUTVA to account for possible interference among segments, but we do not consider this extension here.

*Positivity.* The positivity assumption states that there is a nonzero probability of receiving every level of treatment for every combination of values of exposure and covariates that occur among individuals in the population [Rubin (1978); Hernan and Robins (2006)]. We make the positivity assumption since, in principle, each segment can be assigned any level of *PMR* treatment.

*Unconfoundedness.* The treatment mechanism is said to be unconfounded given a set of covariates  $x_i$ , if the treatment is conditionally independent of the potential

---

<sup>6</sup>Except in cases where the functional mechanism of causation is known, this is called token causation; see Pearl (2000), Chapter 7.

outcomes given the covariates

$$(4.2) \quad t_i \perp\!\!\!\perp S(0), S(1) \mid x_i.$$

In a randomized experimental setting,  $t_i$  would be unconditionally independent of the potential outcomes by design. In the current setting this is not the case, but the treatment assignment can be made conditionally independent of the potential outcomes by balancing on observed covariates. This requires modeling the treatment assignment mechanism as explained below.

4.1.1. *Treatment assignment mechanism.* Let  $P(T_i = 1 \mid X_i)$  be the propensity score. The propensity scores are used in assignment of the treatment to the segments in order to achieve balance. Following Rosenbaum and Rubin (1983), treatment assignment is strongly ignorable given a vector of covariates  $X$  if unconfoundedness and common overlap hold:

$$(4.3) \quad S(0), S(1) \perp\!\!\!\perp T \mid X,$$

$$(4.4) \quad 0 < P(T = 1 \mid X) < 1.$$

In the current setting, the treatment assignment mechanism can be assumed to consist of two parts. In the first part, pavement markings are applied by transportation agencies at different segments with a similar level of retroreflectivity (usually falling into the category *High*). In the second part, the markings are left to deteriorate over a period of 2.5 years. The *PMR* levels decrease due to stress on the pavement marking material from vehicle passes and natural factors such as weather. Thus, it can be assumed that nature assigns a level of *PMR* based on the time elapsed since the initial application period (*AGE*) of the pavement marking, number of vehicle passes and weather conditions. The assignment of *PMR* levels for each segment depends on the *AGE* of the marking within the segment and the number of vehicle passages over the segment within that period. Apart from this, *PMR* levels may also depend on the location of the segment (due to differences in weather conditions), the percentage of trucks that compose the traffic volumes (stress on the marking material is generally greater due to heavier vehicles) and the number of lanes in a segment (the *PMR* levels used are the average of the different pavement markings present in a segment, and multi-lane segments generally have at least one extra marking when compared to two lane segments). All of these variables are included to form a rich propensity score model that specifies the assignment mechanism.

4.2. *Inverse propensity score weighting and regression adjustment.* Below is a description of a particular implementation of the PO framework that estimates ACE and ensures that the assumptions outlined in the previous sections are satisfied. This implementation is a form of doubly robust estimation; see Bang and Robins (2005). The same two steps are repeated to estimate the ACE for each of the three comparisons:

Step 1 Estimate the propensity score model  $\pi = P(T = 1 | X)$  and achieve balance, via *inverse propensity score weighting (IPW)*; see Hirano and Imbens (2001).

Step 2 Estimate ACE, via *regression adjustment method*.

*Inverse propensity score weighting*: The Generalized Boosting Model [GBM, McCaffrey, Ridgeway and Morral (2004)], a multivariate nonparametric technique, was used to estimate the propensity scores. Although logistic regression is the most common way to estimate the propensity scores, studies have shown that other methods can offer considerable improvement [e.g., see Lee, Lessler and Stuart (2009)]. The analysis was carried out using the “twang” package in R [Ridgeway, McCaffrey and Morral (2006)]. Weights were computed from the estimated propensity scores and balance in the data is tested using the estimated weights. Balance is tested by comparing the distributions of key covariates in the treatment and control groups of the weighted data using the Kolmogorov–Smirnov (KS) test statistic. A weight of  $\frac{1}{\pi}$  is assigned to the treatment group and a weight of  $\frac{1}{1-\pi}$  to the control group, where  $\pi$  is the estimated propensity score.<sup>7</sup> Hirano, Imbens and Ridder (2003) show that the use of a nonparametric estimate of propensity score to estimate weights, rather than the true propensity score, can lead to an efficient estimate of ACE. If balance is not achieved, the propensity score model is re-specified and the process is repeated. The model is re-specified by changing the tuning parameters of the boosting model. The tuning parameters are the number of trees used to fit the model, the shrinkage parameter and the interaction depth; see McCaffrey, Ridgeway and Morral (2004) for details. In the selected propensity score models, we used an interaction depth of 2 (i.e., the model fits all two-way interactions), the shrinkage parameter was set at 0.01 and the number of trees were set at 15,000.

*Regression adjustment*: Once the samples (segments) are divided into control and treatment groups and balance is achieved, several methods exist to estimate the ACE of the treatment [Schafer and Kang (2008)]. We applied the inverse propensity weighted regression adjustment. In this method, a model for the safety outcomes (henceforth referred to as *the outcome model*) under both treatment and control application is estimated<sup>8</sup> using weighted regression; the weights come from the estimated propensity scores. We again make use of GBM to estimate a single regression model by using an indicator for the treatment. All covariates listed in Figure 1 are included in the model, and we choose the interactions implicitly. Overfitting is avoided by using 10-fold cross-validation and out-of-bag

---

<sup>7</sup>These correspond to the population weights (ATE weights in the twang package).

<sup>8</sup>One could also specify two separate outcome models, one for the outcome under treatment and one for the outcome under control. We follow this approach in the matching implementation, which is described in the supplementary material [Karwa, Slavković and Donnell (2011)].

estimation [Ridgeway (2007)]. The GBM model of the outcomes is used to predict the probability distributions of  $S_i(1)$  and  $S_i(0)$ .

Once the model for the outcomes is specified, we used two approaches to estimate the ACE that is the causal risk ratio of equation (4.1). The standard errors for these estimates were obtained by bootstrapping. In the *combined prediction* approach, the complete data are used to predict the outcome under treatment application using the outcome model, and the complete data are used to predict the outcome under control application. In the *individual prediction* approach, only treated data are used to predict the outcome under treatment application, and only control data are used to predict the outcome under the control application. In other words, in the combined method, the complete data are used to estimate the potential outcomes under treatment and control application, irrespective of the actual assignment,<sup>9</sup> whereas in the individual method the potential outcomes under treatment and control are estimated by using that part of the data which actually received the treatment and control assignment, respectively. In both methods, the final ACE is estimated as the ratio of expected outcome under treatment and expected outcome under control [cf. equation (4.1)]. The results of the two methods and their comparisons are presented in the next section.

**4.3. Results.** This section presents the results for each of the three comparisons in terms of achieved balance and estimates of ACE.

*Balance of key covariates.* Figure 1 summarizes the effect of weighting on achieving the balance; more detailed statistics are provided in the supplementary materials [Karwa, Slavković and Donnell (2011)]. The graph shows the  $p$ -value for the Kolmogorov–Smirnov test statistic before and after the weighting of key covariates for each of the three comparisons. Figure 1(a) shows that there was a considerable improvement in the balance after IPW for the Medium to Low comparison. However, the variables Truck, Terrain, ADT and Right shoulder width remain unbalanced even after weighting. Terrain was also unbalanced in the High to Medium comparison; see Figure 1(b). Figure 1(c) shows that all covariates seem reasonably well balanced for the High to Low comparison.

*Estimation of ACE.* The estimates of ACE for each of the comparisons are shown in Table 3 for both individual and combined predictions. For the High to Low comparison, the results from both prediction methods suggest that application of High *PMR* significantly reduces the risk of a target crash in comparison to application of Low *PMR*. Based on the results from both groups, the risk of a target crash on segments with Low *PMR* is 3.09 times that of High *PMR*. Furthermore, we can be 95 percent confident that the expected risk of a crash on segments with low *PMR* is between 2.22 and 4.52 times that on segments with application

---

<sup>9</sup>It must be noted that the combined prediction is similar in spirit to the “do” operator, which will be introduced in the Causal Diagrams section; see also the discussion section. Also, the individual prediction can be considered to be an “analytical simulation” of a randomized experiment.

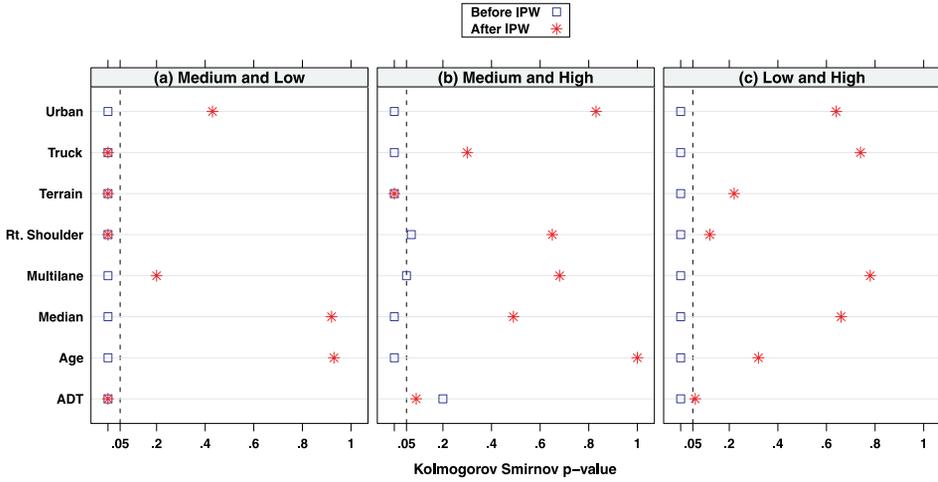


FIG. 1. Balance of covariates before and after weighting for (a) Treatment = Medium and Control = Low, (b) Treatment = Medium and Control = High, and (c) Treatment = Low and Control = High. The vertical dashed line indicates the cutoff p-value at 0.05 level.

of High *PMR*. Similarly, based on the results from the prediction from individual groups, the risk of a target crash on segments with Low *PMR* is 2.53 times that of High *PMR* with the 95 percent confidence interval [2.31, 3.60]. The ACE point estimates from the two methods are quite different, but there is considerable overlap in the confidence intervals, with the interval from the combined prediction being slightly wider than from the individual prediction. This is due to model-based extrapolation in the combined method.

The results for High to Med comparison, from both methods, also suggest that application of High *PMR* significantly reduces the risk of a target crash in comparison to application of Med *PMR*, but the expected risk is smaller in magnitude than for High to Low *PMR* comparisons. The point estimates of the ACE from two methods are close to each other with considerable overlap in the confidence intervals: combined ACE of 1.88 (95% CI: 1.5, 2.02) and individual ACE of 1.82 (95% CI: 1.67, 1.97).

TABLE 3  
Estimate of ACE based on PO framework

Change in visibility level	Individual prediction		Combined prediction	
	Point estimate	95% limits	Point estimate	95% limits
High to Low	2.53	[2.31, 3.60]	3.09	[2.22, 4.52]
Medium to Low	1.21	[0.9, 1.4]	1.35	[0.8, 1.81]
High to Medium	1.82	[1.67, 1.97]	1.88	[1.5, 2.02]

For the Med to Low comparison, the results indicate that there may be no significant effect of changing the PMR level from Low to Med. The point estimates of the ACE and the confidence intervals from the two methods are close to each other. From the combined prediction, the risk of a target crash on segments with Low *PMR* is 1.35 (95% CI [0.8, 1.81]) times that of Med *PMR*. From the individual prediction, the risk of a target crash on segments with Low *PMR* is 1.21 (95% CI [0.9, 1.4]) times that of Med *PMR*. However, it must be noted that this was the most difficult data subsample to attain the balance on the covariates, and the risk ratios may be biased, due to lack of balance over important key covariates.

**5. Causal diagrams.** In this section we present the Causal Diagrams (CD) framework to estimate ACEs. We use discrete Causal Bayesian networks (CBN) to implement the CD framework as described in Section 5.1. Section 5.2 briefly discusses the algorithms used to recover a CBN from observational data, the required assumptions and estimation procedures for the ACE. Section 5.3 presents the results of the analysis.

5.1. *Causal diagrams and components of a causal model.* In the CD setting, a causal model is used as the fundamental element to estimate causal effects, in contrast with the PO model, where potential outcomes are the fundamental quantities. Let  $V$  denote the set of variables representing the attributes of a road segment which includes both the treatment assigned to a segment and its safety outcome. A Causal Model describes the causal relations (in the form of conditional independence) among the variables in  $V$ . The qualitative part of the model is represented by a graph using a set of nodes and edges, and the quantitative part by a set of conditional probability distributions associated with each node in the graph.

In our analysis, we represent the Causal Model by using a discrete Causal Bayesian Network (CBN) for implementing the CD framework.<sup>10</sup> A CBN consists of a directed acyclic graph (DAG) and a set of probability distributions associated with each node, represented by a conditional probability table (CPT). Figure 2 shows an example of such a graph, where, for instance, Safety ( $S$ ) is a *child* ( $ch$ ) of  $ADT$  and  $PMR$ , and, thus, these are its *parents* ( $pa$ ). For more details on graphs and graphical models, see Lauritzen (1999). The discrete versions of the variables as defined in Table 2 were used.

The problem of causal inference involves learning the causal structure, represented by a DAG and a CPT, from data. The ACE of a treatment under intervention is estimated using *intervention theory*, as explained in the next section.

---

<sup>10</sup>CBNs with continuous variables are possible, but algorithms for handling arbitrary continuous distributions are not well developed. Also, many algorithms cannot handle mixed BNs (mixed here refers to the combination of continuous and discrete variables) that have continuous parents of discrete children.

5.1.1. *Causal diagrams as models of intervention.* According to Pearl (2000), CBNs can be regarded as models of interventions if it is assumed that a DAG models the causal mechanism which generated the data. (See Section 5.2 for a review of this assumption.)

Under the above assumption, the edges in a DAG are used to specify the changes in the joint distribution of variables  $V$  due to external intervention. For instance, in Figure 2, forcing the node  $PMR$  to take a particular value, say,  $Low$ , amounts to lifting the existing mechanism on  $PMR$  and putting it under the influence of a new mechanism whose action is to force  $PMR$  to the value  $Low$ , keeping everything else constant. This action is mathematically represented by  $do(PMR = Low)$ . The effect of “setting” a node to a fixed value corresponds to applying the low  $PMR$  treatment to all the segments in the sample. Such interventions are modeled in a DAG  $G$  by creating a new mutilated DAG  $G_{PMR}$  from  $G$ . In  $G_{PMR}$ , the links between  $PMR$  and its parents are removed, keeping the rest of the graph the same. The distribution imposed by the new graph  $G_{PMR}$  under the condition  $PMR = Low$  represents the effect of intervention and is called the post-intervention distribution; for example, see Figure 2.

5.1.2. *Causal effect and ACE.* Given the safety outcome  $S$  and the treatment variable  $PMR$ , the causal effect of  $PMR$  on  $S$ , denoted by  $P[S|do(PMR = i)]$ , where  $i \in \{Low, Med, High\}$ , is a function from  $PMR$  to the space of probability distribution on  $S$ . For each realization of  $PMR$ ,  $P[S|do(PMR = i)]$  gives the probability of  $S = s$  induced from the mutilated graph  $G_{PMR}$  and substituting the value of  $PMR$  as  $i$  in this graph.

Given a causal diagram in which all the parents of manipulated variables are observed, the causal effect can be estimated from passive or noninterventional data. However, when some parents of a child node  $ch$  are not observed,  $P(ch|pa_i)$  may not be estimable in all cases. A graphical test has been provided by Pearl (2000), Chapter 3, to find out when  $P[S|do(PMR)]$  is estimable from the observed data. In the present case, we make the assumption that all potential confounders are included in the analysis. This is a strong assumption and must come from subject matter experts. These assumptions are reviewed in Section 5.2.

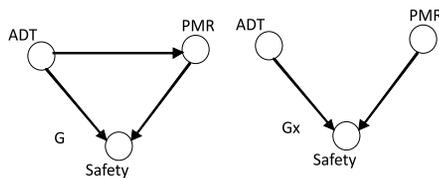


FIG. 2. An example of the interventional distribution. The graph  $G$  represents the original DAG. The mutilated graph  $G_x$  under the intervention of forcing  $PMR$  to take a particular value is obtained by deleting the arcs between  $PMR$  and its parents (e.g.,  $ADT$ ).

When  $PMR$  has two possible states ( $Low$  and  $Med$ ), the ACE is given by the following equation:

$$(5.1) \quad ACE_{Med \text{ to } Low} = \frac{P(S = 1 | do(PMR = Low))}{P(S = 1 | do(PMR = Med))},$$

where  $P[S = 1 | do(PMR = Low)]$  is the marginal probability in  $G_{PMR}$ <sup>11</sup> of  $S = 1$  under the intervention  $PMR = Low$ ; similar expressions are used for the other two comparisons,  $High$  to  $Low$  and  $High$  to  $Med$ . Also, notice the similarity to equation (4.1) from the PO framework. In the next section we describe the algorithms that were used to learn the components of the CBN.

5.2. *Learning CBNs from data and estimation of ACE.* Structure learning algorithms are used to recover a DAG  $G$  and parameter learning algorithms are used to estimate the CPTs, which then lead to estimation of ACE.

5.2.1. *Structure learning.* Learning the structure of causal networks from observational data has received a thorough treatment in the literature; see Pearl and Verma (1991), Heckerman (2008), Spirtes and Glymour (1991). The most common strategies fall into two different classes called *constraint based learning* and *score based learning*. We adopt a simple combination of both approaches to learn the structure of the CBN. Our approach is similar in principle to Tsamardinos, Brown and Aliferis (2006). The PC algorithm [Spirtes, Glymour and Scheines (2001)] is used as a constraint based strategy to recover a DAG from the data. This DAG is supplied as an initial input to the score based learning strategy, which then attempts to find an optimum DAG. The simulated annealing strategy of Hartemink (2005) and the scoring function proposed by Heckerman, Geiger and Chickering (1995) are used to search for optimum scored CBN. The scoring search is implemented in Java using the BANJO library [Hartemink (2005)] and the constraint search is performed using the BNT toolbox in Matlab [Murphy (2001)]. Since the scoring method need not produce the globally optimum structure of the CBN, we used the 10 best networks recovered by the algorithm and performed Bayesian model averaging to estimate the ACE. For details on the averaging, refer to Madigan and Raftery (1994), Hoeting, Adrian and Volinsky (1998) and Heckerman, Geiger and Chickering (1995).

Irrespective of the strategy used, a DAG can be recovered from observational data, up to  $d$ -separation equivalence [Pearl (2000), Chapter 1], only if the three assumptions outlined below are satisfied. Causal interpretation of CBN is possible because of these assumptions, which are in general untestable from observational data and must come from subject matter experts.

---

<sup>11</sup>Conditioning on so-called colliders can actually introduce bias in the ACE; see Pearl (2003). In simple terms, a variable is a collider if it has two arrows into it. In the present case, there are no such colliders on the path between Safety and  $PMR$ .

*Causal Markov Assumption:* The Causal Markov Assumption (CMA) states that given the values of a variable's immediate causes (i.e., its parents), the variable is independent of its nondescendants [Pearl (2000), Chapter 1]. This assumption implies that we must include in the model every variable that is a cause of two or more other variables. It also implies Reichenbach's [Reichenbach (1956)] common cause assumption, which states that, if any two variables are dependent, then one is a cause of the other or there is a third variable causing both.

The natural question that arises is what are the immediate factors that affect the safety of a segment? For instance, is driving at a high speed considered an immediate cause of reduced safety? To understand the CMA in light of safety, we need to consider factors that can cause a crash. These factors can be divided into three broad categories: road user (driver), the vehicle, and roadway characteristics (environmental conditions, roadway volume, etc.). Generally, information on the factors related to drivers and the vehicle is available only for vehicles involved in a crash, and not for noncrash vehicles. Thus, in a driver level analysis, most of the data would be missing. Also, the immediate causes of crashes (75 percent of which are due to human error [Stanton and Salmon (2009)]) become very specific to a particular crash and are governed by complex human behavior which is difficult to model and predict. To avoid these issues, analysis is done at the segment level. Only stable attributes of a roadway segment are included in the analysis; specific human factors are included in the error terms considered to be stochastic in nature. Thus, CMA is treated as a guiding principle rather than an assumption, where it defines the granularity of the model being considered, ensuring that all relevant causes, as defined by subject matter experts and past experiments, are included in the analysis.

*Faithfulness:* The faithfulness assumption ensures that the population that generated the DAG has exactly those independence relations specified by the DAG structure and no additional independencies. If there are any independence relations in the population that are not a consequence of the Causal Markov condition, then the population is unfaithful. By assuming Faithfulness, we eliminate all such cases from consideration.<sup>12</sup>

*Latent variables:* This assumption states that there are no hidden variables in the model that violate the causal Markov condition. That is, all of the variables that effect more than two variables in the model are observed and included in the database. Again, this is a strong assumption, whose validity could be ensured by verification from subject matter experts. For instance, the definition of safety ensures that the causes due to driver and weather factors do not influence the outcome, or else these would have to be entered into the model as latent variables.

---

<sup>12</sup>This assumption is controversial; see the discussion section.

5.2.2. *Parameter learning.* The parameters of the CPT are modeled using *Dirichlet distributions* and the usual assumptions of parameter independence are made. For details on parameter learning, see Heckerman, Geiger and Chickering (1995). The Bayesian Dirichlet Equivalent Uniform Priors (BDEU) were used to compute the parameters of the CBN.

The Dirichlet hyper parameters  $\alpha_{x_i, \pi_i}$  are specified by the following equation:

$$(5.2) \quad \alpha_{x_i, \pi_i} = \alpha \times p(x_i, \pi_i),$$

where  $\alpha_{x_i, \pi_i}$  pertains to variable  $X_i$  in a state  $x_i$  given that its parents are in joint state  $\pi_i$ , for  $i = 1, \dots, n$ , where  $\alpha$  is the number of pseudo-counts, and  $p$  is a (marginal) prior distribution of pseudo-counts; this ensures the likelihood-equivalence of Markov equivalent structures [Heckerman, Geiger and Chickering (1995)]. The value of  $\alpha$  is taken to be 1. The distribution  $p$  is chosen to be uniform between 0 and 1 for all variables (representing noninformative prior), that is, for any CPT, each parent-child combination is given an equal probability.

5.2.3. *Estimation of ACE.* There are several methods in the literature [Cowell (1998)] to efficiently perform inference in a CBN. We computed the marginal probability of  $S$  by using the *junction tree algorithm* that performs exact inference. Recall that the ACE of  $PMR$  on  $S$  is estimated as the ratio of the expected value of safety under the intervention level corresponding to the treatment and the expected value of safety under the intervention level corresponding to control [cf. equation (5.1)]. A full Bayes model was specified and the confidence in the value of the ACE was estimated by computing a 95 percent Bayesian credible interval.

5.3. *Results.* The DAG with the highest score is shown in Figure 3. Notice that there is no *Low speed* variable in this DAG. This could be because given the combination of variables like *ADT*, *Median* and *Multilane*, the value of *Low speed* is completely determined, and the discretization of *ADT* into two levels makes it highly collinear with *Low speed*. It was surprising to see the safety of a segment directly unaffected by *ADT* in this particular DAG, since it is commonly observed that the higher the *ADT*, the higher the probability of a target crash on a segment. However, two of the top 10 graphs show that *ADT* does indeed affect safety. A possible reason could again be the discretized *ADT* variable, which is also highly correlated with the *Multilane* variable; segments with more than two lanes generally have high *ADT*. Similar problems were encountered in the PO framework. This could be the reason why the *Multilane* indicator affects safety in 8 out of the 10 highest scoring models.

Figure 4 shows the mutilated DAG used to model the effect of intervention on the  $PMR$  levels. As noted earlier, the mutilated DAG is formed by deleting all the edges from the original DAG that direct into the  $PMR$  variable, and fixing the value of  $PMR$  at a particular level. The marginal probability distribution of safety in such a DAG represents the effect of manipulating the  $PMR$  variable on  $S$ . We computed

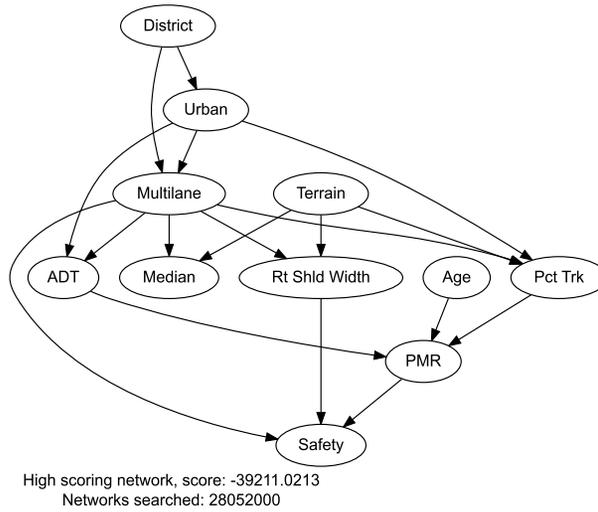


FIG. 3. The best scoring DAG recovered by the search algorithm.

the full Bayesian posterior of ACE using Monte Carlo simulation, averaging over the 10 best selected networks.

Table 4 shows the final results of the effect of *PMR* on safety, computed via the Causal Diagrams approach. The results suggest that higher *PMR* levels correspond to significantly lower risk of a target crash on all comparisons. In particular, the risk of a target crash on a segment with Low *PMR* is 3.12 times that of High *PMR* with a 95% CI of [2.32, 4.11] and 1.79 times that of Med *PMR* with a 95% CI of [1.31, 2.28]. Similarly, the risk of a target crash on a segment with Med *PMR* is 1.86 times that of High *PMR* with a 95% CI of [1.60, 2.17].

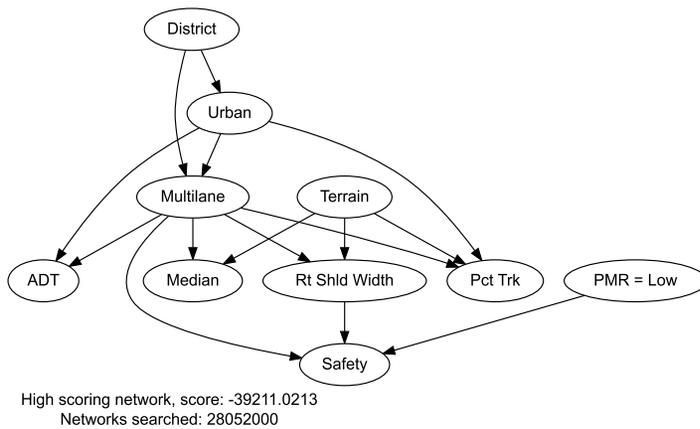


FIG. 4. The mutilated graph under the manipulation of *PMR* = Low.

TABLE 4  
Estimate of ACE based on CD framework

Change in visibility level	Point estimate	95% limits
High to Low	3.12	[2.32, 4.11]
Medium to Low	1.79	[1.31, 2.28]
High to Medium	1.86	[1.60, 2.17]

**6. Comparison of results and discussion.** This section compares and discusses the analyses and results from the PO and CD frameworks. In addition, it also includes the results from the popular PO alternative to IPW, the propensity score matching, whose implementation details are available in the supplementary document [Karwa, Slavković and Donnell (2011)].

6.1. *Comparison of results and implications.* Figure 5 shows the point estimates and confidence intervals of ACEs from both frameworks. Specifically, it shows the overlap among the estimates of ACEs from the following methods: Combined and Individual approaches, using Inverse Propensity Score weighting (IPW) and using Propensity Score Matching (M) with regression adjustment from the PO framework, and Causal Bayesian Networks (CBN) from the CD framework. There are a couple of important points to be noted.

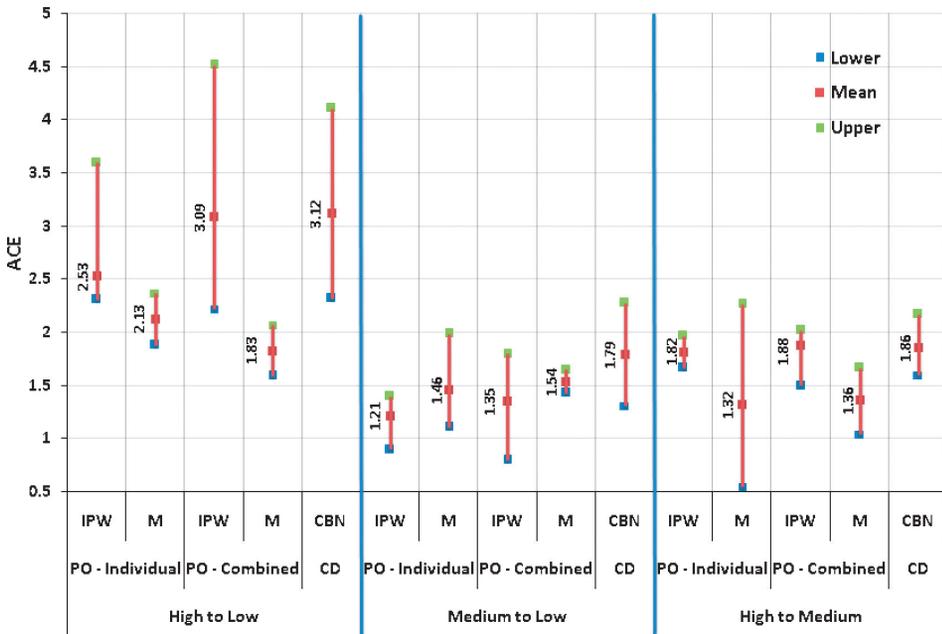


FIG. 5. Overlap between the confidence intervals of ACE from both frameworks.

In terms of a general trend, the results of all methods are consistent with each other; they show that increased *PMR* levels generally lead to a reduced or unchanged risk of a target crash which agrees with engineering expectations. In terms of magnitude of the effect, however, there is a noticeable and in some cases statistically significant difference across different implementations. The CBN point estimates are consistently higher than both the IPW and M point estimates.

In general, the 95% confidence regions from the three methods show significant overlap across the three comparisons, and thus lead to the same conclusions in terms of the expected strength of the causal effect. The most consistent result, statistically and with respect to transportation engineering logic, is for the High to Low comparison where all three methods indicate that there is a significant reduction in risk with application of High *PMR* in comparison to Low *PMR* even though the matching results are significantly lower in magnitude, in particular, for the PO-combined M estimate. The IPW and CBN results display strong correspondence with the exception of Med to Low comparison. The inference based on IPW results implies that the risk of a target crash is not significantly lower on segments with the application of Med *PMR* levels when compared to Low *PMR* levels, whereas both CBN and M based confidence intervals imply statistically significant risk reduction and are more in line with engineering intuition.

The most curious result with respect to engineering expectation occurs for the High to Med comparison. Here, it was expected that there would be a small or no statistically significant effect. Only the matching results support this assertion, and, in particular, only the PO-individual M result claims no effect. The PO-combined M estimates are closer to IPW and CBN results, with the latter two having a very strong correspondence that implies a significant reduction in safety risk with the application of High *PMR* compared to Med *PMR*. It should be noted, however, that the PO-individual M estimate has the highest variability, with a significant CI overlap with other estimated effects, thus, it may be difficult to render a solid practical decision based solely on statistical significance or lack thereof. Alternatively, one can argue that PO-combined, IPW and the CBN estimates are based on extrapolation when compared to the PO-individual M estimates due to differences in data subsamples being used. Further examination of the results is needed, which is discussed below.

*6.2. Discussion of results.* There could be several reasons for the differences in the results from the PO framework and the CD framework. One is distributional support. The variables used in the CBN setting were discretized to ensure the use of efficient algorithms. On the other hand, there was little discretization performed in the PO model. Specifically, the variable “age” was used as a continuous variable

in the IPW estimation, and “age,” “Rt. Shoulder width,” “Percentage Trucks” and “ADT” were used as continuous variables in the Matching estimation.<sup>13</sup>

As noted, the CBN and combined-IPW results are close to each other. This is due to the similarity in empirical estimation of ACE. In both frameworks, causal effect is defined as a contrast between the outcomes of the *same* unit with and without treatment. However, to avoid the problem of missing potential outcomes, the individual IPW estimator compares *similar* units, whereas the combined estimator compares the *same* units. In the latter case, the problem of unobservable potential outcomes is avoided by using the outcome model for prediction. This is similar to applying the “do” operator to all units and contrasting the outcomes under different manipulations.

The differences between results of matching and results of CBN and IPW are due to differences in the data set, more specifically, due to loss of data in matching. To ensure overlap and balance, matching discards data. Dehejia and Wahba (1999) and Heckman, Ichimura and Todd (1998) point out that without overlap the results would be sensitive to the specification of either the propensity score or outcome model. While our matching procedure attained both balance and sufficient overlap for all comparisons, there was a significant loss of observations; see the supplementary document for details. For instance, for High to Med case, matching discards 83% of the data. Thus, the PO-individual M estimate of no significant change for High to Med may not be representative of the whole population. On the other hand, matching has the smallest loss of information for Med to Low case. The results from the CBN and matching show that there could be some significant effect in changing the *PMR* level from Med to Low, whereas the IPW results indicate that there may not be a significant effect. However, the IPW results may be biased because of poor balance achieved in the data for this comparison (cf. Figure 1).

Some of the differences between IPW and CBN could also be due to the use of Bayesian estimation in CBN. For instance, in CBN, Bayesian model averaging was used to account for the possibility of recovering a local minima. Moreover, all of the variables in CBN were discrete, which lead to large and sparse CPTs. The estimates from the PO framework were also affected by sparseness, although in part mitigated by the presence of continuous variables.

6.3. *Which method to use?* The choice of method is greatly influenced by the assumptions made. No method can be completely assumption free, in fact, all causal inferences (from observational data) must be based on causal assumptions. The major causal assumptions in each framework were the Causal Markov

---

<sup>13</sup>These choices were based on two (conflicting) policies, the first was to ensure similarity to the variables used in the CBN setting, and the second to ensure balance, the second one being given more priority.

Assumption (CMA) in CD and the unconfoundedness assumption in the PO framework. We conjecture that CMA is a stronger assumption than unconfoundedness, as it pertains to all the variables in the problem at hand, whereas the latter relates to the treatment variable and the potential outcomes. Furthermore, recent work has shown that unconfoundedness alone may not be sufficient to identify appropriate covariates for inclusion in the propensity score model; see Pearl (2000, 2009).

The PO framework requires overlap and balance assumptions, whereas the CD framework does not. In the broadest sense, the balance assumption ensures that there are no-pretreatment differences between the groups being compared, and the overlap assumption ensures that the estimates do not rely too much on the functional specification of the model. However, in the case of CD, the (qualitative) causal model is assumed to be known completely.<sup>14</sup> Moreover, the framework does not require any explicit functional specification of relation between the variables, given the CMA [and some additional assumptions; see Pearl (2000), Chapter 3, for technical details and mathematical proofs]. Based on this discussion, we suggest the following guidelines for deciding which method to use.

If little is known about the data generating process or the causal mechanism, the analyst should go as “nonparametric” as possible. For example, one could use matching (preferably by specifying the propensity score model using a nonparametric estimator such as GBM), ensure sufficient overlap, and compute average causal effect from the observed data by using individual prediction. However, this may depend on the specification of the propensity score model, and, more importantly, in the case of significant loss of data (as in our case), it is not clear as to which subpopulation the ACE estimates apply. It must also be noted that this strategy may not guard against the inclusion of inappropriate covariates in the propensity score model (such as colliders and bias amplifying covariates).

One could attempt to recover the data generating process from observational data under further assumptions of faithfulness, combined with partial expert knowledge. However, Robins and Wasserman (1999) show that when the probability that variables in the causal model have no common unobserved causes is small relative to the sample size, analysis carried out using faithfulness can lead to inappropriate conclusions.

On the other hand, if the data generating mechanism is known (even qualitatively), the mechanism can be summarized in the form of a causal diagram. The causal diagram may incorporate the mechanism related to treatment assignment and/or the response to the treatment. Such a causal model can be used as a guide to estimate the propensity score model as well as the outcome model (which can then be used with other adjustment methods such as weighting, matching, etc). The dependence on the functional form between variables can be reduced by using

---

<sup>14</sup>This may not be always the case, especially in the social sciences.

categorical variables<sup>15</sup> and/or by using nonparametric estimators such as GBM, as in our case.

In a real data setting, it is always better to compare the results from different methods. In the present study, it is clear that for different comparisons of PMR levels, different methods show consistency based on which assumptions are being violated. For instance, in the High to Low *PMR* case, all methods show good agreement. In the Med to Low case, Matching and CBN show good agreement (IPW results may be biased due to lack of balance). In the Med to High case, IPW and CBN show good agreement (Matching may be biased due to significant loss of data).

The CBN results could also be biased if the causal model recovered by the data is not close to the truth. However, there is evidence that the CBN may be less biased when compared to other methods. In the comparisons where the data are well balanced (High to Low and High to Med) and there is considerable overlap (High to Medium), the CBN results are in close agreement with the IPW, which indicates that the model may be close to the truth. Since the true ACE is unknown, the only test for validity of the results is by implicit agreement of results from different methods. If different methods provide the same answer, the answer must be close to the truth, or, in the worst case, all methods fail to capture the same aspect of the true model.

6.4. *Future work.* To obtain a better comparison of the methods, future studies should aim at using data from simulation. The true causal effect of the population would be known a priori, and the quality and size of data can be controlled. Other advances on this exploratory work can be made by using more complex causal modeling methods. For instance, discretization of the *PMR* treatment variable can be avoided by using a dose-response model [Hirano and Imbens (2004)]. In the current study, temporal/spatial correlations may exist, though evidence was not found in these data. The *PMR* treatment can be modeled as a time varying treatment [Lok et al. (2004)] to take into account such correlations. Specification of the assignment mechanism for the *PMR* treatment variable is convenient when compared to other possible countermeasures, such as roadway lighting. The assignment mechanism for lighting is generally influenced by factors such as local design policies, complaints from residents and may also be related to past crash history. Such assignment mechanisms may prove difficult to model and may require the use of latent variables. Also in the current study, we did not explicitly consider uncertainty in the imputed (using ANN) PMR levels and uncertainty in the measurement process; rather a mean estimate was used. Both of these are important issues that should be carefully considered as part of future work. Sampling zeros were encountered in both the PO model as well as the CBN setting. In the PO

---

<sup>15</sup>This may come with a set of its own problems, for example, sparseness.

framework, such sampling zeros created problems in achieving balance over interactions of covariates (specifically in the matching estimator). In the CBN setting, the use of Bayesian Inference in part addressed this problem. A similar approach in the PO framework would be to use a full Bayes model of both the propensity scores as well as safety outcomes [Rubin et al. (2008)]. These explorations are left to the scope of future work.

**7. Conclusion.** The examination of causal inference methods to transportation safety data reveals that there is considerable scope of their application to estimate safety effects of a countermeasure. A comparison was made between the PO framework and the CD framework. More specifically, the results based on three different implementations of these frameworks on a real data set were compared: Inverse Propensity Score Weighting with regression adjustment and Propensity Score Matching with regression adjustment versus Causal Bayesian Network.

Although the general trend of results seem to be consistent, we found that the magnitude of ACEs are sensitive to the method used and to the assumptions being violated. In real data sets, it is very likely that some assumptions will be violated. Depending upon which assumptions are appropriate, different methods should be used. Assumptions should be considered a priori. If possible, the analyst should run multiple implementations to compare the results for consistency. In conclusion, we suggest the use of the PO framework supplemented by a qualitative causal diagram as a rich framework to estimate the safety effects of countermeasures in transportation studies.

**Acknowledgments.** The authors are grateful to the Editor, Associate Editor and an anonymous referee whose comments have greatly improved the content of this work. We are also thankful to the North Carolina Department of Transportation for providing the data, and Dr. Kenneth Opiela from the Federal Highway Administration Office of Safety R&D for coordinating with the North Carolina Department of Transportation to obtain the data.

#### SUPPLEMENTARY MATERIAL

**Supplement to “Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams”** (DOI: 10.1214/10-AOAS440SUPP; .pdf). This document contains additional details about the Matching and Inverse Propensity score estimators and the top ten graphs recovered by the graph learning algorithm.

#### REFERENCES

- BAHAR, G., MASLIAH, M., ERWIN, T., TAN, E. and HAUER, E. (2006). Pavement marking materials and markers: Real-world relationship between retroreflectivity and safety over time. NCHRP Web Only Document 92, Transportation Research Board, National Research Council, Washington DC.

- BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973. MR2216189
- BUREAU OF TRANSPORTATION STATISTICS, U.S. (2007). National transportation statistics. Technical report.
- COLE, S. R. and FRANGAKIS, C. E. (2009). The consistency statement in causal inference: A definition or an assumption? *Epidemiology* **20** 3–5.
- COWELL, R. (1998). Introduction to inference for Bayesian networks. In *Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models* 9–26. Kluwer Academic, Norwell, MA.
- DAVIS, G. A. (2000). Accident reduction factors and causal inference in traffic safety studies: A review. *Accident Analysis & Prevention* **32** 95–109.
- DAVIS, G. A. (2004). Possible aggregation biases in road safety research and a mechanism approach to accident modeling. *Accident Analysis & Prevention* **36** 1119–1127.
- DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 1053–1062.
- DONNELL, E. T., KARWA, V. and SATHYANARAYANAN, S. (2009). Analysis of effects of pavement marking retroreflectivity on traffic crash frequency on highways in North Carolina. *Transportation Research Record: Journal of the Transportation Research Board* **2103** 50–60.
- EELLS, E. (1991). *Probabilistic Causality*. Cambridge Univ. Press, Cambridge. MR1120269
- FIENBERG, S. E. and SFER, A. M. (2006). Randomization, models, and the estimation of causal effects. Unpublished manuscript.
- HARTEMINK, A. J. (2005). Banjo: Bayesian network inference with Java objects. Software package, available at <http://www.cs.duke.edu/amink/software/banjo>.
- HECKERMAN, D. (2008). A tutorial on learning with Bayesian networks. In *Innovations in Bayesian Networks* (D. Holmes and L. Jain, eds.) 33–82. Springer, Berlin.
- HECKERMAN, D., GEIGER, D. and CHICKERING, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20** 197–243.
- HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1998). Matching as an econometric evaluation estimator. *Rev. Econom. Stud.* **65** 261–294. MR1623713
- HERNAN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60** 578–586.
- HIRANO, K. and IMBENS, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* **2** 259–278.
- HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (X.-L. Meng and A. Gelman, eds.) 73–84. Wiley, Chichester. MR2134803
- HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826
- HOETING, J., ADRIAN, D. M. and VOLINSKY, C. T. (1998). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models* 77–83. AAAI Press.
- HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–960. MR0867618
- HONG, G. and RAUDENBUSH, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis* **27** 205–224.
- KARWA, V. (2009). Safety effects of pavement marking retroreflectivity: An application of causal bayesian networks. Master's thesis, Pennsylvania State Univ., University Park, PA.
- KARWA, V. and DONNELL, E. T. (2011). Predicting pavement marking retroreflectivity degradation using artificial neural networks: An exploratory analysis. *Journal of Transportation Engineering* **137** 91–103.

- KARWA, V., SLAVKOVIĆ, A. and DONNELL, E. T. (2011). Supplement to “Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams.” DOI: 10.1214/10-AOAS440SUPP.
- LAURITZEN, S. L. (1999). Causal inference from graphical models. In *Complex Stochastic Systems* 63–107. Chapman & Hall/CRC Press, Boca Raton, FL. MR1893411
- LEE, B. K., LESSLER, J. and STUART, E. A. (2009). Improving propensity score weighting using machine learning. *Stat. Med.* **29** 337–346.
- LOK, J., GILL, R., VAN DER VAART, A. and ROBINS, J. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statist. Neerlandica* **58** 271–295. MR2157006
- MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam’s window. *J. Amer. Statist. Assoc.* **89** 1535–1546.
- MALDONADO, G. and GREENLAND, S. (2002). Estimating causal effects. *Int. J. Epidemiol.* **31** 422–429.
- MCCAFFREY, D., RIDGEWAY, G. and MORRAL, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* **9** 403–425.
- MURPHY, K. P. (2001). The Bayes net toolbox for MATLAB. *Comput. Sci. Statist.* **33** 2001.
- NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION (2007). Traffic safety facts 2007. Technical report.
- PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, Cambridge. MR1744773
- PEARL, J. (2003). Statistics and causal inference: A review. *Test* **12** 281–318. MR2044313
- PEARL, J. (2009). On a class of bias-amplifying covariates that endanger effect estimates. Technical report, Univ. California, Los Angeles.
- PEARL, J. and VERMA, T. S. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference* **11** 441–452. Morgan Kaufmann, San Mateo, CA. MR1142173
- REICHENBACH, H. (1956). *The Direction of Time*. Univ. California Press, Berkeley.
- RIDGEWAY, G. (2007). Generalized boosted models: A guide to the GBM package. Available at <http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf>.
- RIDGEWAY, G., MCCAFFREY, D. and MORRAL, A. (2006). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. RAND Corporation, Santa Monica, CA.
- ROBINS, J. M. and WASSERMAN, L. (1999). On the impossibility of inferring causation from association without background knowledge. In *Computation, Causation, and Discovery* (C. Glymour and G. F. Cooper, eds.) 305–321. AAAI Press, Menlo Park, CA. MR1689948
- ROSENBAUM, P. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152
- RUBIN, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.
- RUBIN, D. B. (1998). Estimation from nonrandomized treatment comparisons using subclassification on propensity scores. *Ann. Internal Medicine* **8** 757–763.
- RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071
- RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Statist.* **2** 808–840. MR2516795
- RUBIN, D. B., WANG, X., YIN, L. and ZELL, E. R. (2008). Bayesian causal inference: Approaches to estimating the effect of treating hospital type on cancer survival in Sweden using principal

- stratification. In *Handbook of Applied Bayesian Analysis* (T. O'Hagan and M. West, eds.). Oxford Univ. Press, Oxford.
- SCHAFFER, J. L. and KANG, J. (2008). Everage causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods* **13** 279–313.
- SFER, A. M. (2005). Randomization and causality. Technical report, Facultad de Ciencias Económicas, Universidad Nacional de Tucumán, San Miguel de Tucumán, Argentina.
- SPIRITES, P. and GLYMOUR, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* **9** 62–72.
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (2001). *Causation, Prediction, and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675
- STANTON, N. A. and SALMON, P. M. (2009). Human error applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science* **47** 227–237.
- TSAMARDINOS, I., BROWN, L. E. and ALIFERIS, C. F. (2006). The max–min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65** 31–78.

V. KARWA  
A. B. SLAVKOVIĆ  
DEPARTMENT OF STATISTICS  
PENNSYLVANIA STATE UNIVERSITY  
UNIVERSITY PARK, PENNSYLVANIA 16802  
USA  
E-MAIL: vishesh@psu.edu  
sesa@stat.psu.edu

E. T. DONNELL  
DEPARTMENT OF CIVIL AND  
ENVIRONMENTAL ENGINEERING  
PENNSYLVANIA STATE UNIVERSITY  
UNIVERSITY PARK, PENNSYLVANIA 16802  
USA  
E-MAIL: edonnell@enr.psu.edu