Observational Studies 5 (2019) 1-20 Submitted 11/18; Published 3/19

Double-Robust Estimation in Difference-in-Differences with an Application to Traffic Safety Evaluation

Fan Li **frank.li@duke.edu**
Department of Biostatistics and Bioinformatics
Duke University
Durham, North Carolina, 27705, USA

Fan Li **fli@stat.duke.edu**
Department of Statistical Science
Duke University
Durham, North Carolina, 27708, USA

Abstract

Difference-in-differences (DID) is a widely used approach for drawing causal inference from observational panel data. Two common estimation strategies for DID are outcome regression and propensity score weighting. In this paper, motivated by a real application in traffic safety research, we propose a new double-robust DID estimator that hybridizes regression and propensity score weighting. We particularly focus on the case of discrete outcomes. We show that the proposed double-robust estimator possesses the desirable large-sample robustness property. We conduct a simulation study to examine its finite-sample performance and compare with alternative methods. Our empirical results from a Pennsylvania Department of Transportation data suggest that rumble strips are marginally effective in reducing vehicle crashes.

Keywords: Causal inference; crash modification factor; difference-in-differences; doubly robust; propensity score; transportation safety research

1. Introduction

Difference-in-differences (DID) is a popular evaluation strategy used across a range of disciplines. It uses data with a time dimension to control for unobserved but fixed confounding, and identifies causal effects by contrasting the change in outcomes pre- and post-treatment, among the treated and control groups (Ashenfelter, 1978; Ashenfelter and Card, 1985). The most common DID setting is a before-after design, in which the treated and control units are genuinely comparable. For example, DID often exploits a policy shift that occurred in one region but not in an adjacent region (Card and Krueger, 1994). The key assumption of DID is *parallel trend*, that is, the counterfactual trend behavior of treatment and control groups, in the absence of treatment, is the same (Heckman et al., 1997).

The target causal estimand in DID is a version of the average treatment effect for the treated (ATT). Estimation of ATT in DID is traditionally tied with a fixed-effects outcome regression model (Angrist and Pischke, 2009). Though flexible, the regression method relies on strong assumptions such as homogenous and additive effects, and can be sensitive to model misspecification. Alternatively, Abadie (2005) proposed a semiparametric estimator

for DID based on inverse probability weighting (IPW) where only a model for the propensity score but not the outcome is required. The IPW estimator does not require assumptions on the outcome distribution, but may be inefficient compared to a correctly-specified outcome model. Outside the DID context, the double-robust (DR) method (Bang and Robins, 2005) that augments an IPW estimator by an outcome regression have received much attention in causal inference. A DR estimator is consistent if either the outcome model or the propensity score model, but not necessarily both, is correctly specified (Scharfstein et al., 1999), and it is semiparametrically efficient when both models are correctly specified (Robins et al., 1994, 1995; Robins and Ritov, 1997). However, most DR methods focus on the average treatment effect (ATE) estimand rather than ATT. In this paper, the intrinsic connection between Abadie's DID estimator and the IPW technique motivates us to devise a new double-robust DID estimator for ATT, and we show it possesses the desirable large-sample robustness property.

Our method is motivated from a real application to traffic safety research. Specifically, we wish to evaluate the impact of installing rumble strips—a low-cost traffic safety countermeasure—on vehicle crashes. Due to ethical and practical constraints with roadway safety experimentation, observational studies are routinely used for such evaluations. The state-of-the-art method in traffic safety evaluation—the Empirical Bayes (EB) approach—adopts a treatment-control before-after design (Hauer, 1997), where the crash outcomes in a number of comparable treated and control sites were recorded both before and after the safety countermeasure was installed. This design fits naturally into the DID framework, but to our knowledge, the connection was never made in the literature. In fact, the EB approach is entirely regression-based and comes without a causal interpretation. As a robust alternative to EB, a recent stream of research advocated propensity score methods to the after period data alone (Karwa et al., 2011; Wood et al., 015a,b; Wood and Donnell, 2016). However, ignoring the data in the before period may present a major information loss and fail to adjust for the time trend. In contrast, our proposed double-robust estimator combines the virtues of the regression-based and the propensity score weighting estimators for before-after studies. Because the outcome is count data in the transportation application, we particularly focus on the case of discrete outcomes in our modeling and estimation.

The rest of the paper is organized as follows. Section 2 defines the causal estimands, and introduces DID estimators: outcome regression, propensity score weighting and the proposed double-robust estimators. Section 4 presents the application to highway crash data collected by the Pennsylvania Department of Transportation. Section 3 further illustrates the DID estimators through simulations mimicking the traffic safety study. Section 5 concludes.

2. Causal Inference via Difference-in-Differences

2.1 Causal Estimands

We introduce the notation in the context of the evaluation of rumble strips (i.e., treatment). We consider the basic two-period two-group DID design. Assume a sample of traffic sites—units of analysis—indexed by $i = 1, \ldots, N$, belong to one of the two groups, with $G_i = 1$ indicating that rumble strips were applied in the after period, i.e. the treatment group, and $G_i = 0$ indicating that rumble strips were not applied in either period, i.e., the control group.

Units in both groups are followed in two periods of time, with $T = t$ and $T = t+1$ denoting the before and after period, respectively. For each unit i, let D_{iT} be the observed treatment status at period T. Since none of the traffic sites received treatment in the before period, we have $D_{it} = 0$ for all i. Because the treatment is only administered to one group ($G_i = 1$) in the after period, $D_{i,t+1} = 1$ for all units in group $G_i = 1$ and $G_i = D_{i,t+1}$ for all i. Similar to prior traffic safety evaluation studies (Karwa et al., 2011; Wood and Donnell, 2017), we make the Stable Unit Treatment Value Assumption (SUTVA), meaning no interference between units and no different versions of the treatment. This assumption is more reasonable when the traffic sites are far apart from one another, but may be questionable when the sites are in close proximity. We will proceed with this assumption and discuss in Section 5 the implications when SUTVA is violated. Under SUTVA, each unit has two potential crash counts in each period, $Y_{iT}(0)$ and $Y_{iT}(1)$, and only the one corresponding to the observed treatment status, $Y_{iT} = Y_{iT}(D_{iT})$, is observed. The DID design allows us to write $Y_{it} = Y_{it}(0)$ and $Y_{i,t+1} = (1 - G_i)Y_{i,t+1}(0) + G_i Y_{i,t+1}(1)$. A vector of p pre-treatment variables, $\mathbf{X}_i$, is also observed for each unit. We denote the collection of observed data by $\mathcal{Z}_i = \{Y_{it}, Y_{i,t+1}, G_i, \mathbf{X}_i\}$, and assume that the $\mathcal{Z}_i$'s are independent and identically distributed from some common distribution $\mathbb{F}(\mathcal{Z})$.

In traffic safety studies, safety countermeasures are usually applied only to selected pilot sites before rolling out to a larger scale. The safety effectiveness is usually evaluated in a multiplicative fashion using the Crash Modification Factor (CMF, AASHTO, 2010), which can then be used to understand the expected change in crash frequency after a traffic safety countermeasure is implemented. In our traffic application, the rumble strip installation is implemented as part of a national safety improvement program, and the interest lies in quantifying its potential effectiveness among the sites where the rumble strips were installed. Similar to Wood and Donnell (2017), we formally define the CMF as a causal estimand that characterizes the ratio between the expected observed outcome after the installation and the expected counterfactual outcome had the countermeasure not been installed in the pilot sites. Using the potential outcome notation, we define the CMF

$$\tau_{\text{CMF}} \equiv \frac{\mathbb{E}[Y_{i,t+1}(1)|G_i = 1]}{\mathbb{E}[Y_{i,t+1}(0)|G_i = 1]} = \theta_1/\theta_0, \tag{1}$$

where we denote $\theta_1 = \mathbb{E}[Y_{i,t+1}(1)|G_i = 1]$ and $\theta_0 = \mathbb{E}[Y_{i,t+1}(1)|G_i = 0]$. Because the crash outcomes are count data, τ_{CMF} is a causal rate ratio that quantifies the relative average change in crash counts due to rumble strip installation among the treated. The scale-free τ_{CMF} is a ratio version of the usual average treatment effect among the treated (ATT) estimand. Here, to characterize the causal rate difference in the absolute scale, we also define an additive version—the Crash Frequency Difference (CFD):

$$\tau_{\text{CFD}} \equiv \mathbb{E}[Y_{i,t+1}(1) - Y_{i,t+1}(0)|G_i = 1] = \theta_1 - \theta_0. \tag{2}$$

We argue that using the pair of parameters $(\tau_{\text{CFD}}, \tau_{\text{CMF}})$ instead of τ_{CMF} alone presents a more complete picture of the effectiveness of safety countermeasure.

2.2 Assumptions

Estimands τ_{CFD} and τ_{CMF} are functions of θ_1 and θ_0. Under SUTVA, θ_1 is nonparametrically identified: $\theta_1 = \mathbb{E}[Y_{i,t+1}|G_i = 1]$, with a consistent moment estimator

$$\hat{\theta}_1 = \sum_{i=1}^{N} G_i Y_{i,t+1} \Big/ \sum_{i=1}^{N} G_i. \tag{3}$$

In contrast, θ_0—the expected counterfactual outcome in the absence of treatment at time $T = t+1$—must rely on additional restrictions to identify. Following the convention in DID design, we impose the parallel trend assumption,

Assumption **1** *(Parallel Trend) For each unit $i = 1, \ldots, N$,*

$$\mathbb{E}[Y_{i,t+1}(0) - Y_{it}(0)|\mathbf{X}_i, G_i = 1] = \mathbb{E}[Y_{i,t+1}(0) - Y_{it}(0)|\mathbf{X}_i, G_i = 0].$$

Assumption 1 imposes that, conditional on the pre-treatment covariates $\mathbf{X}_i$, the average outcomes in the treated and control groups, in the absence of treatment, would have followed a parallel path over time. The quantity θ_0 is therefore identified under Assumption 1 as

$$\begin{aligned}
\theta_0 =& \mathbb{E}_{\mathbf{X}}\{\mathbb{E}[Y_{i,t+1}(0)|\mathbf{X}_i, G_i = 1]|G_i = 1\} \\
=& \mathbb{E}_{\mathbf{X}}\{\mathbb{E}[Y_{it}(0)|\mathbf{X}_i, G_i = 1] + \mathbb{E}[Y_{i,t+1}(0) - Y_{it}(0)|\mathbf{X}_i, G_i = 0]|G_i = 1\} \\
=& \mathbb{E}[Y_{it}|G_i = 1] + \mathbb{E}_{\mathbf{X}}\{\mathbb{E}[Y_{i,t+1} - Y_{it}|\mathbf{X}_i, G_i = 0]|G_i = 1\},
\end{aligned} \tag{4}$$

where both terms of the right hand side of the equation involve only expectations of observed data and are identified.

It is important to note that a direct DID estimator that uses

$$\hat{\theta}_0^{\text{direct}} = \frac{\sum_{i=1}^{N} G_i Y_{it}}{\sum_{i=1}^{N} G_i} + \frac{\sum_{i=1}^{N}(1 - G_i)(Y_{i,t+1} - Y_{it})}{\sum_{i=1}^{N}(1 - G_i)}. \tag{5}$$

to estimate θ_0 neglects the pre-treatment covariate information and is subject to selection bias. In fact, $\hat{\theta}_0^{\text{direct}}$ is only consistent to θ_0 under the unconditional version of the parallel trend assumption, i.e., $\mathbb{E}[Y_{i,t+1}(0) - Y_{it}(0)|G_i = 1] = \mathbb{E}[Y_{i,t+1}(0) - Y_{it}(0)|G_i = 0]$, which is arguably stronger than Assumption 1. On the other hand, unlike the standard unconfoundedness condition usually assumed for the cross-sectional data, Assumption 1 does not necessarily assume that $\mathbf{X}$ controls for all sources of confounding. Indeed, DID allows for unobserved confounders to affect treatment assignment as long as their impact on the potential outcomes is both separable and time-invariant (Lechner, 2011). Assumption 1 is generally untestable and may be questionable in practice. As an indirect way to assess the plausibility of parallel trend, in this application, we will conduct a "no treatment" evaluation by performing DID analyses for crash outcomes from two pre-treatment periods ($T = t - 1$ and $T = t$). Specifically, if the parallel trend assumption is plausible, that is,

$$\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|\mathbf{X}_i, G_i = 1] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|\mathbf{X}_i, G_i = 0],$$

then the estimated CFD and CMF based on time $T = t - 1, t$ should be close to 0 and 1, respectively, because in reality rumble strips were not applied until after time t and should

have no causal effect for the pre-treatment outcomes. This idea is similar to the falsification endpoints or negative control idea in assessing unconfoundedness (Rosenbaum, 2002).

As in most ATT estimation, we also assume *weak overlap*, that is, each unit has a nonzero probability of receiving the control, $e(\mathbf{X}_i) \equiv \Pr(G_i = 1|\mathbf{X}_i) < 1$, where $e(\mathbf{X}_i)$ is the propensity score. The weak overlap assumption is directly testable by visually comparing the estimated propensity score distributions between the treatment groups.

2.3 Extant Methods: Regression and Weighting

Two main classes of existing estimating methods of DID are outcome regression and propensity score weighting. We first introduce a regression-based estimator specifically for count outcomes. To identify θ_0, we need to identify all components on the right hand side of equation (4). Similar to $\hat{\theta}_1$, the first term $\mathbb{E}[Y_{it}|G_i = 1]$ in θ_0 can be consistently estimated by a moment estimator, $\sum_{i=1}^N G_i Y_{it} / \sum_{i=1}^N G_i$. The second term in θ_0 requires a regression model for the difference in crash counts $Y_{i,t+1} - Y_{it}$ given $\mathbf{X}_i$ among the control sites. Given that a regression model for the difference in counts is difficult to obtain, we separately assume a negative binomial model for each of the cross-sectional counts

$$\begin{aligned} \{Y_{it}(0)|\mathbf{X}_i, G_i = 0\} &\sim \mathrm{NB}(\mu(\mathbf{X}_i;\boldsymbol{\beta}), \phi), \\ \{Y_{i,t+1}(0)|\mathbf{X}_i, G_i = 0\} &\sim \mathrm{NB}(\nu(\mathbf{X}_i;\boldsymbol{\gamma}), \psi), \end{aligned} \tag{6}$$

where μ, ν are known smooth mean functions with parameter $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and the variances are $\mathbb{V}(Y_{it}(0)|\mathbf{X}_i, G_i = 0) = \mu(\mathbf{X}_i;\boldsymbol{\beta}) + \mu^2(\mathbf{X}_i;\boldsymbol{\beta})/\phi$ and $\mathbb{V}(Y_{i,t+1}(0)|\mathbf{X}_i, G_i = 0) = \nu(\mathbf{X}_i;\boldsymbol{\gamma}) + \nu^2(\mathbf{X}_i;\boldsymbol{\gamma})/\psi$, with potentially different dispersion parameters ϕ and ψ. Model (6) is called the crash frequency model in traffic safety research (AASHTO, 2010). When the dispersion parameters approach infinity, model (6) reduces to Poisson regression. As is evident from equation (4), the crash frequency model is only required for the control group, but not the treatment group. We obtain the maximum likelihood estimates of the parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ using the control sites data. Under SUTVA and Assumption 1, equation (4) suggests the following estimator for θ_0,

$$\hat{\theta}_0^{\mathrm{reg}} = \frac{\sum_{i=1}^N G_i Y_{it}}{\sum_{i=1}^N G_i} + \frac{\sum_{i=1}^N G_i \{\nu(\mathbf{X}_i;\hat{\boldsymbol{\gamma}}) - \mu(\mathbf{X}_i;\hat{\boldsymbol{\beta}})\}}{\sum_{i=1}^N G_i}. \tag{7}$$

When the crash frequency model (6) is correctly specified, $\hat{\theta}_0^{\mathrm{reg}}$ is a consistent estimator of θ_0, and thus $\hat{\tau}_{\mathrm{CFD}}^{\mathrm{reg}} = \hat{\theta}_1 - \hat{\theta}_0^{\mathrm{reg}}$ and $\hat{\tau}_{\mathrm{CMF}}^{\mathrm{reg}} = \hat{\theta}_1 / \hat{\theta}_0^{\mathrm{reg}}$ are consistent for τ_{CFD} and τ_{CMF}, respectively.

The second estimator is based on weighting. Specifically, Abadie (2005) showed that under Assumptions 1 and weak overlap,

$$\theta_0 = \frac{1}{\pi}\mathbb{E}\Big\{ G_i Y_{it} + \frac{(1 - G_i)(Y_{i,t+1} - Y_{it})e(\mathbf{X}_i))}{1 - e(\mathbf{X}_i)} \Big\}. \tag{8}$$

where $\pi = Pr(G_i = 1)$.

If the propensity score is correctly estimated by $e(\mathbf{X}_i;\hat{\boldsymbol{\alpha}})$, where $\boldsymbol{\alpha}$ is the parameter of the propensity score model, equation (8) suggests the following weighting estimator for θ_0:

$$\hat{\theta}_0^{\mathrm{wt}} = \frac{\sum_{i=1}^N G_i Y_{it} w_i}{\sum_{i=1}^N G_i} + \frac{\sum_{i=1}^N (1 - G_i)(Y_{i,t+1} - Y_{it}) w_i}{\sum_{i=1}^N G_i}, \tag{9}$$

where $w_i = 1$ for the treated group and $w_i = e(\mathbf{X}_i;\hat{\boldsymbol{\alpha}})/[1-e(\mathbf{X}_i;\hat{\boldsymbol{\alpha}})]$ for the control group. This further gives $\hat{\tau}^{\text{wt}}_{\text{CFD}} = \hat{\theta}_1 - \hat{\theta}^{\text{wt}}_0$ and $\hat{\tau}^{\text{wt}}_{\text{CMF}} = \hat{\theta}_1/\hat{\theta}^{\text{wt}}_0$. Re-weighting the observed crash counts by these ATT weights, we create a pseudo-population in which the covariates are balanced between treatment groups (Li et al., 2018); the covariate balance consists the basis of valid group comparison. The weighting estimator avoids specifying the distributions of outcomes, but is in general not as efficient as outcome regression if the outcome model is correctly specified.

2.4 Double-Robust Estimation

The consistency of the regression estimator and the weighting estimator depends on the correct specification of the outcome model and propensity score model, respectively. Here, we propose a new hybrid DID estimator that augments weighting with regression:

$$\hat{\theta}^{\text{dr}}_0 = \hat{\theta}^{\text{wt}}_0 + \frac{1}{\sum_{i=1}^N G_i}\sum_{i=1}^N \frac{(G_i - e(\mathbf{X}_i;\hat{\boldsymbol{\alpha}}))\{\nu(\mathbf{X}_i;\hat{\boldsymbol{\gamma}}) - \mu(\mathbf{X}_i;\hat{\boldsymbol{\beta}})\}}{1-e(\mathbf{X}_i;\hat{\boldsymbol{\alpha}})}. \tag{10}$$

This estimator can alternatively be written as a regression estimator augmented with weighting as

$$\hat{\theta}^{\text{dr}}_0 = \hat{\theta}^{\text{reg}}_0 + \frac{\sum_{i=1}^N (1-G_i)(\hat{R}_{i,t+1} - \hat{R}_{it})w_i}{\sum_{i=1}^N G_i}, \tag{11}$$

where the residuals are defined as $\hat{R}_{i,t+1} = Y_{i,t+1} - \nu(\mathbf{X}_i;\hat{\boldsymbol{\gamma}})$, $\hat{R}_{it} = Y_{it} - \mu(\mathbf{X}_i;\hat{\boldsymbol{\beta}})$. Based on these two equivalent formulations, we establish the following large-sample robustness property in Proposition 1, and include the proof in the Appendix.

Proposition 1 *As the sample size $n \to \infty$, the proposed estimator $\hat{\theta}^{\text{dr}}_0$ converges in probability to θ_0 if either $e(\mathbf{X}_i;\hat{\boldsymbol{\alpha}})$ is consistent to the true propensity score or both $\nu(\mathbf{X}_i;\hat{\boldsymbol{\gamma}})$ and $\mu(\mathbf{X}_i;\hat{\boldsymbol{\beta}})$ are consistent for the true mean functions.*

By proposition 1, we can obtain the DR estimators for τ_{CFD} and τ_{CMF} by $\hat{\tau}^{\text{dr}}_{\text{CFD}} = \hat{\theta}_1 - \hat{\theta}^{\text{dr}}_0$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = \hat{\theta}_1/\hat{\theta}^{\text{dr}}_0$. In fact, estimator (11) extends the DR estimator for ATT in the cross-sectional setting by Mercatanti and Li (2014), who point out that the DR estimator may serve as a diagnostic tool in practical applications. Specifically, if the DR estimate differs from the regression estimate but is similar to the weighting estimate, it may suggest a potential misspecification of the regression function or lack of covariate overlap; if the DR estimate is close to the regression estimate but differs from the weighting estimate, it may suggest a potential misspecification of the propensity score model. We will exploit this diagnostic property of the DR estimate in the traffic safety study in Section 4.

Although our presentation of the DR estimator is centered around the traffic safety application, the DR estimator should apply equally well in conventional program evaluation studies, such as estimating the causal effect of job training program on earnings (Heckman et al., 1997; Heckman, 1998). In that case, the causal estimand is usually defined on the additive scale similar to (2). However, since the earning outcomes are treated as continuous variables, the predicted mean functions ν and μ in the DR estimator could simply be obtained from the two-way fixed-effects model used in Ashenfelter and Card (1985) and Abadie (2005) rather than from (6).

For estimating the additive causal estimand, τ_{CFD}, in the traffic study, the DR estimator $\hat{\tau}^{\text{dr}}_{\text{CFD}}$ differs from the existing double-robust estimator for ATE, in the sense that $\hat{\theta}^{\text{dr}}_0$ only requires estimating the outcome model among the control group but not the treated group. Further, the proposed estimator $\hat{\tau}^{\text{dr}}_{\text{CFD}}$ is indeed a member of the augmented inverse probability weighting (AIPW) estimator (Robins et al., 1994), but is distinct from the most efficient member, which necessarily requires an outcome model for the treated group. To obtain the most efficient AIPW-DID estimator, one could adapt the corresponding efficient AIPW estimator for estimating ATT designed for cross-sectional data (Yang and Ding, 2018), by essentially replacing their cross-sectional outcome with the before-after difference. Despite the efficiency advantage, we caution that such an estimator is not double-robust since it fails to be consistent to the target estimand once the propensity score model is misspecified. In traffic safety applications where the treated group often includes only a small number of pilot sites, the efficient DID estimator is less attractive because a count regression (e.g. negative binomial regression is routinely used in traffic safety studies) model tends to be unstable with non-convergence issues. For these reasons, we focus on the double-robust DID estimator $\hat{\theta}^{\text{dr}}_0$ instead.

Since our estimator $\hat{\tau}^{\text{dr}}_{\text{CFD}}$ differs from the most efficient AIPW-DID estimator, we suspect that $\hat{\tau}^{\text{dr}}_{\text{CFD}}$ may not guarantee to be asymptotically more efficient than the weighting estimator when all models are correct. This is formalized in Proposition 2 with an additive causal estimand and when the true propensity score is known. Proposition 2 is not directly useful for inference as it assumes known propensity scores, but may provide insights for efficiency comparisons of the DID estimators in simulation experiments. Its proof is given in the Appendix.

Proposition 2 *For estimating the additive causal estimand, assuming the true propensity score is known, the ith influence function of the weighting estimator is*

$$\varphi^{wt}_i = \frac{(G_i - e(\mathbf{X}_i))(Y_{i,t+1} - Y_{it})}{\pi(1 - e(\mathbf{X}_i))} - \tau_{CFD}.$$

Further assuming the regression functions are known, the ith influence function of the double-robust estimator is

$$\varphi^{\text{dr}}_i = \frac{(G_i - e(\mathbf{X}_i))(Y_{i,t+1} - Y_{it} - \{\nu(\mathbf{X}_i) - \mu(\mathbf{X}_i)\})}{\pi(1 - e(\mathbf{X}_i))} - \tau_{CFD}.$$

The double-robust estimator is asymptotically at least as efficient as the weighting estimator only when $\mathbb{V}(\varphi^{wt}_i) - \mathbb{V}(\varphi^{dr}_i) \geq 0$. *However, this inequality does not always hold. The full expression of* $\mathbb{V}(\varphi^{wt}_i) - \mathbb{V}(\varphi^{dr}_i)$ *is provided in the Appendix.*

Even though the DR estimator is more robust to model misspecification in the DID design, Proposition 2 suggests that it may be asymptotically less efficient than the weighting estimator even if all models are correctly specified. This is in sharp contrast to the existing results developed for estimating the average treatment effect (ATE). In the latter case, it is well-known that the double-robust estimator is asymptotically at least as efficient as the propensity score weighting estimator when all models are correctly specified (Bang and Robins, 2005; Tsiatis, 2006).

In the traffic safety application, we use a logistic regression to estimate the propensity scores. Since both the logistic and negative binomial models are smooth parametric models, we use the nonparametric bootstrap (Efron and Tibshirani, 1993) to obtain the associated $(1-\alpha)$ confidence interval (CI) and hence account for the uncertainty in estimating the nuisance parameters. For example, the following two steps are carried out to arrive at the CI estimator for $\hat{\tau}^{\text{dr}}$. First, we re-sample with replacement from the empirical distribution $\hat{\mathbb{F}}_N(\mathcal{Z})$ to obtain the bth ($b = 1, \ldots, B$) bootstrap replicate, $\{\mathcal{Z}_j^b, j = 1, \ldots, N\}$, from which we compute $\hat{\tau}^{\text{dr},b}$. We then estimate the $\alpha/2$th and $(1-\alpha/2)$th quantiles of the collection of the bootstrap estimates, $\{\hat{\tau}^{\text{dr},b}, b = 1, \ldots, B\}$, to form the lower and upper confidence limits for $\hat{\tau}^{\text{dr}}$. Since Y_{it} and $Y_{i,t+1}$ are repeated measurements from the same site in the before and after periods, there may be non-zero residual correlation between these crash counts. An advantage of the bootstrap procedure is that the correlation between repeated measurements are automatically taken into account by re-sampling the entire observed data vector $\mathcal{Z}_i$.

3. Simulations

To illustrate the performance of different DID estimators, we conduct a small simulation study that mimics the real rumble strip application. Specifically, we simulate under a two-period two-group design. Each simulation has $N = 2000$ units. Each unit has a binary covariate X_1 and a continuous covariate X_2, generated as follows:

$$X_1 \sim \text{Bernoulli}(0.25), \quad X_2|X_1 \sim \text{Normal}(2 + 6X_1, 2^2).$$

We simulate the treatment group label G_i independently from a Bernoulli distribution with success probability being the propensity score:

$$\text{logit}\{e(\mathbf{X})\} = -2.0 + X_1 - 0.2X_2 + 0.04X_2^2. \tag{12}$$

Under the true propensity score model, the marginal treatment prevalence is approximately 20%, resembling our real application.

We generate the potential crash counts from negative binomial models, with different mean functions but same dispersion parameter $\phi = 2.5$. Specifically, we assume

$$\begin{aligned} Y_t(0)|\mathbf{X}, G = 0 \sim \text{NB}(\mu_{00}(\mathbf{X}), \phi), \quad Y_t(0)|\mathbf{X}, G = 1 \sim \text{NB}(\mu_{01}(\mathbf{X}), \phi), \\ Y_{t+1}(0)|\mathbf{X}, G = 0 \sim \text{NB}(\nu_{00}(\mathbf{X}), \phi), \quad Y_{t+1}(1)|\mathbf{X}, G = 1 \sim \text{NB}(\nu_{11}(\mathbf{X}), \phi), \end{aligned}$$

with

$$\begin{aligned} \mu_{00}(\mathbf{X}) &= \exp(-2.0 + 0.4X_1 + 0.43X_2 - 0.022X_2^2), \\ \mu_{01}(\mathbf{X}) &= \exp(-3.0 + 0.3X_1 + 0.43X_2 - 0.022X_2^2), \\ \nu_{00}(\mathbf{X}) &= \exp(-1.9 + 0.5X_1 + 0.43X_2 - 0.022X_2^2), \\ \nu_{11}(\mathbf{X}) &= \exp(-2.5 + 0.1X_1 + 0.43X_2 - 0.022X_2^2). \end{aligned} \tag{13}$$

Under the parallel trend, the mean function of the counterfactual crash outcome for the treated sites is $\nu_{01}(\mathbf{X}) = \nu_{00}(\mathbf{X}) + \mu_{01}(\mathbf{X}) - \mu_{00}(\mathbf{X})$. The coefficients of the mean functions

in (13) are informed by regression fit from analyzing the total crashes from the traffic safety application, and ensure that $\nu_{01}(\mathbf{X})$ is positive over the support of $\mathbf{X}$. The true values of CFD and CMF, evaluated in large samples, are -0.078 and 0.862, respectively.

We simulate 500 replicates based on the models specified above. For each replicate, we use $\hat{\theta}_1$ to estimate θ_1, but use different estimators for θ_0. We first use the direct moment estimator based on the observed sample averages given in equation (5). This estimator ignores pre-treatment covariate information and is only valid when there is no selection bias, namely, when the parallel trend holds unconditionally on the pre-treatment covariates. It is used here to quantify the selection bias in the data generation process. Further, the following estimators are compared.

Outcome regression: we adopt the regression estimator in equation (7) with correctly specified mean functions for $\mu(\mathbf{X})$ and $\nu(\mathbf{X})$. We also study the regression estimator with incorrectly specified mean functions that omit the linear term X_1 and the quadratic term X_2^2 in $\mu(\mathbf{X})$ and $\nu(\mathbf{X})$. These two estimators are labeled by REG and REG-mis, respectively.

Propensity score weighting: we consider the weighting estimator in equation (9) with the correctly specified propensity score model, as well as the weighting estimator with an incorrectly specified propensity score model that omits X_1 and X_2^2 in model (12). These two estimators are labeled by WT and WT-mis, respectively.

Double-Robust methods: we compare the DR estimator in equation (10) with correctly specified propensity score and outcome models (DR), the DR estimator with correctly specified outcome regression model but incorrectly specified propensity score model that omits X_1 and X_2^2 (DR-po), the DR estimator with correctly specified propensity score model but incorrectly specified outcome regression model that omits X_1 and X_2^2 (DR-ps), and the DR estimator with propensity score and outcome regression models being both incorrectly specified (DR-mis).

Table 1: Absolute bias (Bias $\times 10^2$), root mean squared error (RMSE $\times 10^2$) and coverage of the 95% confidence interval (Coverage) associated with each estimator for estimating τ_{CFD} and $\log(\tau_{\text{CMF}})$ in the simulations. The confidence intervals are computed based on 500 bootstrap samples from each simulated data set.

	τ_{CFD}			$\log(\tau_{\text{CMF}})$		
	Bias	RMSE	Coverage	Bias	RMSE	Coverage
Direct	13.4	14.5	33.4	27.6	30.5	38.4
REG	0.4	13.4	94.8	1.9	26.6	94.8
REG-mis	10.6	20.0	90.0	14.3	31.3	90.4
WT	0.2	14.1	95.6	2.6	27.7	95.6
WT-mis	4.7	10.0	90.8	9.8	20.7	91.0
DR	0.5	14.5	95.4	2.2	28.6	95.4
DR-po	0.4	13.4	94.6	2.0	26.6	94.8
DR-ps	2.6	15.8	95.8	1.1	30.0	95.6
DR-mis	7.0	16.7	91.8	9.2	27.6	92.0

Table 1 presents the absolute bias, root mean squared error (RMSE) of each point estimator and the coverage of the corresponding 95% bootstrap confidence intervals. Among all the estimators, the direct estimator is associated with the largest bias and RMSE and the lowest coverage in estimating both τ_{CFD} and $\log(\tau_{\text{CMF}})$. This is as expected because X_1 and X_2 affect both the treatment assignment and the potential outcomes, and induce selection bias. The DID regression, weighting, and DR estimators all present small and comparable bias when the corresponding models are correctly specified. When the outcome regression functions $\mu(\mathbf{X})$ and $\nu(\mathbf{X})$ are misspecified, the regression estimator shows inflated bias and RMSE, with reduced coverage. Similarly, misspecification of the propensity score model also leads to increased bias and sub-nominal coverage for the DID weighting estimator. In this simulation, the substantial reduction in the variance of the weights from a misspecified propensity score model appears to outweigh the inflation in bias, which explains the decreased RMSE associated with WT-mis relative to WT.

The simulation also demonstrates the double robustness property of the DID-DR estimator: when either the propensity score model or outcome model is misspecified, the DR estimator (DR-po and DR-ps) leads to small bias and nominal coverage for both estimands. Interestingly, in estimating the additive effect τ_{CFD}, the outcome model appears have a bigger impact on the DR estimator than the propensity score model. Specifically, when only the outcome model is correctly specified, the DR estimator performs very close to the DR estimator with both models being correctly specified, but the DR estimator under-performs much if only the propensity score is correct. Similar phenomenon was previously observed in the DR estimation of ATT and ATE in the cross-sectional setting (e.g. Li et al., 2013). This pattern is not obvious for ratio estimand $\log(\tau_{\text{CMF}})$, likely because that the bias—an additive and scaled quantity by definition—of a scale-free ratio quantity does not fully capture the true discrepancy in estimating θ_1 and θ_0. Lastly, when both the propensity score and outcome models are misspecified, the DR estimator (DR-mis) results in inflated bias and under-coverage; nonetheless, even under this scenario, the misspecified DR estimator still outperforms the corresponding misspecified regression estimator with 46% and 6% reduction in relative bias for estimating the additive and ratio estimands, respectively. In addition, we observe in the simulations that the Monte Carlo variance of the DR estimator, when both models are correctly specified, is very close to that of the weighting estimator with a correct propensity score model. This phenomenon may be partially explained by Proposition 2. Specifically, under the current data generating process mimicking the traffic safety application, we found that the Monte Carlo estimate of $N^{-1}\{\mathbb{V}(\varphi_i^{\text{wt}}) - \mathbb{V}(\varphi_i^{\text{dr}})\} < 0$ is negative and close to zero (averaged across simulation iterations).

4. Application to the Pennsylvania Rumble Strip Data

4.1 The Data

Our application is based on the Federal Highway Administration Evaluation of Low-Cost Safety Improvements Pooled Fund Study (Lyon et al., 2015). The study embraced a broader scope and focused on quantifying the safety effectiveness of the combined application of centerline and shoulder rumble strips in mitigating crash outcomes among two-lane rural road locations in Kentucky, Missouri, and Pennsylvania. We obtained the subset of traffic safety records from the Pennsylvania Department of Transportation (PennDOT; `http:`

//www.penndot.gov/), which includes vehicle crash counts for traffic sites within the state of Pennsylvania up to 2012. Since each traffic site is a roadway segment, we use these two terms interchangeably. From 2009 to 2011, centerline and shoulder rumble strips were installed in 331 rural, undivided two-lane roadway segments for a total of over 200 miles. The control group consists of five times more sites that did not receive rumble strips before 2012 but had similar traffic volume. Therefore, the data we analyze consist of around 2000 rural highway segments, approximately 17% of which received the treatment. We define year 2008 as the before period and year 2012 as the after period.

We consider four types of crash outcomes: (1) fatal-plus-injury (FI)—crashes that involve at least one fatal or injured person; (2) property-damage-only (PDO)—crashes where no occupant was injured; (3) run-off-the-road (ROR)—crashes where a vehicle travels outside the trafficway and collides with a natural or artificial object in an area not intended for vehicles; this is a subset of the first two crash types; (4) total number of fatal-plus-injury and property-damage-only crashes (TOT). Table 2 presents the aggregated crash counts for each type among both treated and control sites in the before and after periods.

Table 2: Crash counts by type for both treated and control sites in the before and after periods. FI: fatal-plus-injury; PDO: property-damage-only; ROR: run-off-the-road; TOT: total.

	Treated ($N_1 = 331$)		Control ($N_0 = 1,655$)	
	Before	After	Before	After
FI	78	77	441	436
PDO	61	41	350	321
ROR	22	21	123	143
TOT	139	118	791	757

The pre-treatment covariates we consider are site-specific characteristics often suggested in constructing crash frequency models (AASHTO, 2010). These variables include the operational characteristic of a roadway segment, the speed limit (high speed if the posted limit is above 45 mph and low speed otherwise), as well as geometric features of a roadway: segment length in miles, pavement width (three categories), average shoulder width (three categories), number of driveways (three categories), existence of intersections (two categories), existence of curves (two categories) and average degree of curvature. An important covariate is AADT—the average annual daily traffic volume. Although strictly speaking AADT is a time-varying covariate, we found that in this application the AADT of the before and after periods are very similar across all sites; thus we assume AADT is time-invariant and take the before period value as the covariate. Table 3 presents the descriptive statistics of the covariates.

4.2 Model Specification

We estimate the propensity score by logistic regression including all the pre-treatment site characteristics. We adopt the power series specification for the continuous variables

Table 3: Definition of variables and their descriptive statistics by treatment group. Mean (st. dev), [Min, Max] values are given for each continuous variable and the number of traffic sites (percentages) are given for each level of the categorical variables. Total sample size $N = 1,986$.

Variable	Definition	Treated	Control
AADT	Annual average daily traffic volume; vehicles per day	3,520 (2,628) [818, 15,033]	3,636 (2,495) [678, 15,379]
Length	Roadway segment length in miles	0.47 (0.16) [0.01, 0.75]	0.48 (0.13) [0.03, 0.76]
Width	Pavement width in feet		
	width ≤ 20	20 (6.0)	346 (20.9)
	$20 <$ width ≤ 23	169 (51.1)	828 (50.0)
	otherwise	142 (42.9)	481 (29.1)
Speed	Posted speed limit		
	low if limit ≤ 45 mph	216 (65.3)	956 (57.9)
	high otherwise	115 (34.7)	696 (42.1)
Shoulder	Average shoulder width in feet		
	width ≤ 3	88 (26.6)	867 (52.4)
	$3 <$ width ≤ 6	191 (57.7)	643 (38.8)
	otherwise	52 (15.7)	145 (8.8)
Driveways	Number of driveways		
	no driveway	24 (7.2)	101 (6.1)
	$1 \leq$ number of driveways ≤ 10	235 (71.0)	1,100 (66.5)
	otherwise	72 (21.8)	454 (27.4)
Intersections	Inclusion of intersections		
	No intersections	257 (77.6)	1,162 (70.2)
	At least 1 intersection	74 (22.4)	493 (29.8)
Curves	Existence of horizontal curves		
	No curves	143 (43.2)	701 (42.4)
	At least 1 curve	188 (56.8)	954 (57.6)
Curvature	Average degree of curvature	2.81 (3.59) [0, 25.28]	3.92 (7.59) [0, 132.30]

(AADT, Length and Curvature) with the optimal order of terms $1 \leq l \leq 5$ selected by leave-one-out-cross-validation. We choose to include up to the third-order terms of the continuous variables in the propensity score model since $l = 3$ corresponds to the lowest mean squared error for predicting treatment. The fitted propensity score model suggests that road segments with wider pavement and shoulder, low speed limit, at least one driveway, no intersections nor curves are more likely to receive rumble strip installation.

For both the regression and DR estimators, we model the cross-sectional means of potential outcomes among the reference sites during each period. AADT and segment length are transformed to the log scale, as is common practice in developing crash frequency models in traffic safety research (AASHTO, 2010). To allow for over-dispersion, we use negative binomial regression to estimate model parameters. Specifically, we assume the conditional distributions of $Y_{it}(0)$ and $Y_{i,t+1}(0)$ given $\mathbf{X}_i$ in the reference group as in (6), where

$$\begin{aligned} \mu(\mathbf{X}) &= L^{\beta_L} \cdot \text{AADT}^{\beta_{\text{AADT}}} \cdot \exp\Big(\beta_0 + \sum_{j=1}^{J} \beta_j X_j\Big), \\ \nu(\mathbf{X}) &= L^{\gamma_L} \cdot \text{AADT}^{\gamma_{\text{AADT}}} \cdot \exp\Big(\gamma_0 + \sum_{j=1}^{J} \gamma_j X_j\Big). \end{aligned} \tag{14}$$

In (14), L denotes the segment length, AADT is the traffic volume and J is the number of remaining covariates (including dummy variables). We adopt the log-linear specification for the outcome model since it performs as well as its power series counterpart regarding mean squared error estimated by leave-one-out-cross-validation, and yet is computationally convenient without convergence issues.

4.3 Assessment of Overlap, Balance and Parallel Trend

We assess the weak overlap assumption by visually checking the overlap in the histograms of the estimated propensity scores for the treated and control sites (Figure 1). The histogram suggests satisfactory overlap between the two groups.

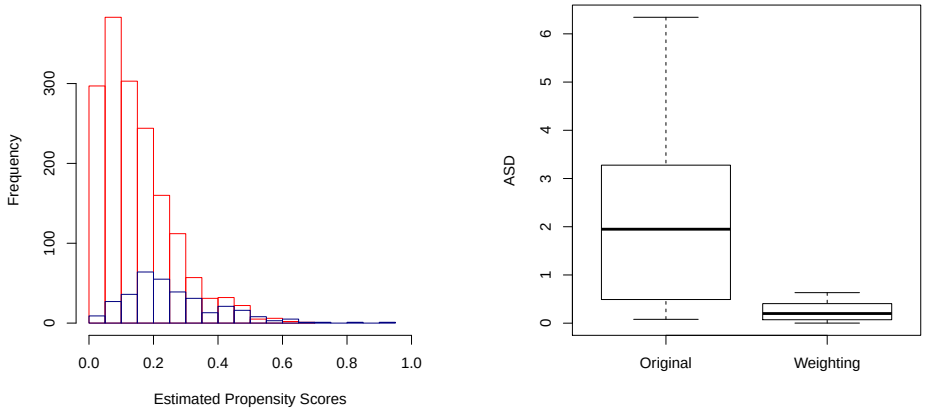

Figure 1: Left panel: histogram of the estimated propensity score for the treated sites (blue) and the control sites (red). Right panel: boxplot of the absolute standardized difference all covariates in the original and weighted data.

We further check the covariate balance in the original and weighted sample by calculating the absolute standardized difference (ASD) of each covariate (including up to the third-order

term for each continuous variable) between the two treatment groups, defined as

$$\text{ASD} = \left| \frac{\sum_{i=1}^{N} G_i X_i w_i}{\sum_{i=1}^{N} G_i w_i} - \frac{\sum_{i=1}^{N} (1-G_i) X_i w_i}{\sum_{i=1}^{N} (1-G_i) w_i} \right| \Big/ \sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}, \tag{15}$$

where N_1, N_0 are the number of treated and reference sites, s_1^2, s_0^2 are the variances of the unweighted covariate in the treated and control group, respectively. The weight $w_i = 1$ for all sites in the original data and the ASD is the standard two-sample t-statistic. For the weighted data, w_i is the ATT weight introduced in Section 2.3. The right panel in Figure 1 presents the boxplots of the ASD; it shows that propensity score weighting substantially improves the covariate balance, with the largest ASD value equal to 0.63 in the weighted sample compared to 6.34 in the unweighted sample (the standard threshold for significant difference is 1.96). The good covariate balance supports that the propensity scores are well estimated.

Table 4: Estimated CFD ($\hat{\tau}_{\text{CFD}}$) and CMF ($\hat{\tau}_{\text{CMF}}$) and the corresponding 95% confidence intervals for all crash types in the "no treatment" evaluation using different DID estimators.

		Direct	REG	WT	DR
FI	$\hat{\tau}_{\text{CFD}}$	-0.060	-0.026	-0.029	-0.028
		(-0.144,0.026)	(-0.104,0.062)	(-0.121,0.073)	(-0.123,0.067)
	$\hat{\tau}_{\text{CMF}}$	0.798	0.902	0.891	0.892
		(0.587,1.107)	(0.651,1.331)	(0.621,1.370)	(0.623,1.341)
PDO	$\hat{\tau}_{\text{CFD}}$	0.008	0.010	0.026	0.027
		(-0.066,0.078)	(-0.066,0.086)	(-0.052,0.108)	(-0.052,0.110)
	$\hat{\tau}_{\text{CMF}}$	1.045	1.059	1.164	1.169
		(0.692,1.594)	(0.693,1.686)	(0.738,2.054)	(0.744,2.106)
ROR	$\hat{\tau}_{\text{CFD}}$	-0.016	0.001	0.007	0.009
		(-0.064,0.029)	(-0.051,0.046)	(-0.043,0.053)	(-0.044,0.054)
	$\hat{\tau}_{\text{CMF}}$	0.809	1.014	1.121	1.150
		(0.431,1.522)	(0.491,2.565)	(0.524,3.096)	(0.521,3.234)
TOT	$\hat{\tau}_{\text{CFD}}$	-0.052	-0.015	-0.003	-0.002
		(-0.146,0.073)	(-0.117,0.109)	(-0.119,0.140)	(-0.115,0.138)
	$\hat{\tau}_{\text{CMF}}$	0.890	0.965	0.993	0.996
		(0.714,1.192)	(0.761,1.333)	(0.764,1.427)	(0.762,1.425)

To indirectly assess the key parallel trend assumption, we perform a DID analysis of the crash outcomes for two pre-treatment periods. Specifically, we obtain the crash outcome, $Y_{i,t-1}$, during the year of 2004 for each traffic site and treat it as the proxy-before observation; the crash outcome, Y_{it}, during the year of 2008 are then regarded as the proxy-after data. As discussed in Section 2.2, if the parallel trend assumption is plausible, then the

estimated CFD and CMF based on the proxy-before-after observations should be close to 0 and 1, respectively, because in reality rumble strips were not applied until after 2008.

Table 4 presents the results of this "no treatment" analysis. For all crash types, DR and weighting estimators produce similar estimates for CFD and CMF. Overall, the confidence intervals for CFD include 0 for all types of crashes regardless of the choice of DID estimator. However, it is worth noting that the CFD estimates from DR and weighting for the ROR and total crashes are close to 0, which further support the plausibility of the parallel trend. By contrast, there is a potential for violation of parallel trend regarding FI and PDO crashes since the DR estimates for CFD tend to deviate from the null. Nevertheless, the lack of statistical significance may still permit the subsequent DID analyses.

Table 5: Estimated CFD ($\hat{\tau}_{\text{CFD}}$) and CMF ($\hat{\tau}_{\text{CMF}}$) and the corresponding 95% confidence intervals for all crash types with before and after data using different DID methods.

		Direct	REG	WT	DR
FI	$\hat{\tau}_{\text{CFD}}$	0.000	-0.022	-0.009	-0.008
		(-0.087,0.078)	(-0.118,0.060)	(-0.110,0.077)	(-0.108,0.079)
	$\hat{\tau}_{\text{CMF}}$	1.000	0.912	0.963	0.966
		(0.706,1.470)	(0.629,1.318)	(0.657,1.458)	(0.660,1.474)
PDO	$\hat{\tau}_{\text{CFD}}$	-0.043	-0.037	-0.056	-0.058
		(-0.113,0.026)	(-0.106,0.036)	(-0.134,0.022)	(-0.137,0.018)
	$\hat{\tau}_{\text{CMF}}$	0.743	0.770	0.687	0.681
		(0.455,1.223)	(0.462,1.384)	(0.409,1.196)	(0.404,1.180)
ROR	$\hat{\tau}_{\text{CFD}}$	-0.015	-0.022	-0.039	-0.039
		(-0.060,0.029)	(-0.066,0.022)	(-0.099,0.006)	(-0.096,0.006)
	$\hat{\tau}_{\text{CMF}}$	0.808	0.746	0.617	0.619
		(0.437,1.592)	(0.417,1.382)	(0.328,1.085)	(0.331,1.090)
TOT	$\hat{\tau}_{\text{CFD}}$	-0.043	-0.060	-0.065	-0.066
		(-0.145,0.063)	(-0.174,0.049)	(-0.188,0.052)	(-0.191,0.052)
	$\hat{\tau}_{\text{CMF}}$	0.893	0.856	0.845	0.844
		(0.684,1.189)	(0.651,1.154)	(0.627,1.166)	(0.626,1.154)

4.4 Results

We analyzed crash outcomes in 2008 and 2012 using different DID estimators for all crash types and present the results in Table 5. As observed in the simulations, the direct estimator (5) is subject to selection bias and tends to give different results from the rest. Across all four crash types, the DR estimator produces CFD and CMF results similar to the weighting estimator, but sometimes different from the regression estimator. Given the satisfactory overlap indicated in Figure 1, the difference in estimates suggests that the outcome regression model may be mildly misspecified. The CFD and CMF for FI crashes estimated by the DR approach are both close to the null values, implying negligible effect

from rumble strips on mitigating FI crashes. The application of rumble strips seems to reduce the PDO crashes with CFD and CMF estimated to be $\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.058$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.681$ using the DR approach. However, cautions need to be exercised to interpret these estimates since the parallel trend assumption may be questionable, as discussed previously. The parallel trend is deemed plausible for the total crashes, and we find that rumble strips have a potentially beneficial effect on total crash frequency ($\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.066$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.844$), but the 95% CIs include the null values. Additionally, the application of rumble strips suggests a potential causal effect on mitigating the ROR crashes, with $\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.039$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.619$ estimated by the DR approach, but the CIs cover 0 and 1. Overall, our analysis only finds beneficial but statistically insignificant effects of rumble strips on reducing crashes. This agrees with the empirical findings of several other traffic safety studies based on alternative data sources and modeling strategies (Griffith, 1999; Gårder and Davies, 2006; Khan et al., 2015).

5. Discussion

In this paper, we draw causal inference in traffic safety before-after studies within the DID framework and propose a new double-robust DID estimator. The primary concern for observational traffic safety data is related to bias, which may be due to confounding, site selection or model misspecification, among others. Our DR estimator grants two chances for consistent estimation of the causal effect and has been demonstrated to have small bias from misspecification of either the propensity score model or the outcome model. Applying the DR method and several alternative methods to a real data, we find that rumble strips have a moderate but statistically insignificant beneficial effect in reducing vehicle crashes. These insignificant findings may be partially due to the limited number of crash events over a one-year period, a limitation of our available data. It would be of interest to update the CFD and CMF estimates with longer before and after periods.

Though the causal rate ratio estimand, CMF, dominates the traffic safety studies, we recommend assessing alternative estimand such as the causal rate difference, CFD to offer a more comprehensive picture of the effectiveness. This is because that CMF is scale-free and does not inform the absolute change in the expected crash frequency. For example, in our application, the CFD estimate suggests a modest absolute change in crash frequency ($\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.039$) for the ROR crashes, which can be translated into an average reduction of 4 crashes per 100 road segments due to rumble strips. On the other hand, the CMF estimate is $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.617$, which indicate a large proportional change. This slight discrepancy comes from the fact that the ROR crashes constitute a small fraction of the total crashes.

There are several limitations of this study. First, a limitation of the DID framework is that the parallel trend assumption is scale-dependent. For example, it may hold for the original Y but not for a nonlinear monotone transformation of Y, such as $\log(Y)$. A common alternative scale-free identification condition for the before-after design is the ignorability assumption conditional on the lagged outcomes. In the context of linear models, Angrist and Pischke (2009) show that the DID estimate and lagged-outcome regression estimates have a bracketing relationship. Namely, if ignorability is correct, then mistakenly assuming parallel trend will overestimate a true positive effect; in contrast, if parallel trend is correct, then mistakenly assuming ignorability will underestimate a true positive effect. Thus, one

can treat the estimate under each assumption as the upper and lower bounds of the true effect in practice. It is particularly relevant to traffic safety studies—where the outcome is usually counts—to evaluate whether such a bracketing relationship holds more generally beyond the linear setting.

Second, the SUTVA may be violated and such a violation could lead to a biased average causal effect estimate. The violation is more likely if the traffic sites are adjacent to each other allowing for a potential spillover effect. For example, it is possible that a drowsy and fatigued driver was alerted in a roadway segment with shoulder rumble strips and hence was less likely to have a run-off-the-road crash in a nearby reference site, thus biasing the causal estimate towards the null. It is also likely that *crash migration* leads to violation of SUTVA. For instance, a vehicle travelling through a reference site with low visibility may end up in a crash in a consecutive site with rumble strips. However, the reporting officer usually traced the location where the crash was initiated by a careful analysis of the available evidence at the crash site, and may mitigate such concerns. In any case, the extension of the DID analysis that accounts for interference between roadway segments in the spirit of Hudgens and Halloran (2008) would be of interest.

We have adopted smooth parametric models to estimate the propensity score and the crash counts. In this case, the resulting DID estimators are all asymptotically linear, and the nonparametric bootstrap enables valid inference (Shao and Tu, 2012). This also underlies why the bootstrap CI maintains nominal coverage for the DR-po and DR-ps estimators in our simulation study. In practice, since well-estimated propensity score and outcome models are critical for the consistency of the DR estimator, an appealing strategy is to leverage data-adaptive machine learning techniques for estimating the propensity scores and for predicting the crash counts. Specifically, one could use boosting to estimate the propensity score (McCaffrey et al., 2004, 2013), which has been demonstrated to maintain adequate weighted covariate balance (Lee et al., 2009), or use random forest to better predict the counterfactual safety outcomes (Breiman, 2001; Liaw and Wiener, 2002). However, the nonparametric bootstrap CI may not guarantee to carry nominal coverage in those cases since the resulting estimator may no longer be asymptotically linear.

Finally, we have only developed double-robust estimation within the canonical two-period DID design with panel data. More complicated before-after data structure may arise in other policy evaluation contexts. For example, the treatment (policy) could be administered to a small number of states, and repeated cross sections or surveys are taken at the household level or individual level to measure the before and after outcomes. When repeated cross sections or random surveys are taken in both the before and after periods (rather than panel observations for the same group of units), the proposed double-robust DID estimator may not directly apply since the identification condition differs from equation (8). Abadie (2005) provided the revised identification condition and suggested the corresponding weighting estimator (Section 3.2.1 of the paper). Therefore, an appropriate double-robust estimator is obtained by modifying the propensity score weighting component along the lines of Abadie (2005). Additionally, one must address the within-state correlations among households or individuals when the treatment is applied at the state level. In particular, valid bootstrap should proceed by resampling the states so that the within-state correlation structure is preserved (Li et al., 2013). In other program evaluation applications with staggered adoption and multiple time periods, the two-way fixed-effects model repre-

sents a standard regression estimator for causal inference. Callaway and Sant'Anna (2018) recently defined several new average causal estimands appropriate for the multiple-period DID design, and studied a propensity score weighting estimator. An important avenue for future research is to provide a double-robust DID extension by combining the weighting estimator of Callaway and Sant'Anna (2018) and the two-way fixed-effects outcome model for improved inference with multiple-period data.

Acknowledgments

We are grateful to the Pennsylvania Department of Transportation (PennDOT) and Eric Donnell (Pennsylvania State University) for providing the traffic safety application, to Peng Ding (University of California, Berkeley) for insightful discussions. We thank the Editor, Associate Editor and an anonymous referee for helpful comments that improved an earlier version of this manuscript.

Appendix

Proof of Proposition 1

The DR estimators are constructed as $\hat{\tau}^{\rm dr}_{\rm CFD} = \hat{\theta}_1 - \hat{\theta}^{\rm dr}_0$ and $\hat{\tau}^{\rm dr}_{\rm CMF} = \hat{\theta}_1/\hat{\theta}^{\rm dr}_0$; the moment-based estimator

$$\hat{\theta}_1 = \frac{\sum_{i=1}^N G_i Y_{i,t+1}}{\sum_{i=1}^N G_i} \xrightarrow{p} \frac{\mathbb{E}[G_i Y_{i,t+1}]}{\pi} = \mathbb{E}[Y_{i,t+1}(1)|G_i = 1] = \theta_1, \tag{16}$$

where $\pi = Pr(G_i = 1) > 0$. To show that $\hat{\tau}^{\rm dr}_{\rm CFD}$ and $\hat{\tau}^{\rm dr}_{\rm CMF}$ are double-robust for estimating $\tau_{\rm CFD}$ and $\tau_{\rm CMF}$, it suffices to show that $\hat{\theta}^{\rm dr}_0$ is double-robust for estimating θ_0.

We first assume that the propensity score model $e(\mathbf{X}; \boldsymbol{\alpha})$ is correctly specified while the outcome model may be subject to misspecification. We assume certain regularity conditions hold (e.g., smooth regression functions and bounded moments for all covariates), and denote the maximum likelihood estimators for model parameters by $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$. Under these assumptions, $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_0$, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$, where $\boldsymbol{\alpha}_0$ is the true value for the correct propensity score model but $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}^*$ may be different from the true values $\boldsymbol{\gamma}_0$, $\boldsymbol{\beta}_0$. By the results of White (1982), $\boldsymbol{\gamma}^*$ and $\boldsymbol{\beta}^*$ are the least false values that minimize the Kullback-Leibler distance between the probability distribution based on the postulated models and the true data generating models. We first observe that the last term on the right hand side of equation (10) converges in probability to zero. To see why, we write

$$\begin{aligned}
&\frac{1}{\sum_{i=1}^N G_i} \sum_{i=1}^N \frac{(G_i - e(\mathbf{X}; \hat{\boldsymbol{\alpha}}))\{\nu(\mathbf{X}_i; \hat{\boldsymbol{\gamma}}) - \mu(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\}}{1 - e(\mathbf{X}; \hat{\boldsymbol{\alpha}})} \\
\xrightarrow{p} &\frac{1}{\pi} \mathbb{E}\Big[\frac{(G_i - e(\mathbf{X}; \boldsymbol{\alpha}_0))\{\nu(\mathbf{X}_i; \boldsymbol{\gamma}^*) - \mu(\mathbf{X}_i; \boldsymbol{\beta}^*)\}}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)}\Big] \\
= &\frac{1}{\pi} \mathbb{E}\Big[\frac{\{\mathbb{E}(G_i|\mathbf{X}_i) - e(\mathbf{X}; \boldsymbol{\alpha}_0)\}\{\nu(\mathbf{X}_i; \boldsymbol{\gamma}^*) - \mu(\mathbf{X}_i; \boldsymbol{\beta}^*)\}}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)}\Big] = 0,
\end{aligned}$$

where the second to last equation is an application of the Law of Iterated Expectation. Therefore by (10), it is immediate that $\hat{\theta}^{\rm dr}_0$ shares the same probability limit with $\hat{\theta}^{\rm wt}_0$,

which is consistent to θ_0 when the propensity score model is correctly specified (Abadie, 2005). This is why $\hat{\theta}_0^{\text{dr}} \xrightarrow{p} \theta_0$.

Alternatively, suppose the outcome model is correctly specified but the propensity score model may subject to misspecification. In this case, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_0$, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^*$ minimizes the Kullback-Leibler distance between the probability distribution based on the postulated model and the true data generating model (White, 1982) and thus may differ from truth data generating model parameter $\boldsymbol{\alpha}_0$. Then the last term on the right hand side of (11) converges in probability to zero. This is because

$$
\begin{aligned}
&\frac{\sum_{i=1}^N (1-G_i)(\hat{R}_{i,t+1} - \hat{R}_{it}) w_i}{\sum_{i=1}^N G_i} \\
=&\frac{\sum_{i=1}^N (1-G_i)}{\sum_{i=1}^N G_i} \frac{1}{\sum_{i=1}^N (1-G_i)} \sum_{i=1}^N \frac{(1-G_i)(\hat{R}_{i,t+1} - \hat{R}_{it}) e(\mathbf{X}; \hat{\boldsymbol{\alpha}})}{1 - e(\mathbf{X}; \hat{\boldsymbol{\alpha}})} \\
\xrightarrow{p}&\frac{1-\pi}{\pi} \mathbb{E}\Big[\frac{(Y_{i,t+1} - Y_{it}) e(\mathbf{X}; \boldsymbol{\alpha}_0)}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)} \Big| G_i = 0\Big] - \frac{1-\pi}{\pi} \mathbb{E}\Big[\frac{\{\nu(\mathbf{X}_i; \boldsymbol{\gamma}_0) - \mu(\mathbf{X}_i; \boldsymbol{\beta}_0)\} e(\mathbf{X}_i)}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)} \Big| G_i = 0\Big].
\end{aligned}
$$

and

$$
\begin{aligned}
&\Big[\frac{(Y_{i,t+1} - Y_{it}) e(\mathbf{X}; \boldsymbol{\alpha}_0)}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)} \Big| G_i = 0\Big] \\
=&\mathbb{E}\Big[\frac{\{Y_{i,t+1}(0) - Y_{it}(0)\} e(\mathbf{X}; \boldsymbol{\alpha}_0)}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)} \Big| G_i = 0\Big] \\
=&\mathbb{E}\Big[\frac{e(\mathbf{X}; \boldsymbol{\alpha}_0)}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)} \mathbb{E}[Y_{i,t+1}(0) - Y_{it}(0) | \mathbf{X}_i, G_i = 0] \Big| G_i = 0\Big] \\
=&\mathbb{E}\Big[\frac{e(\mathbf{X}; \boldsymbol{\alpha}_0)}{1 - e(\mathbf{X}; \boldsymbol{\alpha}_0)} \{\nu(\mathbf{X}_i; \boldsymbol{\gamma}_0) - \mu(\mathbf{X}_i; \boldsymbol{\beta}_0)\} \Big| G_i = 0\Big],
\end{aligned}
$$

where the second to last equality is granted by the Law of Iterated Expectation and the last equality comes from the definition of the regression function. By (11), it follows that $\hat{\theta}_0^{\text{dr}}$ shares the same probability limit with $\hat{\theta}_0^{\text{reg}}$, which is consistent to θ_0 when the cross-sectional crash frequency model is correctly specified. Therefore $\hat{\theta}_0^{\text{dr}} \xrightarrow{p} \theta_0$, and the double-robust property holds.

Proof of Proposition 2

Denote the ith known true propensity score as $e(\mathbf{X}_i)$, then the weighting estimator is

$$
\begin{aligned}
\tau_{\text{CFD}}^{\text{wt}} &= \frac{\sum_{i=1}^N G_i (Y_{i,t+1} - Y_{it})}{\sum_{i=1}^N G_i} - \frac{1}{\sum_{i=1}^N G_i} \sum_{i=1}^N \frac{(1-G_i)(Y_{i,t+1} - Y_{it}) e(\mathbf{X}_i)}{1 - e(\mathbf{X}_i)} \\
&= \left(\frac{N}{\sum_{i=1}^N G_i}\right) \left\{ \frac{1}{N} \sum_{i=1}^N G_i (Y_{i,t+1} - Y_{it}) - \frac{1}{N} \sum_{i=1}^N \frac{(1-G_i)(Y_{i,t+1} - Y_{it}) e(\mathbf{X}_i)}{1 - e(\mathbf{X}_i)} \right\}.
\end{aligned}
$$

Further observe that

$$\begin{aligned}
&\sqrt{N}(\tau_{\text{CFD}}^{\text{wt}} - \tau_{\text{CFD}}) \\
=&\frac{1}{\pi}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{G_i(Y_{i,t+1}-Y_{it}) - \frac{(1-G_i)(Y_{i,t+1}-Y_{it})e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)} - \tau_{\text{CFD}}\right\} + o_p(1) \\
=&\frac{1}{\pi}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\frac{(G_i - e(\mathbf{X}_i))(Y_{i,t+1}-Y_{it})}{\pi(1-e(\mathbf{X}_i))} - \tau_{\text{CFD}}\right\} + o_p(1),
\end{aligned}$$

where $o_p(1)$ is a residual term that converges in probability to zero. We then obtain φ_i^{wt} as the individual summand in the bracket (Tsiatis, 2006). A similar reasoning is used to obtain φ_i^{dr} in Proposition 2. Notice that

$$\begin{aligned}
\varphi_i^{\text{wt}} + \varphi_i^{\text{dr}} &= \frac{1}{\pi}\frac{G_i - e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)}\{2(Y_{i,t+1}-Y_{it}) - (\nu(\mathbf{X}_i) - \mu(\mathbf{X}_i))\} - 2\tau_{\text{CFD}}, \\
\varphi_i^{\text{wt}} - \varphi_i^{\text{dr}} &= \frac{1}{\pi}\frac{G_i - e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)}\{\nu(\mathbf{X}_i) - \mu(\mathbf{X}_i)\},
\end{aligned}$$

and the difference in asymptotic variance

$$\begin{aligned}
&\mathbb{V}(\varphi_i^{\text{wt}}) - \mathbb{V}(\varphi_i^{\text{dr}}) = \mathbb{E}[(\varphi_i^{\text{wt}} + \varphi_i^{\text{dr}})(\varphi_i^{\text{wt}} - \varphi_i^{\text{dr}})] \\
&= \frac{2}{\pi^2}\mathbb{E}\Big[\Big(\frac{G_i - e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)}\Big)^2(Y_{i,t+1}-Y_{it})(\nu(\mathbf{X}_i) - \mu(\mathbf{X}_i))\Big] - \frac{1}{\pi^2}\mathbb{E}\Big[\Big(\frac{G_i - e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)}\Big)^2(\nu(\mathbf{X}_i) - \mu(\mathbf{X}_i))^2\Big],
\end{aligned}$$

since

$$\frac{2\tau_{\text{CFD}}}{\pi}\mathbb{E}\Big[\Big(\frac{G_i - e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)}\Big)\Big(\nu(\mathbf{X}_i) - \mu(\mathbf{X}_i)\Big)\Big] = \frac{2\tau_{\text{CFD}}}{\pi}\mathbb{E}\Big[\Big(\frac{\mathbb{E}(G_i|\mathbf{X}_i)_i - e(\mathbf{X}_i)}{1-e(\mathbf{X}_i)}\Big)\Big(\nu(\mathbf{X}_i) - \mu(\mathbf{X}_i)\Big)\Big|\mathbf{X}_i\Big] = 0.$$

Unfortunately, the expression for $\mathbb{V}(\varphi_i^{\text{wt}}) - \mathbb{V}(\varphi_i^{\text{dr}})$ does not further simplify to more elegant forms. But it is evident that there is no guarantee that this difference is nonnegative since it could not be simplified to the expectation of a quadratic form (this is in sharp contrast to the analogous results developed for estimating the average treatment effect, or ATE). Hence even if all models are correct, the double-robust estimator is not necessarily asymptotically more efficient than the weighting estimator.

References

AASHTO (2010). *Highway Safety Manual*. Washington, D.C., American Association of State Highwayand Transportation Officials (AASHTO).

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72(1):1–19.

Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics: an Empiricist's Companion*. Princeton University Press, Princeton, NJ.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 60(1):47–57.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–972.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Callaway, B. and Sant'Anna, P. (2018). Difference-in-differences with multiple time periods and an application on the minimum wage and employment. *arXiv:1803.09015v2*, pages 1–47.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 82(1):772–793.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY.

Gårder, P. and Davies, M. (2006). Safety effect of continuous shoulder rumble strips on rural interstates in Maine. *Transportation Research Record*, 1953(1):156–162.

Griffith, M. (1999). Safety evaluation of rolled-in continuous shoulder rumble strips installed on freeways. *Transportation Research Record*, 1665(1):28–34.

Hauer, E. (1997). *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Emerald Group Publishing Limited, Oxford, OX, U.K., Pergamon.

Heckman, J. J. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Karwa, V., Slavkovic, A. B., and Donnell, E. T. (2011). Causal inference in transportation safety studies: Comparison of potential outcomes and diagrams. *The Annals of Applied Statistics*, 5(2B):1428–1455.

Khan, M., Abdel-Rahim, A., and Williams, C. J. (2015). Potential crash reduction benefits of shoulder rumble strips in two-lane rural highways. *Accident Analysis and Prevention*, 75:35–42.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, 4(3):165–224.

Lee, B. K., Lessler, J., and Stuart, E. A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine*, (29):337–346.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2:18–22.

Lyon, C., Persaud, B., and Eccles, K. (2015). Safety evaluation of centerline plus shoulder rumble strips. Technical Report Report No. FHWA-HRT-15-048, Federal Highway Administration, McLean, VA.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., , and Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, (9):403–425.

Mercatanti, A. and Li, F. (2014). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Annals of Applied Statistics*, 8(4):2485–2508.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression-coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer Series in Statistics. Springer, New York, NY.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer, New York, NY.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York, NY.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Wood, J. and Donnell, E. T. (2016). Safety evaluation of continuous green T intersections: A propensity scores-genetic matching-potential outcomes approach. *Accident Analysis and Prevention*, 93:1–13.

Wood, J. S. and Donnell, E. T. (2017). Causal inference framework for generalizable safety effect estimates. *Accident Analysis and Prevention*, 104:74–87.

Wood, J. S., Donnell, E. T., and Porter, R. J. (2015a). Comparison of safety effect estimates obtained from empirical Bayes before-after study, propensity scores-potential outcomes framework, and regression model with cross-sectional data. *Accident Analysis and Prevention*, 75:144–154.

Wood, J. S., Gooch, J. P., and Donnell, E. T. (2015b). Estimating the safety effects of lane widths on urban streets in Nebraska using the propensity scores-potential outcomes framework. *Accident Analysis and Prevention*, 82:180–191.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493.